Genetic and non-genetic determinants of cell-growth variation assessed by high-throughput microscopy - Supplementary material

Naomi Ziv, Mark L. Siegal and David Gresham

Note about mixed effect modeling

Model construction

Mixed-effect models contain both fixed and random terms. The levels of fixed-effect terms (such as concentrations of glucose) are repeatable and of interest in and of themselves, rather than instantiations drawn randomly from a larger population. By contrast, the levels of random terms (such as plates) are sampled from a potentially infinite population. For all models the fixed terms were genotype, environment and the genotype-environment interaction and the random terms were wells (explicitly nested within plates), plates and the interactions of plate with genotype and environment.

The parameters of a mixed model are the fixed terms' regression coefficients, the random terms' variances and the error variance. All measurements (after appropriate transformation) were modeled with Gaussian distributions for error and random effects. Parameters were estimated using restricted maximum likelihood (REML), using the *lmer* function in the *lme4* R package. The complete dataset and an R script containing the analysis are provided as supplementary material.

We determined the significance of terms by performing a series of likelihood ratio tests, using the *anova* function, on nested models fit to the same dataset. For analyses of lag duration and absolute growth-rate deviation, log transformations were used to reduce heteroscedasticity and non-normality of the measurements, improving the accuracy of estimations and their associated errors. Table 1 presents datasets and models analyzed in this paper.

Table 1. Datasets and analytical strategy. The complete study was performed and analyzed as four datasets.

Dataset	Structure	Models discussed
LWO – Lab, Vineyard, Oak and Oak-Vineyard F1 strains in a range of glucose concentrations *	4 genotypes X 7 environments X 3 wells per plate X 4 plates	Growth rate Lag duration
WILD – A range of different wild yeast isolates in two glucose concentrations **	12 genotypes X 2 environments X 4 wells per plate X 4 plates	Growth rate Growth-rate deviation Lag duration
CIT1 – GFP fusion in Lab strain background in a range of glucose concentrations ***	1 genotype X 8 environments X 4 wells per plate X 3 plates	Growth rate CIT1 florescence
SOIL – Two wild yeast strains in two glucose concentrations ****	2 genotypes X 2 environments X 24 wells per plate X 2 plates	Growth rate Growth-rate deviation

*Experiments also contained an additional environment (12 wells per plate) of base media (containing no added glucose) in which we variably observed cells growing very slowly or not at all. Data from these wells were excluded from the analysis.

**In addition to strains reported in the main text, we studied a strain that we understood to be DBVPG6765, but we later suspected to be an additional Pennsylvanian Oak strain (possibly YPS129) based on an oak-like phenotypic profile and adjacent well positions in a frozen stock plate. Data for this strain were included in the analysis (i.e., contributed to normalization) but not presented.

***These experiments only contain 24 wells per plate due to long exposure time in the fluorescent channel. Experiments also contained a base media environment with no added

glucose. In contrast to the LWO dataset, the CIT1-GFP strain consistently grew slowly in the base media. Data for these wells were included in the analysis but not presented. ****This dataset was used to assess the reproducibility of growth rate deviation estimates (see below).

Fixed-effects estimation

Each model contains coefficients for each level of the fixed-effect terms. To estimate a specific genotype-environment combination, the relevant parameters were combined. For example, the estimate for strain A in environment 1 would be calculated by adding 4 parameters (the regression intercept and the coefficients for Strain A, Environment 1 and the Strain A-Environment 1 interaction). The error of this estimate is a combination of the errors of the individual parameters. Specifically, the standard error is the square root of the summed parameter variances and twice each of the co-variances. For the example of strain A in environment 1, this calculation would include four variances and six co-variances. 95% confidence intervals were calculated as plus or minus 1.96 multiplied by the standard error. These estimates and confidence intervals are shown in figures 3-6 of the main paper. If data were transformed for modeling, as in the case of log transformations for lag duration and absolute growth rate deviation, estimates and intervals were transformed back to the original scale for presentation.

Some genotype-environment combinations were reproduced in different datasets. Six combinations were shared between the LWO and WILD datasets and four were shared between the WILD and SOIL datasets. Although experiments for different datasets were conducted 6–12 months apart, the combined parameter estimates are highly reproducible (**Figure S7**).

Random-effects estimation

Unlike fixed effects, random effects are not modeled with a parameter for each level. Instead, the variances of the random effects are parameters in the model. However, an

estimate for each level of each random term can be calculated conditional on the model parameters. These conditional means are the best linear unbiased predictors (BLUPs). The assumption of normality of the random effects was confirmed by assessing quantile-quantile plots (**Figure S8**). Confidence intervals on the BLUPs are calculated from the conditional variance-covariance matrices (posterior variances).

Adjusted pooled distributions

The fixed-effect terms represent the factors of interest (genotype and environment) whereas the random-effect terms capture various aspects of technical variation (focus, illumination, media preparation and cell preparation). Although our estimates of the fixed-effect parameters were sufficient for some analyses (e.g., comparison of average growth-rate deviation between the Netherlands and Finland soil strains), other analyses required distributions of measurements for each microcolony (e.g., correlation between CIT1 expression and growth rate within conditions). In the latter cases, the relevant random-effect conditional means were subtracted from each original measurement, creating adjusted values that could be pooled by condition across wells and plates. For example if a specific well designated A1.110316 contained strain A and environment 1, four values would be subtracted from each original microcolony measurement (the effects for well A1.110316, plate 110316, strain A-plate 110316 interaction and environment 1-plate 110316 interaction). After this normalization, we find that the means of all wells with the same condition are nearly identical (**Figure S9**). This approach is analogous to using linear mixed modeling to normalize data from cDNA microarray experiments (Wolfinger et al. 2001).

Levy SF, Ziv N, Siegal ML. 2012. Bet Hedging in Yeast by Heterogeneous, Age-Correlated Expression of a Stress Protectant.Hurst LD, editor. PLoS Biology 10:e1001325.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. 2001. Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. Journal of Computational Biology 8:625–637.

