Assigning function to natural allelic variation via dynamic modeling of gene network induction

Magali Richard[1,2,#], Florent Chuffart[1], Hélène Duplus-Bottin[1], Fanny Pouyet[1], Martin Spichty[1], Etienne Fulcrand[1], Marianne Entrevan[1], Audrey Barthelaix[1], Michael Springer[3], Daniel Jost[2,#] and Gaël Yvert[1,#]

1) Laboratoire de Biologie et de Modélisation de la Cellule, Ecole Normale Supérieure de Lyon, CNRS, Université Lyon 1, Université de Lyon; 69007 Lyon; France.

2) Univ Grenoble Alpes, CNRS, TIMC-IMAG, F38000 Grenoble, France.

3) Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA.

#) corresponding authors: Experiments: magali.richard@univ-grenoble-alpes.fr and gael.yvert@ens-lyon.fr Modeling: daniel.jost@univ-grenoble-alpes.fr and gael.yvert@ens-lyon.fr

Contact Information:

Gael Yvert

Laboratory of Biology and Modeling of the Cell

Ecole Normale Supérieure de Lyon, CNRS,

46 Allée d'Italie, Lyon, F-69007, France

Ph: +33 4 72 72 87 17, Gael.Yvert@ens-lyon.fr

KEYWORDS: Genetic variation, complex traits, association mapping, systems biology, gene regulatory network, stochastic model, yeast, galactose, network dynamics, GAL, inference, single-cell, SNP function, eQTL, personalized medicine, genotype-phenotype.

ABSTRACT

More and more natural DNA variants are being linked to physiological traits. Yet, understanding what differences they make on molecular regulations remains challenging. Important properties of gene regulatory networks can be captured by computational models. If model parameters can be 'personalized' according to the genotype, their variation may then reveal how DNA variants operate in the network. Here, we combined experiments and computations to visualize natural alleles of the yeast *GAL3* gene in a space of model parameters describing the galactose response network. Alleles altering the activation of Gal3p by galactose were discriminated from those affecting its activity (production/degradation or efficiency of the activated protein). The approach allowed us to correctly predict that a non-synonymous SNP would change the binding affinity of Gal3p with the Gal80p transcriptional repressor. Our results illustrate how personalizing gene regulatory models can be used for the mechanistic interpretation of genetic variants.

INTRODUCTION

In the past decade, countless DNA variants have been associated to physiological traits. A major challenge now is to understand how they operate at the molecular level. This is a difficult task because the mechanistic consequences resulting from each variant are not easy to identify. Even when the function of a gene is well documented, investigators need to determine the tissues, cells or organelles in which a mutant allele makes a biological difference, the developmental stage at which this may happen, the metabolic or regulatory network that may be involved, as well as possible molecular scenarios. A mutation may alter the regulation of transcription or mRNA splicing; the enzymatic activity of the target protein; its rate of production, maturation, or degradation; its intracellular localisation; its binding affinity to an interacting partner or the specificity of its molecular interactions. In the vast majority of cases, information from the DNA sequence alone is not sufficient to delimit the perimeter of possible implications.

Systems biology has opened new opportunities to better predict the action of DNA variants. First, 'omics' data that are gathered at various levels (DNA, transcripts, proteins, metabolites...) establish relations between target sequences and functional pathways. Information about molecular and genetic interactions, expression profiles, chromatin landscapes, post-transcriptional and post-translational regulations can be exploited to derive functional predictions of DNA variants. Various methods have been proposed to do this, such as Bayesian genetic mapping[1], visualization of SNPs on relational protein networks[2], prioritization based on negative selection[3], or inference of miRNA:RNA binding defects[4]. In addition, structural data of biomolecules can also highlight functional perturbations in specific domains such as catalytic sites or interaction surfaces[5,6].

Another alternative is to model the quantitative and dynamic properties of molecular reactions and to explore which feature(s) may be affected by a DNA variant. The functional consequences of mutations can then be inferred by considering their impact on specific parameters of the model. In other words, assigning function to a DNA variant may be straightforward after it is linked to parameters of a model. This perspective may also, on the long term, generate developments in personalized medicine: if a model can be *personalized* according to the patient's genotype then it can help predict disease progress or treatment outcome and therefore adapt medical care to the patient's specificities. For this to become reality, the model must be i) informative on the biological trait of interest and ii) identifiable and sufficiently constrained so that model parameters can be reliably inferred, accounting for the patient's specificities. These two requirements antagonize each other regarding the complexity of the model to be used. The former asks for completeness: the molecular control of the trait must be correctly covered by the model, describing known reactions as best as possible. The latter asks for simplicity: if too many parameters are allowed to be adjusted to the data, then the validity of the personalized model is questionable and none of the adjustments are informative. It is therefore important to determine if and how personalizing model parameters can be productive.

For a given molecular network, individuals from natural populations have different genotypes at several nodes (genes) of the network, as well as in numerous external factors that can affect the network properties. Such external factors can modify, for example, global translation efficiencies, metabolic states, or pathways that cross-talk with the network of interest. Adapting model parameters to specific individuals is challenging when so many sources of variation exist. A way to circumvent this difficulty is to study the network

96 experimentally in the context of a more reduced and focused variation. If investigators have
97 access to nearly-isogenic individuals that differ only at specific genes of the network, they can
98 then characterize the differences in network behaviour that result from these specific allelic
99 differences. The numerous external factors affecting the network can then be ignored or
100 drastically simplified in the model because they are common to all individuals. This way, the
101 parameter space is constrained and only potentially-informative parameters are allowed to be
102 adjusted to fit individual-specific data.
103
104 Some model organisms such as the yeast *S. cerevisiae* offer this possibility. They can
105 be manipulated to generate single allelic changes, which provides an ideal framework to link
106 DNA variants to model parameters. In particular, the gene regulatory network controlling the
107 yeast response to galactose (GAL network) is well characterized, both *in vivo* and *in silico*.
108 This circuit controls galactose utilization by upregulating the expression of regulatory and
109 metabolic genes in response to extracellular galactose[7]. Regulation is based on the
110 transcriptional activator Gal4p, the galactose transporter Gal2p, a signal transducer Gal3p and
111 the transcriptional inhibitor Gal80p. In addition, the galactokinase Gal1p, involved in
112 galactose metabolism is also a coinducer of the response[8]. This system can display either a
113 gradual induction (where the rate of transcription progressively increases in each cell
114 according to the timing and intensity of the stimulus) or a probabilistic induction (where the
115 probability of having high/low rate of transcription in each cell varies). This dual behaviour
116 has received a lot of attention and important molecular features have been elucidated by
117 experimental and theoretical approaches[9–12]. In particular, the dynamic response of a
118 population of cells to galactose can be described by two quantities: the inducibility of the
119 network is defined as the proportion of activated cells in the population, and the amplitude of
120 the response refers to the expression level that is reached by induced cells. Regulatory

feedback loops of the network are critical to the switch-like behaviour. They were shown to feed back the dynamics of transcription bursts rather than the levels of expression[13]. They regulate the amplitude response by reducing noise in GAL gene expression[14], they control the inducibility by fine-tuning the timing of the switch[14], and they participate to the memory of previous inductions[15,16]. As a consequence, bimodal distributions of expression of the GAL genes can be observed in isogenic populations exposed to intermediate concentrations of inducer[17–19], and this population heterogeneity can confer a growth advantage during the transition from glucose to galactose metabolism (diauxic shift)[20]. Interestingly, wild yeast isolates present diverse types of induction dynamics during the diauxic shift, ranging from strictly unimodal to transient bimodal distribution of expression levels[21,22]. This indicates that natural genetic variation can modify the network dynamics.

The *GAL3* gene plays a central role in the network. Its protein product Gal3p is activated by binding to galactose and ATP and then binds as a dimer to Gal80p dimers to release the repression on Gal4p at target promoters[7]. The protein is enriched in the cytoplasm prior to stimulation and in the nucleus after the stimulation, although this cyto-nuclear transfer does not account for the dynamics of activation[23,24]. Expression of *GAL3* is itself under Gal4p/Gal80p control (positive feedback). In addition, the sequence of *GAL3* differs between natural isolates of *S. cerevisiae* and this allelic variation was recently associated to different sensitivities of the network to galactose (Lee et al. PLoS Genetics, *in press*). There are multiple ways that a *GAL3* variant could affect the dynamics of induction: by modifying the production or degradation rates of the Gal3p protein or of its messenger RNA, by changing the affinity of Gal3p to galactose or ATP, by changing the capacity of Gal3p to dimerize, by changing the nucleocytoplasmic ratio of Gal3p molecules, or by changing the affinity of Gal3p to Gal80p. A *GAL3* variant may also affect the background expression level

146 of Gal3p prior to stimulation, which is known to be critical for network memory of prior
147 stimulations[25]. Thus, it is difficult to predict the functional consequence of sequence variation
148 in *GAL3*.
149
150 Using the yeast *GAL3* gene as a model framework, we show here that experimental
151 acquisitions combined with network modeling is efficient to predict the effect of sequence
152 variants. The principle of the approach is to link genetic variation to informative changes of
153 parameter values of the model. We show that replacing natural *GAL3* alleles can be sufficient
154 to transform a gradual response into a probabilistic activation, and the approach allowed us to
155 distinguish between different types of *GAL3* alleles segregating in *S. cerevisiae* populations:
156 those altering the activation of Gal3p by galactose, and those altering the strength with which
157 activated Gal3p alleviates the transcriptional inhibition operated by Gal80p. In particular, our
158 approach was efficient to associate a non-synonymous SNP with a change of binding affinity
159 for Gal80p.
160