Table S1. Assessment of alternative models relating growth rate to nutrient concentration

Model	Strain	Parameters			Fit		
		μ_{max}	K_s		AIC	BIC	
$\mu = \frac{\mu_{max}\, s}{K_s + s}$	Lab	0.433	0.202		-161830.7	-161804.7	
	Oak	0.519	0.115		-103657.9	-103633.2	
	Vineyard	0.483	0.189		-169043.2	-169017.2	
	Oak/Vineyard F1	0.520	0.143		-150138.7	-150113	
		μ_{max}	K_s	a	AIC	BIC	P-value*
$\mu = \frac{\mu_{max}\,(s - a)}{K_s + s - a}$	Lab	0.433	0.202	-0.0001	-161828.8	-161794.2	0.729
	Oak	0.515	0.101	0.011	-104353.7	-104320.7	< 2.2e-16
	Vineyard	0.478	0.168	0.0149	-170171.2	-170136.5	< 2.2e-16
	Oak/Vineyard F1	0.518	0.135	0.005	-150355.7	-150321.4	< 2.2e-16
		μ_{max}	K_s	a	AIC	BIC	P-value*
$\mu = \frac{\mu_{max}\, s - K_s\, a}{K_s + s}$	Lab	0.433	0.202	-0.0003	-161828.8	-161794.2	0.729
	Oak	0.515	0.090	0.062	-104353.7	-104320.7	< 2.2e-16
	Vineyard	0.478	0.153	0.046	-170171.2	-170136.5	< 2.2e-16
	Oak/Vineyard F1	0.518	0.130	0.022	-150355.7	-150321.4	< 2.2e-16
		μ_{max}	K_s	a	AIC	BIC	P-value*
$\mu = \frac{\mu_{max}\, s - K_s\, a - s\, a}{K_s + s}$	Lab	0.433	0.202	-0.0003	-161828.8	-161794.2	0.729
	Oak	0.577	0.090	0.062	-104353.7	-104320.7	< 2.2e-16
	Vineyard	0.524	0.153	0.046	-170171.2	-170136.5	< 2.2e-16
	Oak/Vineyard F1	0.540	0.130	0.022	-150355.7	-150321.4	< 2.2e-16
		a	b		AIC	BIC	P-value*
$\mu = a + b \ln(s)$	Lab	0.333	0.073		-142529.7	-142503.7	NA (models not nested)
	Oak	0.434	0.076		-80730.29	-80705.55	
	Vineyard	0.376	0.080		-139222.8	-139196.8	
	Oak/Vineyard F1	0.424	0.081		-119618.4	-119592.7	

* Likelihood-ratio test comparing fit of alternative model to that of first (basic Monod) model.

Figure S1

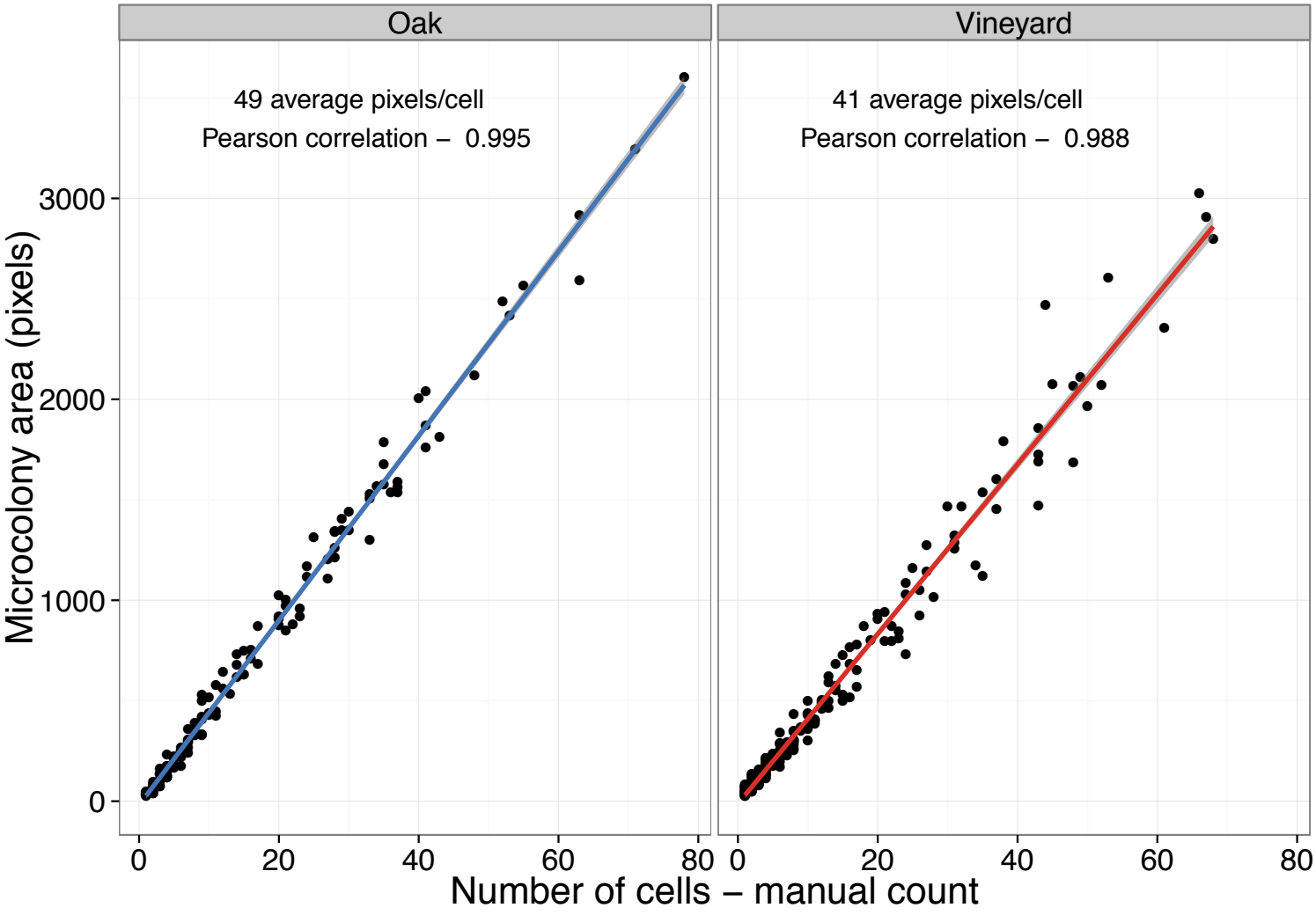

Figure S1 – Microcolony area is correlated with cell number for different strains growing in low glucose conditions.
The number of cells within each microcolony was counted at each time point for ten microcolonies for both the oak (left panel) and vineyard (right panel) strains growing in 0.22 mM glucose.
The average number of pixels per cell and Pearson correlation coefficient between the manual cell counts and automated area estimations are indicated in the upper left corner of each panel.

Figure S2

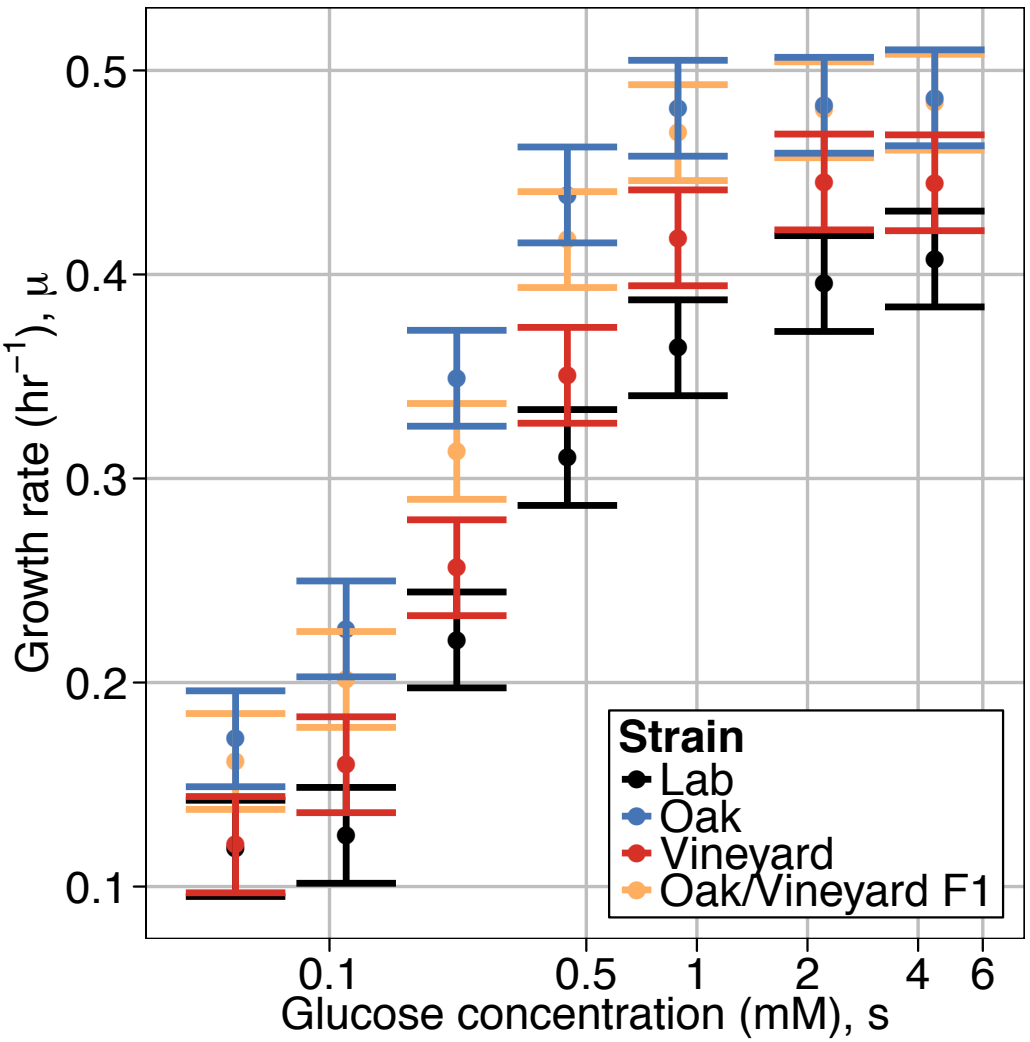

Figure S2 – Growth rate is determined by glucose concentration in a genotype specific manner. Growth rate estimates for each genotype and environment combination based on mixed effect modeling. Estimates are calculated as a combination of relevant fixed effect parameters, estimated using restricted maximum likelihood (REML) (see supplementary note). Glucose concentration is shown on a logarithmic scale for clarity. Error bars represent 95% confidence intervals.