RESULTS

Natural variation in GAL3 affects the dynamics of network induction

We constructed a panel of yeast strains that were all isogenic to the reference laboratory strain BY, except for *GAL3*. At this locus, each strain carried an allele that was transferred from a natural strain of the *Saccharomyces* Genome Resequencing Project[26] (Supplementary Fig. 1). All strains of the panel also harboured a P_{GAL1}-*GFP* reporter of network activity, where the promoter of the *GAL1* gene controlled the expression of a GFP fluorescent protein destabilized by a degradation signal[27,28]. *GAL1* is a paralogous gene of *GAL3*[29] and transcription at its promoter is commonly used as a proxy of GAL network activity[15,20,22]. Using flow cytometry, we monitored the dynamics of network activation in each strain (Fig. 1). This was done by first culturing cells for 3 hours in a medium containing 2% raffinose, a sugar known to be neutral on network activity, adding galactose (0.5% final concentration), and quantifying fluorescence at multiple time points for 4 hours. Significant differences in the dynamics of activation were observed between the strains. Those harbouring the $GAL3^{NCYC361}$, $GAL3^{K11}$, $GAL3^{BY}$, $GAL3^{DBVPG1788}$, $GAL3^{DBVPG1853}$ and $GAL3^{JAY291}$ alleles displayed a gradual response, all cells of the population were induced and responded with similar rate of expression, maintaining population homogeneity (see example shown in Fig. 1a). In contrast, strains harbouring the $GAL3^{Y12}$ and $GAL3^{YJM978}$ alleles displayed a binary response, with a transient co-existence of induced (ON) and uninduced (OFF) cells in the population (example in Fig. 1b).

We quantified induction using two metrics: the mean level of reporter expression in activated cells (response amplitude), and the proportion of activated cells in the population (inducibility of the network). We observed that the response amplitude varied little among the strains, all of them approaching steady state with comparable kinetics (Fig. 1c). In contrast, inducibility of the network differed between strains (Fig. 1d). As expected, in strains showing

187 a gradual response, the fraction of ON cells increased significantly during the first two hours
188 of induction, reaching full inducibility (all cells activated) by the end of the experiment. On
189 the opposite, the strains showing a transient binary response displayed reduced inducibility
190 over time. For instance, 21% of *GAL3*^{Y12} cells were still not induced after 250 minutes of
191 stimulation. These results indicate that natural genetic variation in *GAL3* is sufficient to
192 modify the inducibility of the network and to convert a gradual response into a binary
193 response, or vice versa.
194
195 **A quantitative model of inducibility over time**
196 To examine what functional properties of the *GAL3* gene could determine a gradual or
197 a binary response, we constructed a dynamic stochastic model of the network (Fig. 2a). We
198 based our quantitative model on the following current molecular knowledge. In absence of
199 galactose, a homodimer of the transcription factor Gal4p is constitutively bound to upstream
200 activation sites (UAS) of promoter regions of GAL genes. However, transcription is inactive
201 because of the homodimeric Gal80p inhibition of Gal4p[30,31]. When intracellular galactose
202 binds Gal3p, it changes conformation and associates with Gal80p dimers[32], thereby releasing
203 Gal80p from promoters and allowing Gal4p-mediated transcriptional activation. It was
204 initially thought that activated Gal3p sequestered Gal80p in the cytoplasm, preventing it from
205 its inhibitory role in the nucleus[30]. Later studies revised this view by showing that Gal3p
206 molecules were not exclusively cytoplasmic[23], that forcing Gal3p to be mostly nuclear did not
207 alter the kinetics of induction[23], and that the dynamics of nucleocytoplasmic trafficking were
208 too slow to explain the fast induction of transcription[24]. This implies a direct role of Gal3p in
209 promoting the dissociation of Gal80p from UAS. In addition, the galactokinase Gal1p (a
210 paralog of Gal3p) can also act as a co-inducer of the regulatory circuit, presumably using
211 similar mechanisms as Gal3p[18].

Our model covers the mRNA and protein species of three major players of GAL network induction: GAL1, GAL3 and GAL80, as well as of the reporter gene. We considered that promoters of each GAL gene could switch between an ON state (full transcription) and an OFF state (leaky transcription) at rates that depended on the concentration of Gal80 dimers, activated Gal3p dimers and activated Gal1p dimers. A detailed description of the model is given in Materials and Methods and in Supplementary Text 1. Most of the parameters of the model were fixed at values obtained from previous studies (Table S1).

Stochastic simulations reproduce the two types of induction observed experimentally

We first explored if our model captured the two types of responses of allele-replacement strains (*i.e.* binary and gradual). We ran stochastic simulations[33] that accounted for intrinsic and extrinsic sources of noise (see Supplementary Text 1). We observed that tuning the parameters related to *GAL3*, while keeping all other parameters constant, was sufficient to modify inducibility and to obtain either a gradual (Fig. 2b) or a binary (Fig. 2c) response of the network at a given concentration of galactose. In the gradual system, the simulated single-cell trajectories were all similar; in the binary system, the simulated single-cell trajectories bifurcated with a subset of cells having a stochastic lagging time before responding. The single-cell value of this lag time is directly correlated with the number of potential inducer proteins (Gal1p and Galp3p) present in the cell just before induction (Supplementary Fig. 2). This is in very good agreement with recent single-cell experiments on galactose induction[25].

We then studied the response predicted by the model when stimulating the network with various concentrations of galactose while keeping model parameters constant (Supplementary Fig. 3). Inducibility increased with the concentration of galactose, with low concentrations causing a probabilistic induction (binary) and high concentrations a

237 deterministic one (gradual).

238

239 **Two parameters related to GAL3 control network behaviour**

240 A detailed analysis of the model showed that inducibility of the system was mainly
241 controlled by the average values of promoter switching rates k_{on} and k_{off} at the time of
242 induction (see Material and Methods, Supplementary Text 1 and Supplementary Figs. 2 and
243 4). Rates k_{off} depend only on *GAL80* and are therefore invariant to *GAL3* allelic variation.
244 Rates k_{on} depend on *GAL3* in two ways: via Gal3p*, the amount of galactose-activated Gal3p,
245 and via K_3, which corresponds to an effective concentration encompassing the dissociation
246 constants of the Gal3p-Gal80p interaction and of Gal3p dimerization (see Supplementary
247 Text 1). *Gal3p** is determined by the level of Gal3p and by parameter K_{gal}, which represents
248 the typical concentration of galactose needed to efficiently activate Gal3p. While K_{gal} was
249 identifiable, several other GAL3-related parameters, such as those controlling the level of
250 Gal3p, were not and we grouped them in a meta-parameter, ρ_{Gal3}, which we termed the
251 *strength* of GAL3. ρ_{Gal3} corresponds to the invert ratio between K_3 and the mean
252 concentration of Gal3p at the time of induction, which depends on the leaky transcription rate,
253 the translation rate and the degradation rates of *GAL3* mRNA and protein product.

254

255 This formalism made the network sensitive to only two identifiable GAL3-related
256 parameters, K_{gal} and ρ_{Gal3}. At a fixed concentration of galactose induction, high ρ_{Gal3} values
257 correspond to high numbers of Gal3p dimers that can rapidly be activated to release Gal80
258 repression. The model predicted that high values of ρ_{Gal3} would generate a gradual response
259 (Supplementary Fig. 5a) because the number of potential activators was high enough in each
260 cell to rapidly trigger the GAL1/GAL3-mediated positive feedback loop. In contrast, low
261 values of ρ_{Gal3} would generate a binary response (Supplementary Fig. 5b) because the number

of activators is more stochastic, with many cells having too few initial Gal1p or Gal3p dimers to directly trigger the response. These cells need a lag time before fast activation (Fig.2b,c and Supplementary Fig. 2). The other important parameter, K_{gal}, corresponded to a threshold of galactose concentration below which induction was limited and favoured a binary response, and above which induction was efficient and favoured a gradual response (Supplementary Fig. 3c). In summary, both ρ_{Gal3} and K_{gal} values can determine whether the network adopts a gradual or a binary response at a given concentration of galactose induction.

Linking *GAL3* alleles to specific parameter values

In order to test the predictions of the model, we measured the transcriptional response of the $GAL3^{BY}$, $GAL3^{Y12}$ and $GAL3^{YJM978}$ strains at different galactose concentrations (0.05%, 0.1% and 0.5%). Our experimental observations confirmed that the inducibility increases with galactose concentration (Fig. 3a). We then used this experimental data to infer parameters ρ_{Gal3} and K_{Gal} for each of the three strains. This was done by selecting a set of parameters that minimized a global chi2-score of deviation between the measured and predicted fractions of induced cells at different times after induction and for the different galactose concentrations (for details, see methods and Supplementary Text 1). To evaluate the usefulness of the inferred parameter values, we used the fitted model to predict the behaviour of each strain at a galactose concentration that was not used for model training (0.2%) (Fig. 3b). Finally, to test model predictions, we experimentally monitored $GAL3^{BY}$, $GAL3^{Y12}$ and $GAL3^{YJM978}$ induction at 0.2% galactose. Without any additional fitting procedure, we observed that inducibility (fraction of activated cells over time) differed between strains in a way that was entirely consistent with model predictions. Thus, the differences among parameter values assigned to the different *GAL3* alleles are relevant outside the specific experimental conditions used for parameters estimation.

287

288 **Natural *GAL3* alleles map to distinct locations of the parameter space**

289 We sought to classify *GAL3* alleles based on the parameter values assigned to them.
290 We made experimental measurements on two additional strains ($GAL3^{NCYC361}$ and
291 $GAL3^{DBVPG1788}$) and we determined best-fit ρ_{Gal3} and K_{gal} values to them as for the three
292 strains described above. This data and the corresponding fitted models are shown in
293 Supplementary Fig. 6. Fig. 4a,b shows the obtained parameters, ρ_{Gal3} and K_{Gal}, normalized
294 by the corresponding values of our reference strain $GAL3^{BY}$. Different data points represent
295 results obtained by applying the inference process to models with different GAL3-
296 independent parameters (see Supplementary Text 1).