Figure S3

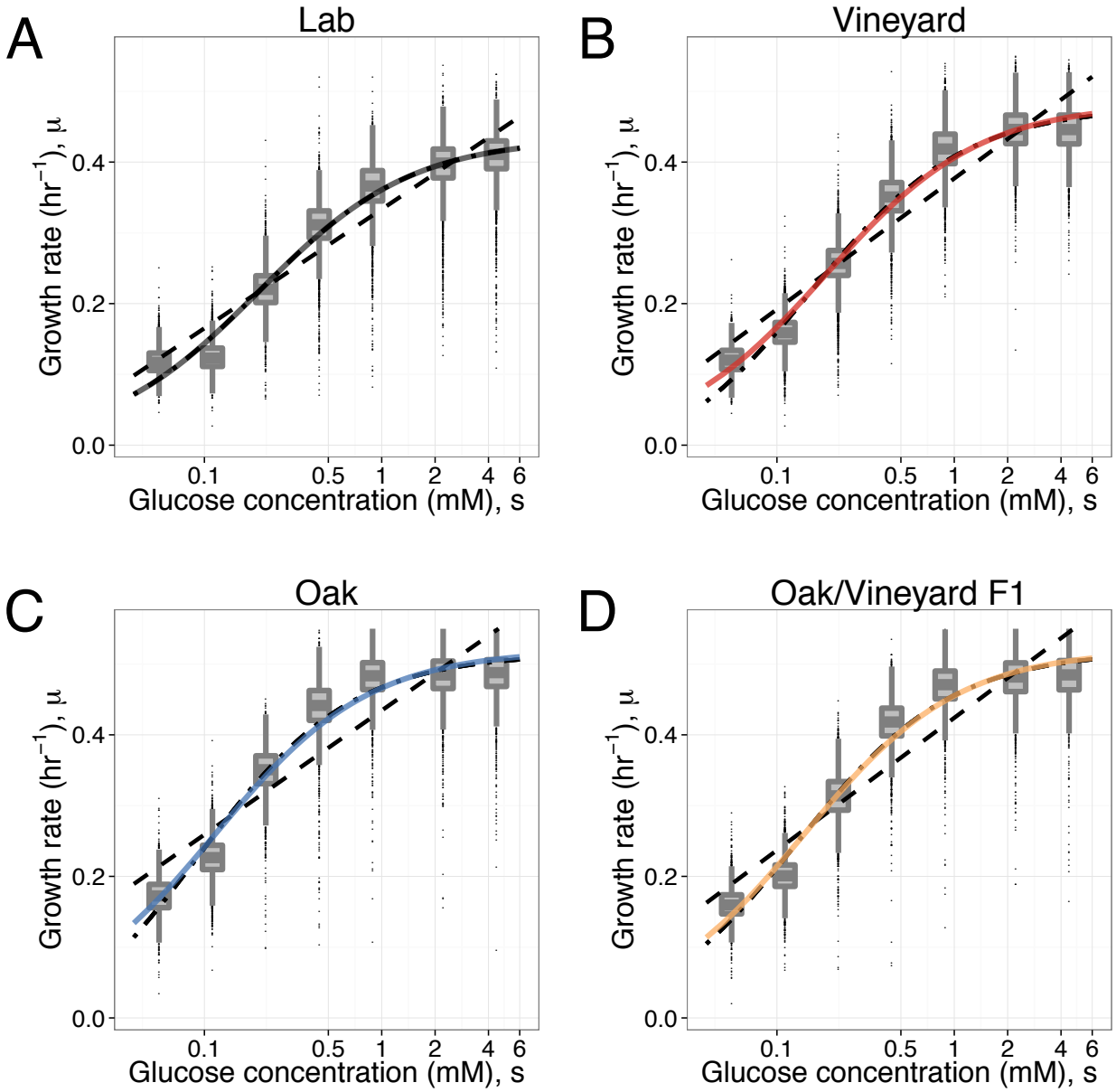

Figure S3 – Growth rate dependence on glucose concentration fits the Monod model for different strains. Growth-rate distributions for four strains (A-Lab, B-Vineyard, C-Oak and D-Oak/Vineyard F1) over a range of glucose concentrations. Solid curves depict the best fit of the Monod equation to the normalized data. Fits of the Westerhoff model are shown as dashed curves. Three variants of the Monod model with an additional parameter give virtually indistinguishable fits and are shown as a single dot-dashed curve in each panel. The glucose concentration is shown on a logarithmic scale for clarity.

Figure S4

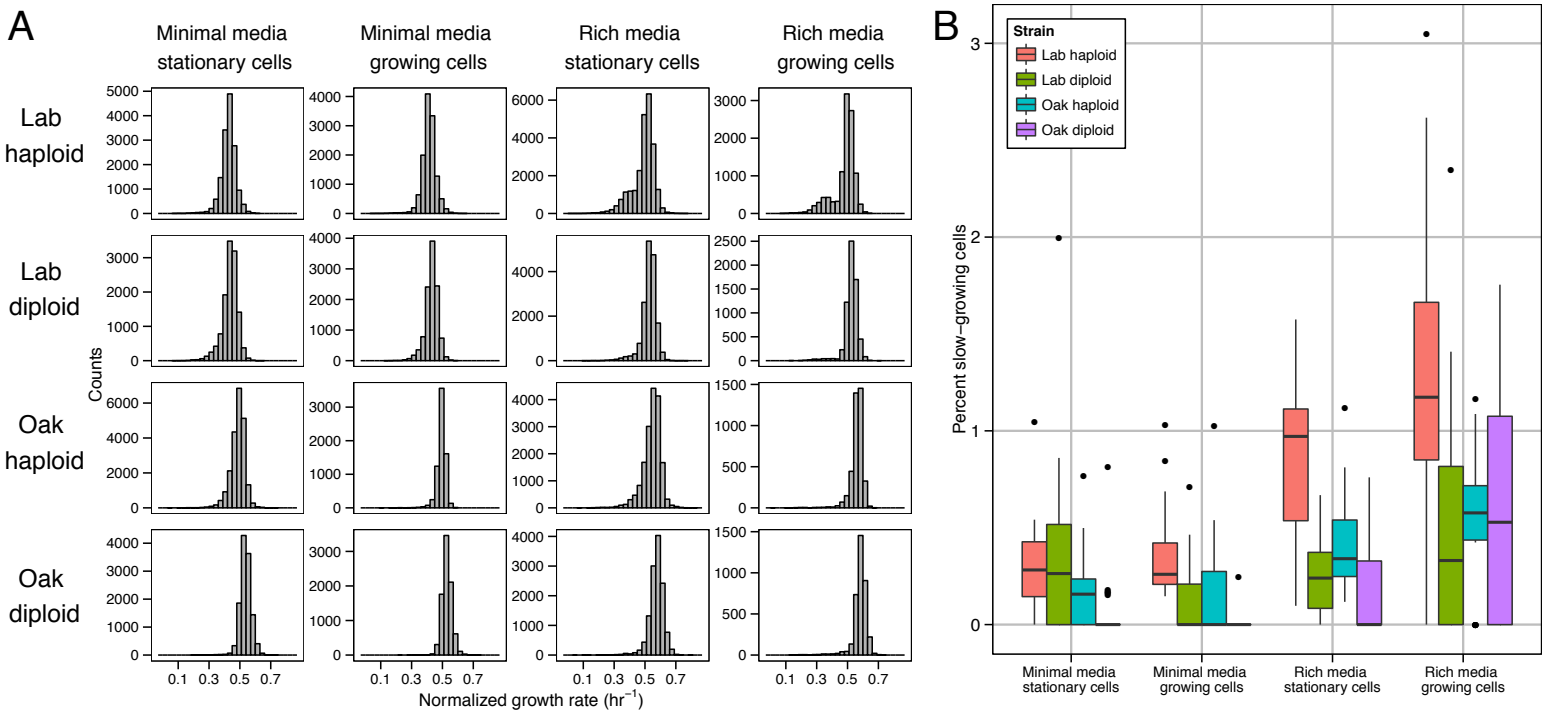

Figure S4 – Factors affecting the shape of the growth-rate distribution.
(A) Growth-rate distributions for four strains (haploid or diploid cells in the lab or oak genetic background) in four media/growth history conditions. Minimal medium is chemically defined carbon-limiting media with 0.08% glucose and rich medium is synthetic complete (SC) with 2% glucose. Cells were plated for imaging either from stationary cultures (~48 hours after 1:300 dilution) or actively growing cultures (4 hours growth after 1:50 dilution of the stationary culture). Distributions are from data pooled across four replicate plates, each with six wells per strain/condition combination, after normalization using mixed modeling. (B) Box-plots of percent slow-growing cells, calculated as previously (Levy et al. 2012) as the percentage of cells growing at less than half the median for each well.

Figure S5

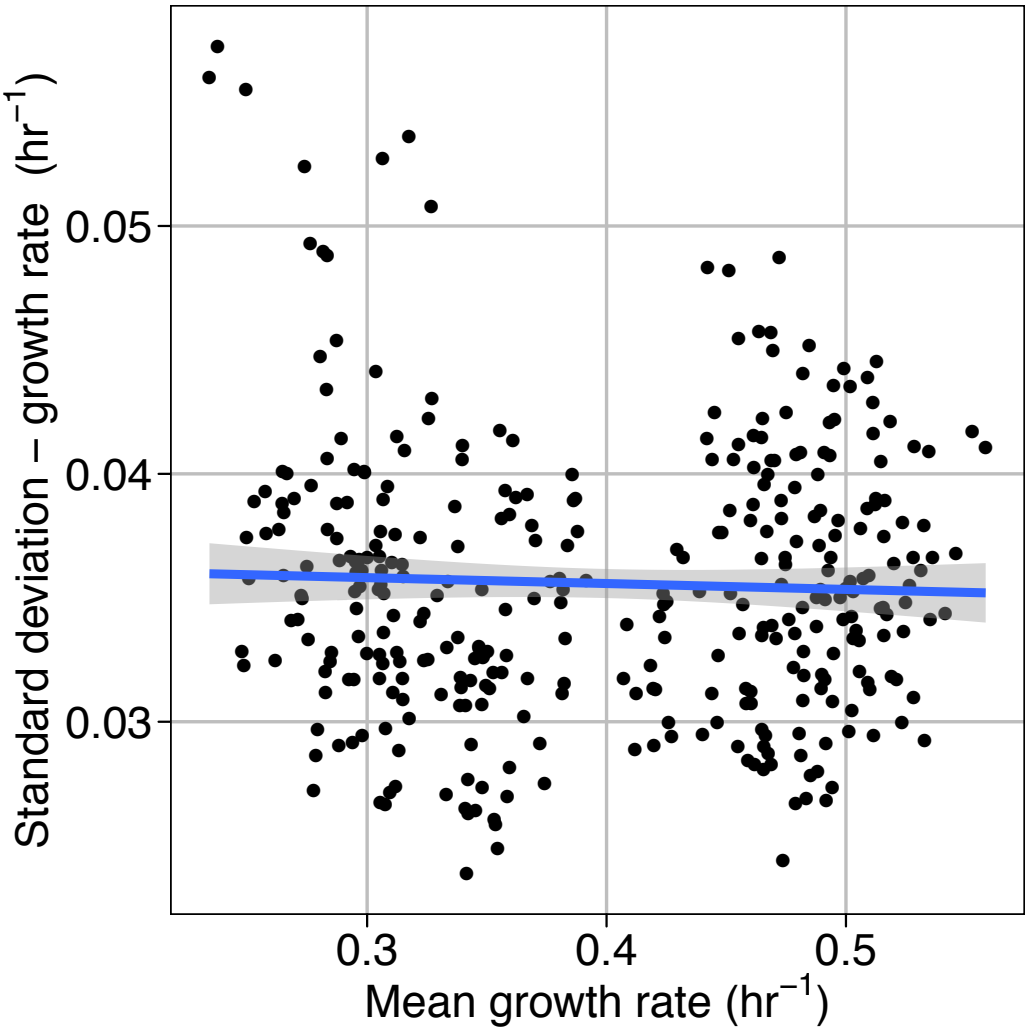

Figure S5 – Absence of mean-variance correlation for growth rate distributions.
Standard deviations verses means for 352 growth rate distributions representing eleven genotypes and two environments. The line depicts linear least squares regression.