297 The fold change of a parameter between two different strains is indicative of the
298 functional nature of the genetic variations between the two *GAL3* alleles. In agreement with
299 the model predictions (Supplementary Fig. 3), we observed that more gradual strains
300 ($GAL3^{NCYC361}$ and $GAL3^{DBVPG1788}$) display a high GAL3 strength ρ_{Gal3} and a low 'typical'
301 galactose concentration K_{Gal}. Interestingly, we observed that ρ_{Gal3} and K_{Gal} can be de-
302 correlated. In particular, although both $GAL3^{YJM978}$ and $GAL3^{Y12}$ strains were binary
303 responders at all galactose concentrations tested, the model attributed this behaviour to
304 different functional effects: a low sensitivity to galactose (high K_{gal}) for the Gal3 protein
305 originating from YJM978 and a reduced strength of the *GAL3* gene originating from Y12.
306 Thus, the induction specificities of the strains can be attributed to distinct GAL3-related
307 parameters.

308 To address the direct relationship between the network properties (gradual or binary
309 response) and the GAL3-related parameters, we positioned each of the tested strains within a
310 phenotypic landscape according to their relative ρ_{Gal3} and K_{Gal} parameters (Fig 4c).
311 According to our model, ρ_{Gal3} and K_{Gal} parameters are sufficient to predict the behaviour

(gradual or binary) associated with a given *GAL3* allele at a given concentration of galactose. As an illustration of these predictions, we specifically observed the dynamics of transcriptional activation of the network for the strain $GAL3^{DBVPG1788}$ (Fig. 4d). The position of the $GAL3^{DBVPG1788}$ allele on the phenotypic landscape corresponded to a transient probabilistic activation at low concentration ([gal]=0.05%) converted into a gradual response at higher concentration ([gal]=0.1% and [gal]=0.5%).

Variation in induction dynamics is consistent with variation in diauxic shift decision

The physiological relevance of the GAL network regulation is to switch from the consumption of glucose (the preferred carbon source) to the consumption of galactose when glucose supply is running out. This diauxic switch is controlled not only by galactose induction but also by glucose-mediated repression. When both sugars are present, their relative concentration ratio determines whether cells activate the switch or not [20,21,34]. At some ratio values, only a fraction of the cells are induced, even at the steady-state. Given this dual regulation, the propensity of a strain to activate GAL metabolism can be quantified by measuring the fraction of induced cells after a prolonged period (8 hours) of simultaneous induction (by galactose) and repression (by glucose). If this measurement is repeated at a given concentration of galactose and various concentrations of glucose, a useful score can be computed (called 'decision threshold' hereafter): the concentration of glucose needed to maintain half the population of cells in the repressed (OFF) state (Fig. 5a). A high decision threshold corresponds to an early activation of GAL genes during the diauxic shift.

A previous study identified *GAL3* as an important genetic determinant for this decision: the concentration ratio at which cells turn GAL expression ON differs between strains harbouring different natural alleles of *GAL3* (Lee et al. PLoS Genetics, *in press*). We

337 asked if this variation was correlated with the variation observed on the dynamics of network
338 induction. We chose four strains that showed different decision thresholds because of
339 different *GAL3* alleles (Fig 5b) (Lee et al. PLoS Genetics, *in press*) and we monitored their
340 dynamics of induction at three different concentrations of galactose (with no glucose). We
341 then used our model to assign ρ_{Gal3} and K_{Gal} parameter values to each strain. Experimental
342 data and model fitting are shown in Fig. 5b and Supplementary Fig. 7. We used the inferred
343 parameter values to visualize the four strains in the parameter space where binary and gradual
344 responses upon stimulation at [gal] = 0.25% are delimited (Fig. 5d). Remarkably, the
345 properties of induction dynamics in absence of glucose were fully consistent with the decision
346 threshold during diauxic shift from glucose to galactose. Strains having a low decision
347 threshold, such as $GAL3^{YJM421}$, displayed a transient binary response, and strain $GAL3^{BC187}$
348 had a high decision threshold and responded gradually. Coordinates of strains in the
349 parameter space indicate that ρ_{Gal3} values are highly informative on the decision threshold
350 (Fig. 5d). Thus, mapping allelic variation to dynamic parameters of induction is also useful to
351 understand trade-offs that are observed at steady-state.
352
353 **A quantitative parameter change predicts a role of H352D SNP on Gal3:Gal80 complex**
354 **formation**
355 We noticed that, at position 352 of the Gal3p protein, all natural strains harboured an
356 aspartic acid, whereas the reference laboratory strain BY harboured a histidine. This aspartic
357 acid was also conserved in *S. mikatae*, *S. paradoxus* and *S. uvarum* protein sequences[35].
358 Given the prevalence of this aspartic acid, we hypothesized that a single H352D amino-acid
359 change could have consequences on Gal3p regulatory function.
360 To test this, we generated an artificial $GAL3^{BY\text{-}H352D}$ allele by introducing the H352D
361 mutation in the $GAL3^{BY}$ strain and we monitored the dynamics of induction of the resulting

strain. At similar concentrations of galactose, induction was faster for the modified strain than for the original strain (compare Fig. 6a with Fig. 2a). We then used our model to make functional predictions. We fitted our model to experimental data of induction as described above for natural alleles. Induction dynamics of the modified strain were fully explained by preserving parameter K_{Gal} and increasing ρ_{Gal3} (Fig. 6b). This suggested that the H352D mutation did not affect activation of Gal3p by galactose but rather the strength of Gal3p, which summarizes six biochemical features: the basal level of *GAL3* transcription prior to induction, its translation and degradation rate, the degradation rate of its coding mRNA, its capacity to homodimerize and the affinity of activated Gal3p for Gal80p.

How the implicated SNP could change either the leaky transcription level prior to induction or the transcription rate during induction is difficult to imagine. In addition, the amino-acid change was not surrounded by any particular peptide motif, nor was it located at the extremity of the protein. This did not support for an effect on translation or degradation rates. Thus, the most plausible interpretation of the parameter change of the model was that the H352D modification would increase either the capacity of Gal3p* to dimerize or the affinity of the Gal3p* dimer for Gal80p.

To explore these possibilities, we analyzed the structure of the heterotetramer $[Gal3p^*]_2$-$[Gal80p]_2$ that was previously solved[32]. We made three important observations. First, His352 is located at the binding interface of the Gal3p* dimer with the Gal80p dimer (Fig. 6c), and distant from the pocket containing galactose and ATP. Secondly, it is spatially close to the Gal80p site where the acidic domain of Gal4p is known to bind[36]. Finally, the Gal80p dimer exhibits a positive electrostatic surface potential in the vicinity of Gal3p-His352, suggesting that the replacement of the neutral His352 by a negatively charged aspartic acid would stabilize the Gal3p*-Gal80p complex. Stabilization refers here to a gain in thermodynamic stability relative to the Gal4p-Gal80p complex, or in other words, to a

387 decrease of Gibbs free energy change (ΔG_{sub}) for the substitution of the Gal4p dimer by the
388 Gal3p dimer as binding partner of the Gal80p dimer. A molecular dynamics simulation of the
389 Asp352 mutant (in a model system of the Gal3p*-Gal80 complex) indicates that two
390 positively charged amino acids, Gal3p-Arg362 and Gal80p-Lys287, are able to form direct
391 salt bridges with Asp352 (Fig. 6d). These attractive interactions of Asp352 with its
392 environment are, however, expected to be partially cancelled out by repulsive interactions
393 with the less proximate, negatively charged amino acids Gal3p-Glu363 and Gal80p-Glu348
394 (Fig. 6d). Also, the polar solution (water + counter ions) could partially reduce the
395 stabilization effect of the H352D mutation because residue 352 is better solvated in the
396 Gal3p* dimer than in the Gal3p*-Gal80p tetramer. Thus, to quantify a possible stabilization
397 effect of the H352D mutation, we computed the change in the Gibbs free energy difference,
398 $\Delta\Delta G_{sub} = \Delta G_{sub}^{D352} - \Delta G_{sub}^{H352}$, with the aid of the thermodynamic cycle depicted in Fig. 6e.
399 The actual free energy calculations (see Methods) yielded $\Delta\Delta G_{sub} = -2.8 \pm 0.9$ kcal/mol,
400 which indicates that the H352D mutation indeed increases the thermodynamic stability of the
401 Gal3p*-Gal80p complex with respect to the Gal4p-Gal80p complex. Thus, as predicted by the
402 dynamic model of network induction, the H352D mutation increases the cellular response by
403 facilitating the formation of the complex.
404

405 DISCUSSION

406

407 We experimentally monitored the induction dynamics of the yeast GAL network in the
408 context of natural genetic variation at the *GAL3* gene. We built a stochastic model of the
409 network and used it to link *GAL3* alleles to functional network parameters. This approach
410 discriminated alleles that increased the strength of activated Gal3p (*e.g.* of strains NCYC361
411 and DBVPG1788) from alleles that desensitized Gal3p to galactose activation (*e.g.* of strain
412 YJM978). Alleles showing different glucose/galactose trade-offs at equilibrium displayed
413 different dynamics of induction, and they were associated to different strength of activated
414 Gal3p. Our approach also predicted a functional effect of a single non-synonymous SNP that
415 was validated by atomistic simulations of the binding interface between Gal3p and Gal80p
416 dimers. These results provide further details on the yeast GAL system and, perhaps more
417 importantly, they constitute a proof-of-concept of the feasibility and usefulness of linking
418 genetic variants to model parameters.