Figure S6

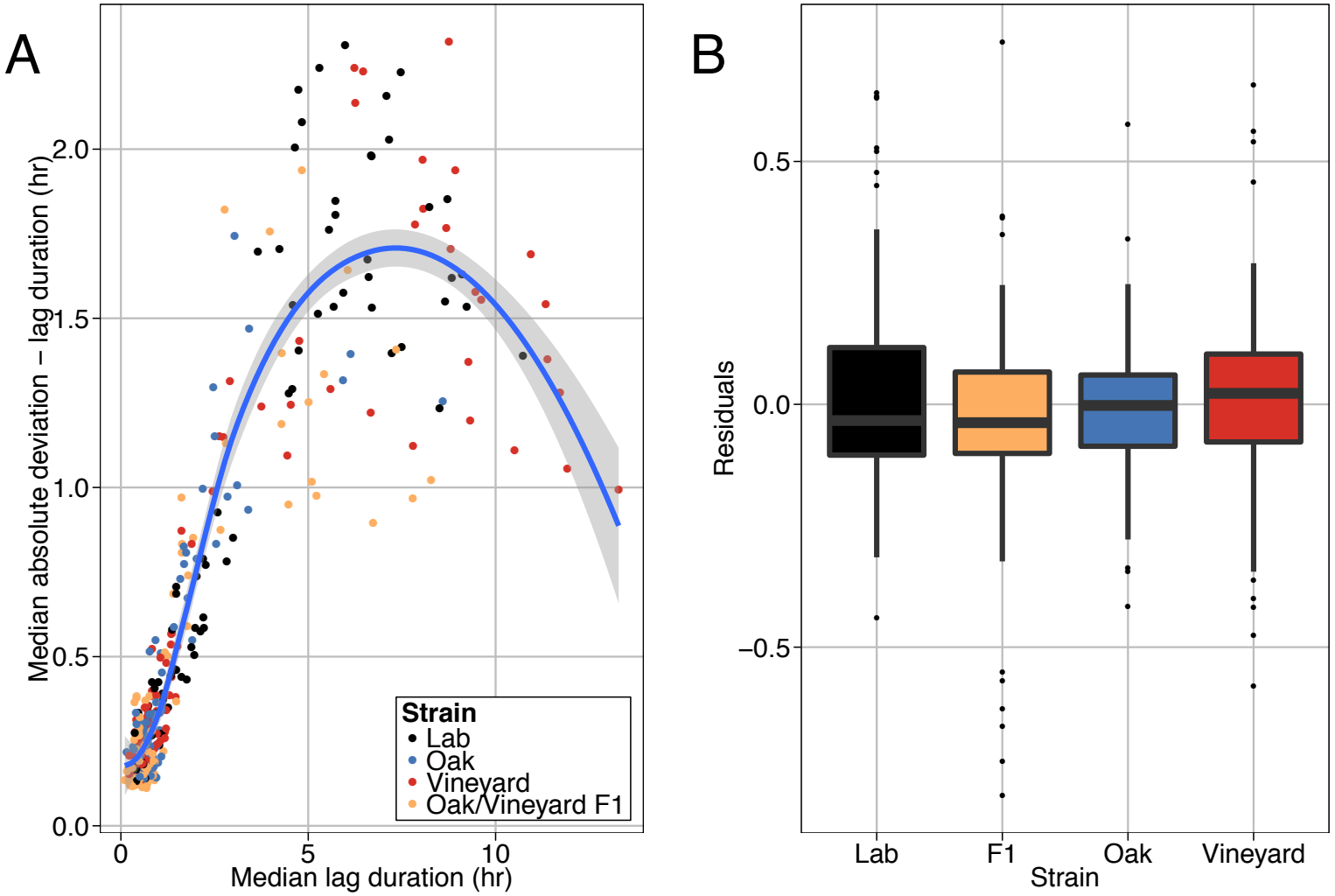

Figure S6 – No median independent genotype specific difference in lag duration variation.
(A) Median absolute deviations (MAD) verses medians for 333 lag duration distributions representing four genotypes and seven environments. The line depicts a local loess regression. (B) Distributions of residuals from loess regression shown in A, grouped by genotype. There is no significant difference between genotypes.