419

420 **Genetic variability of the yeast GAL network**

421

422 Our *in vivo* and *in silico* analysis of the induction kinetics of yeast GAL activation
423 reveals properties of this system and how it is sensitive to genetic variation. Previously,
424 several computational models of the network have been proposed, usually in an effort to
425 understand the properties of the system at steady-state[11,15,18]. Particularly, they highlighted the
426 important role of Gal3p, Gal1p and Gal80p-mediated feedback loops. Our *in silico* analysis
427 suggests that the gradual or binary kinetic response is mainly controlled by the initial number
428 of repressors (Gal80p) and inducers (Gal1p and Gal3p), the efficacy of galactose to activate
429 the inducers and the efficiency of the activated inducers to release the effect of repressors. In

particular, a low mean number of inducers at the time of induction may lead to high cell-to-cell variability in their actual number. Cells with few inducers (as compared to repressors) display a lag time before responding, leading to a binary response pattern at the population level. This prediction from our model is fully consistent with recent experiments that tracked the induction of the network at the single-cell level and showed that the initial concentrations of Gal1p and Gal3p are predictive of the transient bimodal response[25]. We also observed that feedback loops were important to control the strength of cell-to-cell variability before induction (Gal80-mediated negative feedback) and the duration of lag times (Gal3/Gal1-mediated positive feedbacks), which agrees with the previous observation that disabling the Gal80p and Gal3p feedback loops can transform a gradual response into a binary one[14]. Our results on GAL3 genetic variants also complement previous genetic manipulations of the feedback loops, where their effect on bimodality was tested by modulating promoter activities[14,15,18,19]. Here, we showed that a non-synonymous variant affecting Gal3p:Gal80p interaction directly affects the dynamics of transient bimodality. This is a novel experimentally-based observation that is totally coherent with the conclusions of Venturelli *et al.* who showed computationally that steady-state bimodality of the network could rely on protein-protein binding affinities[18].

We also observed that genetic variation at *GAL3* could affect its propensity to be activated by galactose/ATP binding. In particular, the $GAL3^{YJM978}$ allele was associated with increased values of the K_{Gal} parameter (more galactose needed for its activation). This allele harboured 3 non-synonymous SNPs: M179I, R312I and H352D. As shown above, H352D is found in all natural alleles that we tested and it therefore does not explain a change in K_{Gal} specifically for $GAL3^{YJM978}$. According to the structure of the Gal3p:Gal80p tetramer complexed with galactose and ATP[32], the other two polymorphic sites do not map close to the

pocket containing the ligands. Met179 is located at the surface of the complex, distant from any binding interface and distant from the bound galactose (30 Å) and ATP (25 Å). The mutational effect of the rather conservative amino acid change (methionine to isoleucine) on the K_{Gal} parameter might therefore be negligible. In contrast, the non-conservative arginine to isoleucine mutation at site 312 could influence K_{Gal} in several ways: first, the positively charged arginine contributes favorably to the binding of the negatively charged ATP through long-range electrostatic interactions. The charge-neutral Ile312 variant lacks this favorable interaction and may have lower affinity for ATP, thereby penalizing activation by the two ligands. Second, residues Arg312 of the two Gal3p units are in direct contact with each other, and the non-conservative R312I change may affect the dimerization of Gal3p. Lavy *et al.*[32] reported that, in absence of galactose, Gal3p is monomeric in solution and adopts an open conformation that differs from the conformation generating the Gal3p:Gal3p dimeric interface found upon interaction with Gal80p. If the R312I modification alters Gal3p dimerization, this could modify the overall activation by galactose because these processes are coupled.

We observed that genetic variation of the strength of activated Gal3p (ρ_{Gal3}), estimated from the dynamic properties of network activation, was correlated with variation of the glucose/galactose trade-off at steady state. This implies that the two traits co-evolve in natural populations of *S. cerevisiae*. Given the relatively short time-scale of network induction, mild differences in the dynamics of activation alone are unlikely to cause fitness differences unless environmental galactose concentrations are highly dynamic. In contrast, variation in the sensitivity of the network to the ratio of external sugars corresponds to the triggering of an adaptive metabolic process, which is highly related to fitness even for slow environmental changes. The induction dynamics that we observed on short time-scales are

479 probably not themselves under selection, but they provide valuable information on the
480 molecular mechanism affecting a fitness-related trait operating on longer time scales.
481
482 The H352D variant is interesting in this regard. At this position in Gal3p, a histidine
483 residue was found in all laboratory strains (BY4741, CEN.PK, D273-10B, FL100, FY1679,
484 JK9-3d, SEY6210, W303, X2180-1A, YPH499), while nearly all natural isolates as well as
485 distant species possess an aspartic acid. Our results showed the importance of this aspartic
486 acid for interaction with Gal80p, which suggests that its conservation in wild population
487 results from purifying selection. The presence of slightly-deleterious mutations in laboratory
488 strains is well-known. Examples from the reference strain BY/S288c include mutations in
489 *AMN1* [37], *BUL2* [38], *ERC1* [39], *FLO8* [40], *GPA1* [37] and *HAP1* [41]. These mutations likely resulted
490 from a release of purifying selection caused by strong population bottlenecks when
491 propagating yeast on petri dishes. As for the genes listed above, the implication for *GAL3* is
492 that most mechanistic studies refer to a "Wild-Type" protein that is in fact a slightly-
493 hypomorphic allele not found in nature.
494
495 We also noted cases where the specificities of a *GAL3* allele in the context of the BY
496 strain did not reflect the properties of the donor strain. An extreme example of this was the
497 $GAL3^{DBVPG1853}$ allele which improved the response of the BY strain (Fig. 1d) while the
498 DBVPG1853 strain itself did not respond at all to galactose (not shown), presumably because
499 of genetic defects in other genes. Background-specific effects are common and should be
500 taken into account when interpreting the functional impact of natural alleles in their original
501 strain context[42].
502
503 **Linking DNA variants to model parameters: feasibility and potential**

We developed our approach using a model system, the yeast GAL network, which was an ideal context for investigation: molecular players were well known, important network properties had been previously described, genetic engineering could be used to study the effect of a single gene in an otherwise isogenic background, and experimental measurements were relatively cheap. If network modeling had provided no added value in such a context, it would be hard to imagine how it could be useful in more complex frameworks. We report that it did: observing different dynamics experimentally was not sufficient to make functional inferences, but combining data and modeling was. The concept is therefore fruitful and it is interesting now to consider how it can be extended to other biological systems.

First, it is important to realize that inference is based on the wealth of information contained in the dynamics of activation. Evidently, studying the system at equilibrium would not be sufficient. Mapping DNA variants to model parameters is therefore promising for systems where time-course data is available.

Second, even in the simple context of our study, not all parameters of the model were identifiable and it was necessary to aggregate several of them into a meta-parameter (ρ_{Gal3}). We admit that this constitutes a limit of the approach: when the H352D SNP was linked to this meta-parameter, additional assumptions were needed to infer biochemical effects. Similar difficulties will likely be encountered in other systems and the identifiability and sensitivity analysis of the model is therefore crucial to determine the nature of biological information that can be retrieved by the approach.

Third, our method here was to infer function and then to validate a prediction by exploring the structural data of a protein complex. Depending on the system under consideration and the data available, it may be judicious to reverse the approach: scanning

protein structures first in order to identify variants modifying binding affinities and then studying these variants specifically using experimental measurements and model-fitting. This way, a parameter change is first inferred from structural data and a dynamic model of the network then allows one to predict its phenotypic effect. The SAAP database[6], which registers structurally-relevant variants of human proteins, may constitute a very helpful resource to do this.

Fourth, while we based our approach on cell population distributions, tracking the response dynamics of individual cells over time is also possible[25] and can provide more information on the network response. In other contexts, such methods had been very useful to infer parameters associated to individual cells[43]. A variant may then be associated to one parameter by a whole distribution of values, which likely carries more information than a single scalar value as presented here.

Fifth, additional work is now needed to extend the approach to more than one gene. At the level of an entire network, the overall genotype of the individual is a combination of alleles. The number of such combinatorial genotypes of the network segregating in natural populations can be very large and mapping this diversity to the parameter space would be very interesting. In particular, models accounting for genetic changes might predict and explain genetic interactions (epistasis) within the network. The challenge to achieve this will likely reside in the number of free parameters: if the genotype is allowed to vary at too-many genes, parameters cannot be constrained efficiently. Mapping variants one gene at a time, as we did here, and then in combination would maintain this necessary constraint while evaluating epistasis. A more difficult task would be to infer the contribution of genes that are external to the network while nonetheless affecting its behavior (*e.g.* by modifying widely transcription rates or the stability of proteins, or cross-talks with other networks). Studying these factors by our approach is only possible after they are identified and connected to the

554 network. Their identification can be obtained by genetic mapping. For example, we recently
555 identified a locus on yeast chromosome V that affects the variability of the GAL response at
556 transient times of activation[28]. Once identified, these factors must be integrated in the network
557 model, which may be a complex task.
558
559 Network modeling is expected to help the development of personalized medicine and
560 the fact that it is possible, in a yeast system, to *personalize* model parameters according to
561 DNA variants is encouraging. Can the approach described here be applied to human variants?
562 This requires overcoming several difficulties that could be avoided in our framework. First,
563 most regulatory networks of human systems are incompletely known. Second, most of these
564 networks comprise numerous genes, implying many model parameters and, possibly, too-
565 many degrees of freedom for adjustments and identifiability issues. The first task is therefore
566 a careful identifiability and sensitivity analysis of the model and, as much as possible, a
567 reduction of its complexity. The work of Zhao *et al.*[44] is encouraging in this regard. The
568 authors studied the mitochondrial outer membrane permeabilization network controlling entry
569 in apoptosis. Their model comprised ~50 parameters and ~20 molecular species, but the
570 network critical behaviour (bifurcation point) was sensitive to less than half of the parameters.
571 The authors then searched for enrichment of cancer mutations in protein domains involved in
572 molecular interactions and they used molecular dynamics simulations to estimate the affinity
573 changes caused by these mutations. Interestingly, most mutations that were predicted to affect
574 sensitive parameters of the model caused a significant change of affinity in the expected
575 direction, illustrating that the model was able to highlight relevant vulnerabilities. Similarly,
576 Nijhout et al.[45] studied a model of the folate-mediated one carbon metabolism system. They
577 reported that human mutations that strongly perturb enzymatic activities could have little
578 phenotypic effect if they targeted parameters that are poorly sensitive. Another type of

579 difficulties when studying human networks are experimental limitations: manipulating human
580 cells needs more time and funds than manipulating yeast; replacing alleles of specific genes is
581 possible via CRISPR/Cas9 editing but the large physical size of human genes as well as the
582 functional redundancy between paralogues can be problematic; and setting up dynamic
583 experimental acquisitions is often not straightforward. Thus, applying our approach to a
584 minimal network in human cells compatible with genetic editing and time-series acquisitions
585 will probably constitute an important step in the near future.
586
587

588 METHODS

589

590 **Yeast strains and plasmids.**

591 The strains used is this study are listed in Table S3. We used the strain BY4711 (GY145,
592 isogenic to s288c) as BY reference strain. The $P_{GAL1}GFP$ reporter cassette was obtained from
593 plasmid pGY338 previously described[28]. pGY338 was linearized by NheI and integrated at
594 the *HIS3* locus of BY4711 to create strains GY1648 and GY1649, two independent
595 transformants. To replace endogenous $GAL3^{BY}$ allele by natural variants in GY1648 strain, we
596 PCR amplified the *TRP1-GAL3* locus of natural wild isolates using primers 1D28(5'-
597 AGAGGCGGTGGAGATATTCCTTATG-3') and 1D56(5'-
598 ACGTCCGCTATACCTTCGTTTTCTC-3'). The endogenous locus was then replaced by *in*
599 *vivo* homologous recombination and positive transformants were selected on SD-TRP plates.
600 $GAL3^{NCYC361}$, $GAL3^{K11}$, $GAL3^{Y12}$, $GAL3^{DBVPG1788}$, $GAL3^{DBVPG1853}$, $GAL3^{YJM978}$, $GAL3^{JAY291}$
601 were PCR amplified from NCYC3451, NCYC3452, NCYC3445, NCYC3311, NCYC3313,
602 NCYC3458 (wild isolates from the *Saccaromyces* Genome Resequensing Project, SGRP[26,46])
603 and JAY291 (Argueso *et al.*[47]), respectively. The strains used to characterize the effect of
604 natural variants on galactose response were GY1648, GY1689, GY1692, GY1695, GY1698,
605 GY1704, GY1707 and GY1713, all isogenic to S288c except for $GAL3^{BY}$, $GAL3^{NCYC361}$,
606 $GAL3^{K11}$, $GAL3^{Y12}$, $GAL3^{DBVPG1788}$, $GAL3^{DBVPG1853}$, $GAL3^{YJM978}$, $GAL3^{JAY291}$, respectively.
607 Strains genotype was verified by PCR and either high-resolution melting curves, restriction
608 fragment length polymorphism typing or sequencing. The *TRP1-GAL3* locus from BY strain
609 was PCR amplified with primers 1M95 (5'-
610 tctttcattatgtgagagtttaaaaaccagaaactacatcatcgaaaaagggatccAGAGGCGGTGGAGATATTCCT
611 TATG-3') and 1M96 (5'-
612 cgcccaatacgcaaaccgcctctccccgcgcgttggccgattcattaatgcagctgACGTCCGCTATACCTTCGTTT

613 TCTC-3') and cloned into HpaI-linearized plasmid pALREP[39] by homologous recombination
614 in yeast, generating plasmid pGY409. The mutated *GAL3*$^{BY\text{-}H352D}$ allele was synthesized by
615 GeneScript and subcloned into pGY409 using MscI-BstEII restriction sites, generating
616 plasmid pGY418. The *TRP1-GAL3*$^{BY\text{-}H352D}$ locus was PCR-amplified from pGY418 using
617 primers 1D28 and 1D56 and transformed into GY1649 to create strain GY2009. Genotype
618 was validated by PCR and sequencing. Strains of figure 5 were MPJ125-E06 (*GAL3*^{BY}),
619 MPJ143-H01 (*GAL3*^{YJM428}), MPJ143-F01 (*GAL3*^{YJM421}) and MPJ125-A07 (*GAL3*^{BC187}) which
620 were described in another study (Lee et al. PLoS Genetics, *in press*); they all derived from a
621 S288c *hoΔ::GAL1pr-YFP-mTagBFP2-kanMX4; gal3*Δ*::hphNT1* parental strain.
622
623 **Galactose response measurements.**
624 Liquid cultures in synthetic medium with 2% raffinose (Yeast Nitrogen Base w/o amino acids
625 6.7g/L, Raffinose 2%, Dropout Mix 2g/L, adjusted to pH=5.8) were inoculated with a single
626 colony and incubated overnight, then diluted to OD600 = 0.1 (synthetic medium, 2%
627 raffinose) and grown for 3 to 6 hours. The galactose induction experiments were carried out in
628 96-well sterile microplates using a Freedom EVO200 liquid handler (Tecan) equipped with a
629 96-channel pipetting head (MCA), a high precision 8-channel pipetting arm (LiHa), a robotic
630 manipulator arm (RoMa) and a MOI-6 incubator (Tecan). All robotic steps were programmed
631 in Evoware v2.5.4.0 (Tecan). Cells were resuspended in synthetic medium with 2% raffinose
632 and the appropriate galactose concentration (0.01, 0.1, 0.2 and 0.5%) and grown for the
633 desired time (from 0 to 250 minutes). Cells were then washed with PBS1X, incubated for 8
634 min in 2% paraformaldehyde (PFA) at room temperature, followed by 12 min of incubation in
635 PBS supplemented with Glycine 0.1M at room temperature and finally resuspended in PBS.
636 They were then analyzed on a FACSCalibur (BD Biosciences) flow cytometer to record

637 10,000 cells per sample. Each set of data is representative of the results of two independent
638 experiments (each comprising 3 technical replicates).
639 Flow cytometry data was analysed using the *flowCore* package from Bioconductor[48]. Cells of
640 homogeneous size were dynamically gated as follows: (i) removal of events with saturated
641 signals (FSC, SSC or FL1 = 1023 or = 0), (ii) correction by subtracting the mean(FL1) at t=0
642 to each FL1 values, (iii) computation of a density kernel of FSC, SSC values to define a
643 perimeter of peak density containing 60% of events, (iv) cell gating using this perimeter, (v)
644 removal of samples containing less than 3,000 cells at the end of the procedure and (vi)
645 correction of the data according to an eventual experimental bias during cytometer
646 acquisitions. For the twelve time-points (0, 10, 20, 30, 40, 60, 80, 100, 130, 160, 205 and 250
647 minutes) experimental design, the time course for a given strain was acquired on different
648 plates on the flow cytometer. In order to correct an eventual plate effect, we systematically
649 included 24 replicates on each plate acquired on flow cytometer. We then tested the fixed
650 effect of plates using an ANOVA. The FL1 values of each cell were subsequently corrected
651 according to the plate offset of the ANOVA. For the six time-points (0, 30, 60, 80, 130 and
652 210 minutes) experimental design, all the timepoints being acquired on the same experimental
653 plate, we did not apply the normalization filter. The GFP expression values presented here in
654 arbitrary units were the FL1 signal of the retained cells (normalized for the plate effect, if
655 required).
656 **Analysis of flow cytometry distributions.** All statistical analysis were done using R (version
657 3.2.4).
658 *Calculation of the response amplitude.* The response amplitude A was defined as the mean of
659 $P_{GAL1}GFP$ expression in activated cells. First, for each strain, at each time point, we
660 determined by eye if the $P_{GAL1}GFP$ distribution was unimodal ($f_{(X_{ALL})} = N(\mu_{ALL},\sigma_{ALL})$) or
661 bimodal ($f_{(X_{ALL})} = f_{(X_{OFF})} + f_{(X_{ON})}$). If the distribution was unimodal, we calculated:

A = μ_{ALL}. Otherwise, bimodal distributions were considered as mixtures of two normal distributions, such as: $f_{(X_{ALL})} = \rho_{OFF} N(\mu_{OFF,}\sigma_{OFF}) + \rho_{ON} N(\mu_{ON,}\sigma_{ON})$, with A = μ_{ON}. We used the function mixtools::normalmixEM() to calculate A for mixture distributions.

Calculation of inducibility. Inducibility was defined as the proportion of ON cells in the population. The threshold *t* between OFF and ON cells was calculated as follows: i) a subset of OFF cells (all cells acquired at t=0min) and ON cells (activated cells belonging to unimodal distributions, acquired at the latest time point of the experiments) was defined for each experiments , ii) the mean and standard deviation were extracted from each OFF and ON normal distributions using the function `mixtools::normalmixEM()`, iii) these parameters were used to determine *t* such as $\mathbb{P}(X_{ON} < t) = \mathbb{P}(X_{OFF} > t)$, with X_{ON} the observed fluorescence FL1 in ON_cells and X_{OFF} the observed fluorescence FL1 in OFF_cells, iv) we finally calculated I = $\frac{nb\_cells(FL1>t)}{nb\_cell(total)}$ for each time point, for each strain.