Figure S7

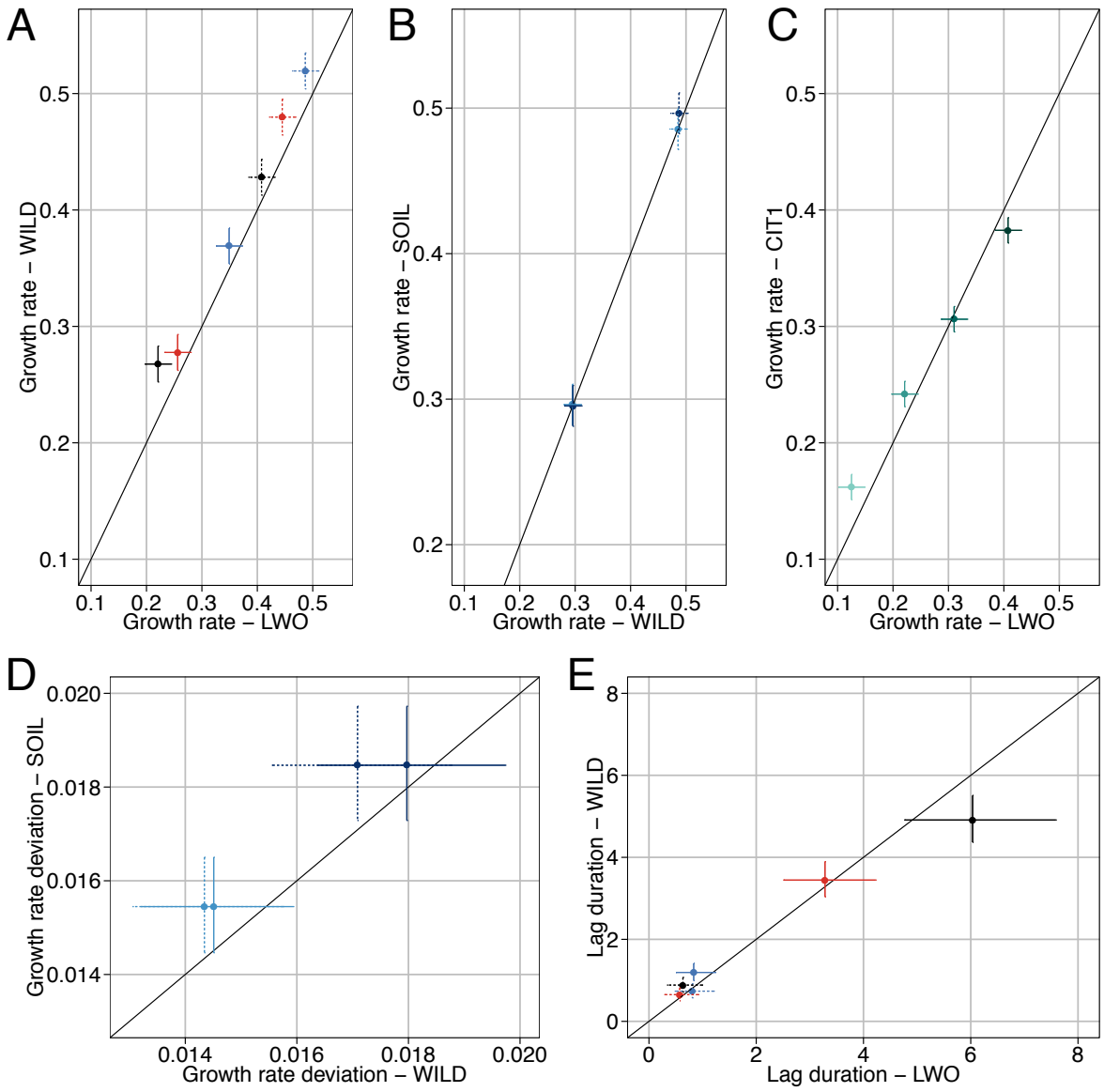

Figure S7 – Reproducibility of fixed effect parameters for growth rate (A-C), growth rate deviation (D) and lag duration (E).

In all panels, error bars represent 95% confidence intervals and the solid black line indicates x=y. (A and E) Comparison of growth rate (A) or lag duration (E) estimates for 6 genotype-environment combinations included in both the LWO and WILD datasets. The line type depicts glucose concentration (solid – 0.22mM, dashed – 4.44mM) and color depicts genotype (blue – Oak, red – Vineyard, black – Lab). (B and D) Comparison of growth rate (B) or growth rate deviation (D) estimates for 4 genotype-environment combinations included in both the WILD and SOIL datasets. The line type depicts glucose concentration (solid – 0.22mM, dashed – 4.44mM) and color depicts genotype (light blue – Finland, dark blue – Netherlands). (C) Comparison of growth rate estimates for 4 similar genotype-environment combinations between the LWO and CIT1 datasets. While the environments compared are identical, the genotype in the two datasets differs slightly: the CIT1-GFP strain differs from the lab strain used in the LWO experiments, as it is heterozygous for the CIT1-GFP fusion and a number of auxotrophies. Color depicts glucose concentration (from light to dark green – 0.11, 0.22, 0.44 and 4.44 mM).