Stochastic modeling

We model the stochastic gene expression of *GAL1, GAL3, GAL80* and of the reporter gene (under a *GAL1* promoter). For each gene we account for the status of the promoter (ON/OFF) and for the production and degradation of mRNAs and proteins. In addition, for the reporter gene, we account for the maturation of the fluorescent protein. The promoter switching rate from ON to OFF for gene *i* is driven by GAL80p: $k_i^{off} = k_o^{off}\left[\left(\frac{Gal80p}{K_{80}}\right)^2\right]^{n_i}$ with n_i the number of strong GAL4p binding sites in the promoter. We assume that GAL80p represses transcription via its dimerized form (with K_{80} encompassing the dimer dissociation constant). The promoter switching rate from OFF to ON is driven by GAL3p and Gal1p: $k_i^{on} = k_o^{on}\left[\left(\frac{Gal1p^*}{K_1}\right)^2 + \left(\frac{Gal3p^*}{K_3}\right)^2\right]^{n_i}$ with $Galp^* = Galp\left(\frac{[gal]/K_{gal}}{1+[gal]/K_{gal}}\right)$ the number of activated proteins at a given galactose concentration [gal] (K_{gal} being the galactose dissociation constant). Here also, we assume that activated Gal3p and Gal1p are mainly found as dimers.

686 K_1 and K_3 encompass the dimer dissociation constants as well as the affinity of activated
687 Gal3p* for Gal80p. For a detailed description of the model see Supplementary Text 1. Most
688 of the parameters of the model (except K_1, K_3, K_{80} and K_{gal}) were fixed based on the literature
689 (see Table S1 in Supplementary Text 1). The model had 7 GAL3-dependent parameters: α_3
690 (leaky transcription rate), γ_3 (translation rate), β_3 (mRNA degradation rate), μ_3 (protein
691 degradation/dilution rate), $\Delta\alpha_3$ (full transcription rate), K_3 and K_{gal}. The phenotypic response
692 of a strain (gradual vs binary) at a given galactose concentration mainly depends on K_{gal} and
693 on the strength of GAL3 defined by $\rho_{Gal3}=\alpha_3\gamma_3/(\beta_3\mu_3K_3)$ (see main text and Supplementary
694 Text 1). For a given set of parameters, the stochastic dynamics of galactose induction was
695 simulated using the stochastic simulation algorithm from Gillespie[33]. The system was first
696 allowed to reach steady-state at [gal]=0. At t=0, galactose is introduced and the parallel -
697 independent - evolution of 5,000 cells is monitored during 250 minutes of real time.
698
699 **Parameter inference**
700 For a fixed set of GAL3-independent parameters, predictions for various values of GAL3-
701 dependent parameters ρ_{Gal3} and K_{gal} were performed at 3 different galactose concentrations
702 (0.05, 0.1 and 0.5%). Parameters were sampled from a 2D logarithmic-grid encompassing the
703 region of interest. Then, for each strain, a global chi2-score between the experimental data
704 and the corresponding model predictions integrating the 3 concentrations were minimized to
705 infer ρ_{Gal3} and K_{gal}. Uncertainties on the parameters reflect the size of the sampling parameter
706 grid. Parameter inference was repeated 6 times for different values of GAL3-independent
707 parameters (see Supplementary Text 1).
708
709 **Molecular dynamics simulations** for free energy calculations were carried out as described
710 in Supplementary Text 2 and Supplementary Fig. 8.

711

712 **Data availability.** All flow cytometry raw data files can be downloaded from

713 http://flowrepository.org under accession number FR-FCM-ZY6Y.

714

715

716

REFERENCES

1. Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13,** R7 (2012).
2. Bauer-Mehren, A., Furlong, L. I., Rautschka, M. & Sanz, F. From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. *BMC Bioinformatics* **10,** S6 (2009).
3. Levenstien, M. A. & Klein, R. J. Predicting functionally important SNP classes based on negative selection. *BMC Bioinformatics* **12,** 26 (2011).
4. Coronnello, C. *et al.* Novel Modeling of Combinatorial miRNA Targeting Identifies SNP with Potential Role in Bone Density. *PLOS Comput. Biol.* **8,** e1002830 (2012).
5. Barenboim, M., Masso, M., Vaisman, I. I. & Jamison, D. C. Statistical geometry based prediction of nonsynonymous SNP functional effects using random forest and neuro-fuzzy classifiers. *Proteins Struct. Funct. Bioinforma.* **71,** 1930–1939 (2008).
6. Al-Numair, N. S. & Martin, A. C. The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC Genomics* **14,** S4 (2013).
7. Sellick, C. A., Campbell, R. N. & Reece, R. J. Galactose metabolism in yeast-structure and regulation of the leloir pathway enzymes and the genes encoding them. *Int. Rev. Cell Mol. Biol.* **269,** 111–150 (2008).
8. Bhat, P. J. & Hopper, J. E. Overproduction of the GAL1 or GAL3 protein causes galactose-independent activation of the GAL4 protein: evidence for a new model of induction for the yeast GAL/MEL regulon. *Mol. Cell. Biol.* **12,** 2701–2707 (1992).
9. Hawkins, K. M. & Smolke, C. D. The Regulatory Roles of the Galactose Permease and Kinase in the Induction Response of the GAL Network in Saccharomyces cerevisiae. *J. Biol. Chem.* **281,** 13485–13492 (2006).
10. Biggar, S. R. & Crabtree, G. R. Cell signaling can direct either binary or graded transcriptional responses. *EMBO J.* **20,** 3167–3176 (2001).
11. Apostu, R. & Mackey, M. C. Mathematical model of GAL regulon dynamics in Saccharomyces cerevisiae. *J Theor Biol* **293,** 219–35 (2012).
12. Song, C. *et al.* Estimating the stochastic bifurcation structure of cellular networks. *PLoS Comput Biol* **6,** e1000699 (2010).
13. Hsu, C. *et al.* Stochastic signalling rewires the interaction map of a multiple feedback network during yeast evolution. *Nat. Commun.* **3,** 682 (2012).
14. Ramsey, S. A. *et al.* Dual feedback loops in the GAL regulon suppress cellular heterogeneity in yeast. *Nat Genet* **38,** 1082–7 (2006).
15. Acar, M., Becskei, A. & van Oudenaarden, A. Enhancement of cellular memory by reducing stochastic transitions. *Nature* **435,** 228–32 (2005).
16. Kundu, S. & Peterson, C. L. Dominant Role for Signal Transduction in the Transcriptional Memory of Yeast GAL Genes. *Mol. Cell. Biol.* **30,** 2330–2340 (2010).
17. Becskei, A., Seraphin, B. & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *Embo J* **20,** 2528–35 (2001).
18. Venturelli, O. S., El-Samad, H. & Murray, R. M. Synergistic dual positive feedback loops established by molecular sequestration generate robust bimodal response. *Proc. Natl. Acad. Sci.* **109,** E3324–E3333 (2012).
19. Peng, W., Liu, P., Xue, Y. & Acar, M. Evolution of gene network activity by tuning the strength of negative-feedback regulation. *Nat. Commun.* **6,** 6226 (2015).
20. Venturelli, O. S., Zuleta, I., Murray, R. M. & El-Samad, H. Population Diversification in a Yeast Metabolic Program Promotes Anticipation of Environmental Shifts. *PLoS Biol* **13,** e1002042 (2015).

766 21. New, A. M. *et al.* Different Levels of Catabolite Repression Optimize Growth in
767 Stable and Variable Environments. *PLoS Biol* **12,** e1001764 (2014).
768 22. Wang, J. *et al.* Natural Variation in Preparation for Nutrient Depletion Reveals a
769 Cost–Benefit Tradeoff. *PLoS Biol* **13,** e1002041 (2015).
770 23. Jiang, F., Frey, B. R., Evans, M. L., Friel, J. C. & Hopper, J. E. Gene Activation by
771 Dissociation of an Inhibitor from a Transcriptional Activation Domain. *Mol. Cell. Biol.* **29,**
772 5604–5610 (2009).
773 24. Egriboz, O., Jiang, F. & Hopper, J. E. Rapid GAL Gene Switch of Saccharomyces
774 cerevisiae Depends on Nuclear Gal3, Not Nucleocytoplasmic Trafficking of Gal3 and Gal80.
775 *Genetics* **189,** 825–836 (2011).
776 25. Stockwell, S. R. & Rifkin, S. A. A living vector field reveals constraints on galactose
777 network induction in yeast. *Mol. Syst. Biol.* **13,** 908 (2017).
778 26. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458,** 337–41
779 (2009).
780 27. Mateus, C. & Avery, S. V. Destabilized green fluorescent protein for monitoring
781 dynamic changes in yeast gene expression with flow cytometry. *Yeast* **16,** 1313–23 (2000).
782 28. Chuffart, F. *et al.* Exploiting Single-Cell Quantitative Data to Map Genetic Variants
783 Having Probabilistic Effects. *PLoS Genet.* **12,** e1006213 (2016).
784 29. Hittinger, C. T. & Carroll, S. B. Gene duplication and the adaptive evolution of a
785 classic genetic switch. *Nature* **449,** 677–681 (2007).
786 30. Peng, G. & Hopper, J. E. Gene activation by interaction of an inhibitor with a
787 cytoplasmic signaling protein. *Proc. Natl. Acad. Sci.* **99,** 8548–8553 (2002).
788 31. Pilauri, V., Bewley, M., Diep, C. & Hopper, J. Gal80 Dimerization and the Yeast
789 GAL Gene Switch. *Genetics* **169,** 1903–1914 (2005).
790 32. Lavy, T., Kumar, P. R., He, H. & Joshua-Tor, L. The Gal3p transducer of the GAL
791 regulon interacts with the Gal80p repressor in its ligand-induced closed conformation. *Genes*
792 *Dev.* **26,** 294–303 (2012).
793 33. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys.*
794 *Chem.* **81,** 2340–2361 (1977).
795 34. Escalante-Chong, R. *et al.* Galactose metabolic genes in yeast respond to a ratio of
796 galactose and glucose. *Proc. Natl. Acad. Sci.* **112,** 1636–1641 (2015).
797 35. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of
798 budding yeast. *Nucleic Acids Res.* **40,** D700–D705 (2012).
799 36. Thoden, J. B., Ryan, L. A., Reece, R. J. & Holden, H. M. The Interaction between an
800 Acidic Transcriptional Activator and Its Inhibitor THE MOLECULAR BASIS OF Gal4p
801 RECOGNITION BY Gal80p. *J. Biol. Chem.* **283,** 30266–30272 (2008).
802 37. Yvert, G. *et al.* Trans-acting regulatory variation in Saccharomyces cerevisiae and the
803 role of transcription factors. *Nat Genet* **35,** 57–64 (2003).
804 38. Kwan, E. X., Foss, E., Kruglyak, L. & Bedalov, A. Natural Polymorphism in BUL2
805 Links Cellular Amino Acid Availability with Chronological Aging and Telomere
806 Maintenance in Yeast. *PLoS Genet* **7,** e1002250 (2011).
807 39. Fehrmann, S. *et al.* Natural sequence variants of yeast environmental sensors confer
808 cell-to-cell expression variability. *Mol. Syst. Biol.* **9,** 695 (2013).
809 40. Kron, S. J. Filamentous growth in budding yeast. *Trends Microbiol.* **5,** 450–454
810 (1997).
811 41. Gaisne, M., Bécam, A. M., Verdière, J. & Herbert, C. J. A 'natural' mutation in
812 Saccharomyces cerevisiae strains derived from S288c affects the complex regulatory gene
813 HAP1 (CYP1). *Curr. Genet.* **36,** 195–200 (1999).
814 42. Gerke, J., Lorenz, K., Ramnarine, S. & Cohen, B. Gene-environment interactions at
815 nucleotide resolution. *PLoS Genet* **6,** (2010).

816 43. Llamosi, A. *et al.* What Population Reveals about Individual Cell Identity: Single-Cell
817 Parameter Estimation of Models of Gene Expression in Yeast. *PLOS Comput Biol* **12,**
818 e1004706 (2016).
819 44. Zhao, L., Sun, T., Pei, J. & Ouyang, Q. Mutation-induced protein interaction kinetics
820 changes affect apoptotic network dynamic properties and facilitate oncogenesis. *Proc. Natl.*
821 *Acad. Sci.* **112,** E4046–E4054 (2015).
822 45. Nijhout, H. F., Best, J. A. & Reed, M. C. Using mathematical models to understand
823 metabolism, genes, and disease. *BMC Biol.* **13,** 79 (2015).
824 46. Louis, E. & Durbin, R. The Saccharomyces Genome Resequencing Project;
825 http://www.sanger.ac.uk/Teams/Team71/durbin/sgrp/. (2007).
826 47. Argueso, J. L. *et al.* Genome structure of a Saccharomyces cerevisiae strain widely
827 used in bioethanol production. *Genome Res* **19,** 2258–70 (2009).
828 48. Hahne, F. *et al.* flowCore: a Bioconductor package for high throughput flow
829 cytometry. *BMC Bioinformatics* **10,** 106 (2009).
830

831

832

833 ACKNOWLEDGEMENTS

834

835 We thank Orsolya Symmons for discussions, Olivier Gandrillon for critical reading of
836 the manuscript, Nicolas Rochette for sequence analysis, Audrey Barthelaix for initial tests on
837 the robotic platform, Sandrine Mouradian and SFR Biosciences Gerland-Lyon Sud
838 (UMS344/US8) for access to flow cytometers and technical assistance, the Pôle Scientifique
839 de Modélisation Numérique and the CIMENT infrastructure (supported by the Rhône-Alpes
840 region, Grant CPER07_13 CIRA) for computing resources, BioSyL Federation and Ecofect
841 Labex for inspiring scientific events, developers of R, bioconductor and Ubuntu for their
842 software. This work was supported by the European Research Council under the European
843 Union's Seventh Framework Programme FP7/2007-2013 Grant Agreement n°281359. DJ
844 acknowledges program AGIR of University Grenoble-Alpes and the Institut Rhône-Alpin des
845 Systèmes Complexes for funding. M. Spr. was supported by grants NSF 1349248 and RO1
846 GM120122-01 from the National Science Foundation (USA).

847

AUTHORS CONTRIBUTION

Performed experiments: M.R., H.D-B., E.F., M.E., A.B.

Contributed analysis tools: M.R., F.C., D.J., and G.Y.

Contributed reagents: M.Spr.

Developed and evaluated pilot versions of the model: F.C. and F.P.

Conceived, implemented and used the model: D.J.

Performed molecular dynamics simulations: M.Spi.

Interpreted results: M.R., M.Spr., M.Spi., D.J. and G.Y.

Conceived and designed the study: G.Y.

Wrote the paper: M.R., D.J. and G.Y.

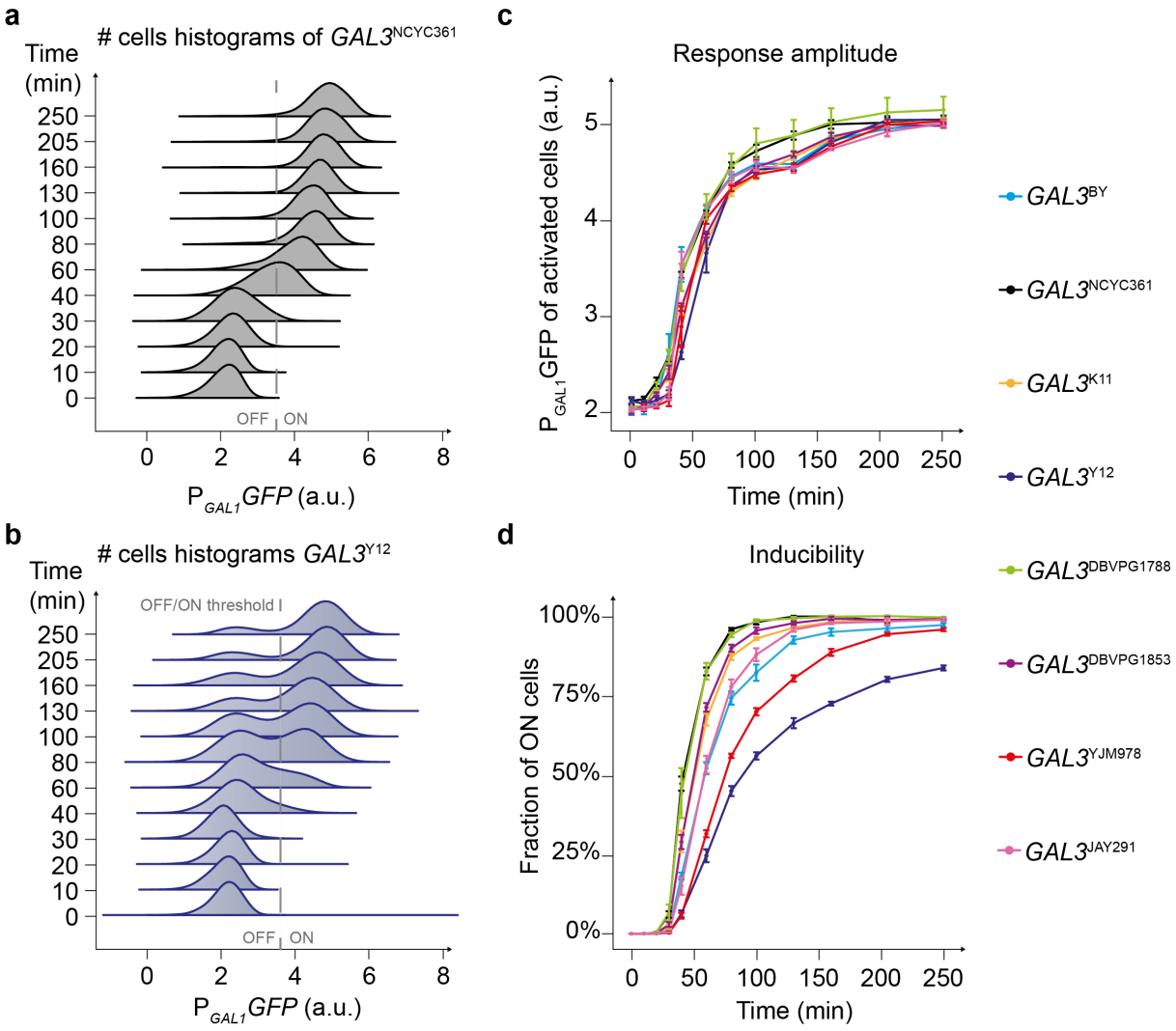

FIGURE 1

Figure 1. **Dynamic response to galactose in the context of *GAL3* variants.** Acquisitions were made on strains where the *GAL3* allele was replaced by the indicated natural alleles. These strains were otherwise isogenic, with a BY background. (**a-b**) Flow-cytometry data obtained on strains harboring the $GAL3^{NCYC361}$ allele (**a**) or the $GAL3^{Y12}$ allele (**b**). Cells were cultured in raffinose 2% and induced at time 0 by adding galactose at a final concentration of 0.5%. a. u., arbitrary units. Grey dashed line, threshold used to distinguish ON cells from OFF cells. (**c**) Amplitude of the response (mean expression) as a function of time for each *GAL3* replacements strain. Error bars represent standard error of the mean (n=6). (**d**) Inducibility of the response (fraction of ON cells) as a function of time for each *GAL3* replacement strain. Error bars represent standard error of the mean (n=6).