Figure S8

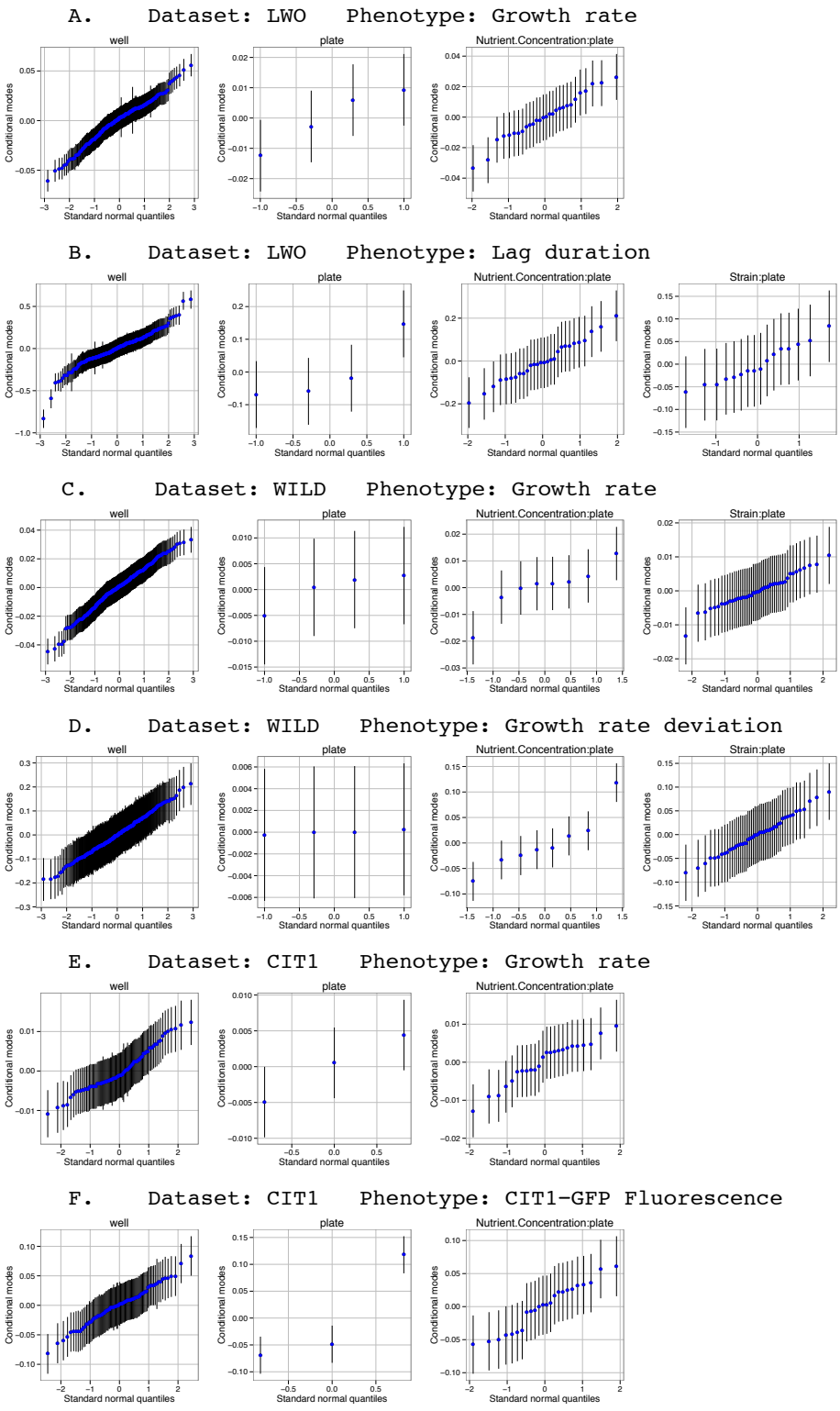

Figure S8 – Random effect conditional means.
Quantile-quantile plots comparing random effect conditional means for each random effect term in 6 distinct mixed models to standard normal quantiles. Panels represent models of growth rate (A, C and E), lag duration (B), growth rate deviation (D) and CIT1-GFP fluorescence (F). Panels also represent the three main datasets: LWO (A and B), WILD (C and D) and CIT1 (E and F). Vertical columns represent different random effect terms, from left to right: well, plate, environment-plate interaction and genotype-plate interaction (not present in all models). Blue dots represent estimates and black lines represent 95% confidence intervals.

Figure S9

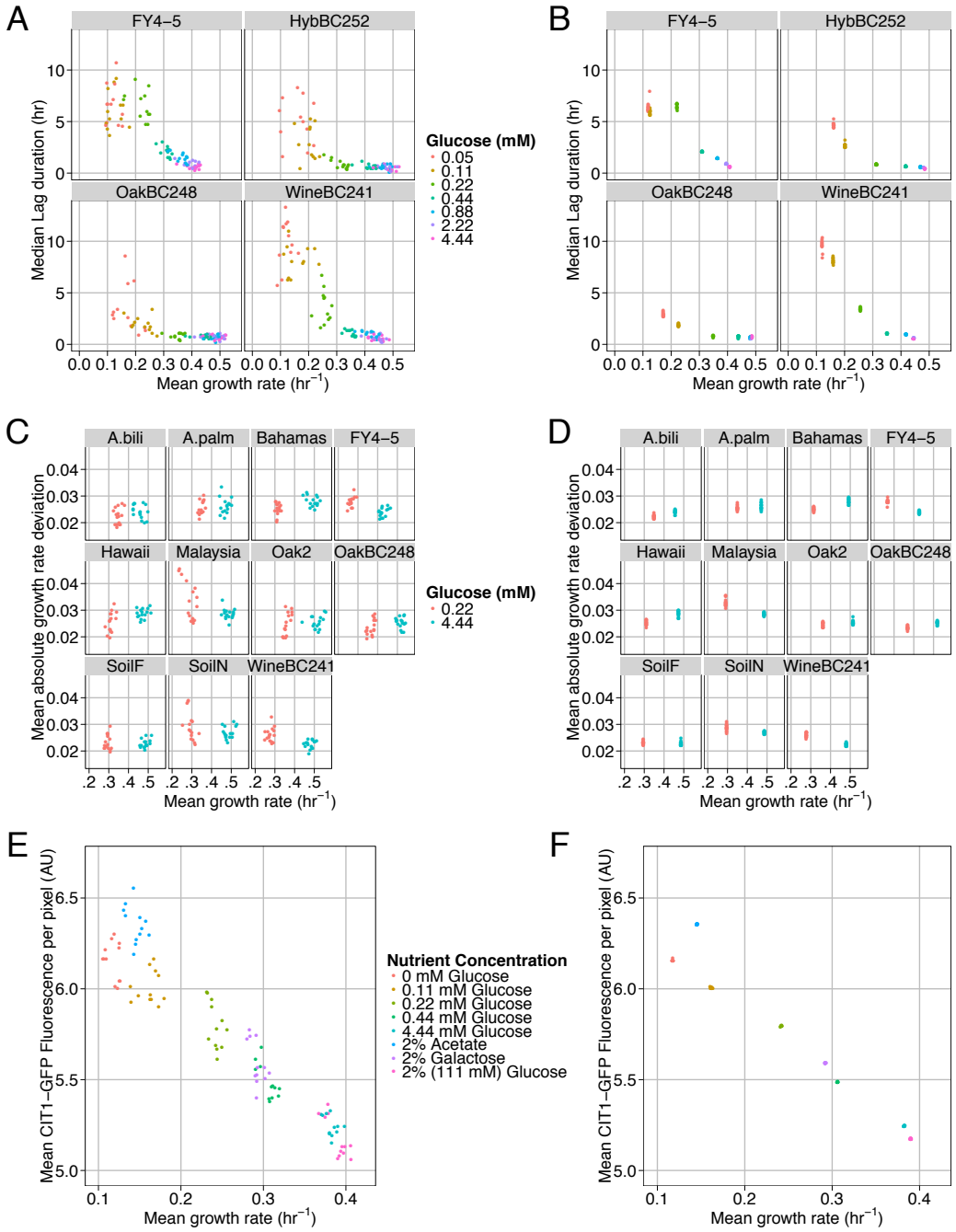

Figure S9 – Consequences of normalization by subtraction of random effect conditional means.
In all panels, a single dot corresponds to mean or median estimates of distributions originating from a single experimental well. Left panels (A, C and E) depict original estimates before adjustment; right side panels (B, D and F) depict corresponding estimates after adjustment. The panels represent estimates of growth rate verses lag duration in the LWO dataset (A and B), growth rate verses growth rate deviation in the WILD dataset (C and D) and growth rate verses CIT1-GFP fluorescence in the CIT1 dataset (E and F). In some panels (A-D), estimates from the same genotype are grouped and represented as different labeled facets. In all panels colors correspond to different environments, which are specified in a legend between each pair of panels. For lag duration and growth rate deviation, estimates are presented on the original scale despite modeling log-transformed distributions, resulting in greater spread between same-condition estimates.