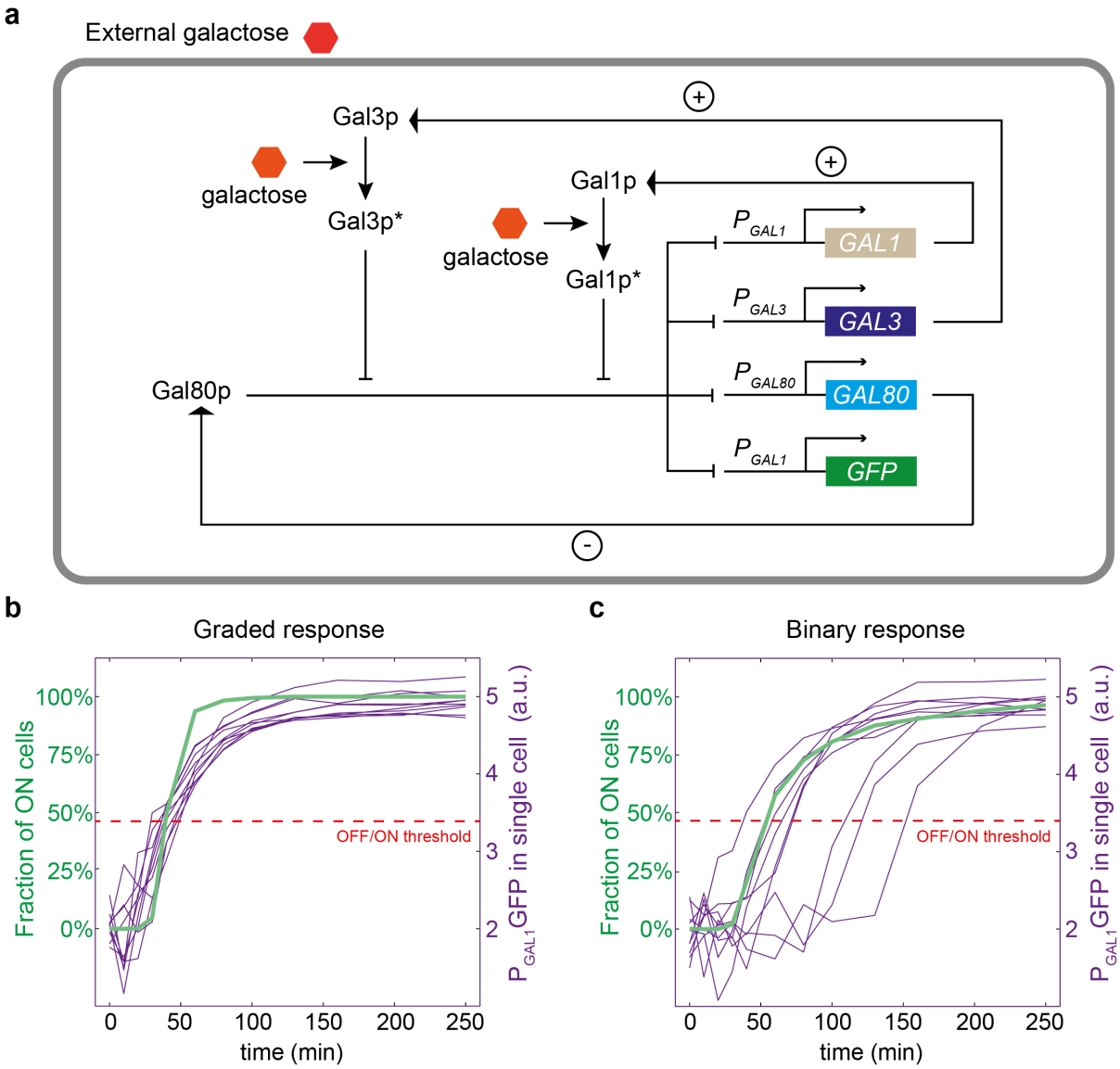

FIGURE 2

Figure 2. ***In-silico* model of network induction.** (**a**) Schematic representation of the model used. Galactose-activated Gal1p and Gal3p proteins become Gal1p* and Gal3p*, respectively. Pointed and blunt arrows represent activation and inhibition, respectively. Positive and negative feedback loops are highlighted by + and - signs. (**b**) Example of a gradual response predicted by the model ([gal]=0.5%, ρ_{Gal3}=140 and K_{Gal}=0.055). Thin violet lines represent stochastic simulations of network activation in individual cells. Dashed red line represents the threshold distinguishing ON from OFF cells. Green thick line indicates the fraction of ON cells as a function time. (**c**) Example of a binary response predicted by the model ([gal]=0.5%, ρ_{Gal3}=40 and K_{Gal}=0.055). Same color code as in (**b**).

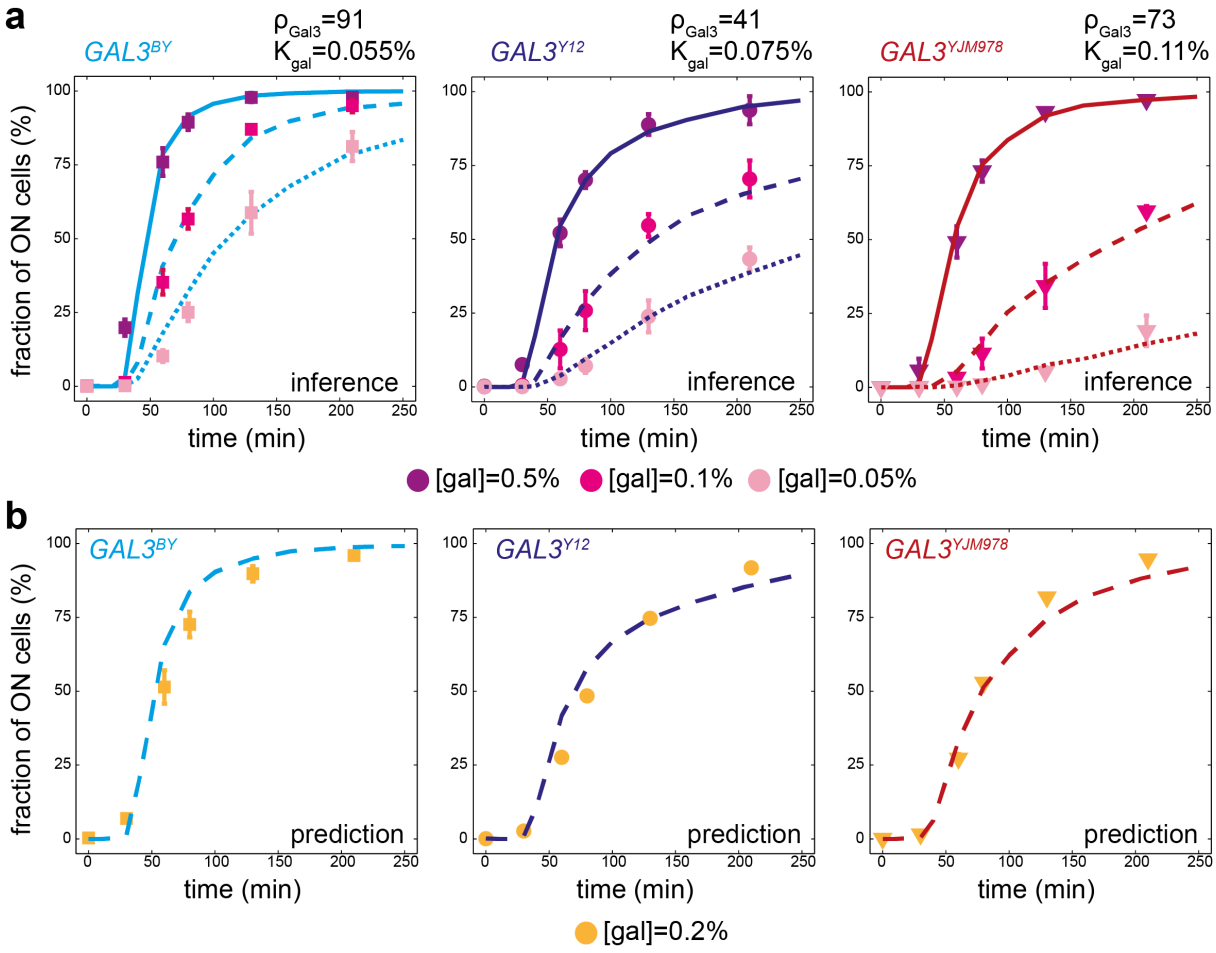

FIGURE 3

Figure 3. **Strain-specific training of the model and validation.**(**a**) Model fitting. Each panel corresponds to one strain carrying the indicated *GAL3* allele. Inducibility was measured by flow cytometry (data points +/- s.e.m.) after stimulating cells with three different concentrations of galactose (points colored according to the concentration). For each strain, this data was used to fit the GAL3-dependent parameters ρ_{Gal3} and K_{Gal}. Inferred parameter values are shown. Lines in plain (resp. dashed and dotted) represent the inducibility predicted by the model at [gal]=0.5% (resp. 0.1% and 0.05%). (**b**)With the parameters inferred in (**a**) we use the model to predict the inducibility of each strain at a galactose concentration of 0.2% (lines), and this prediction was compared to experimental measures (dots +/- s.e.m.).

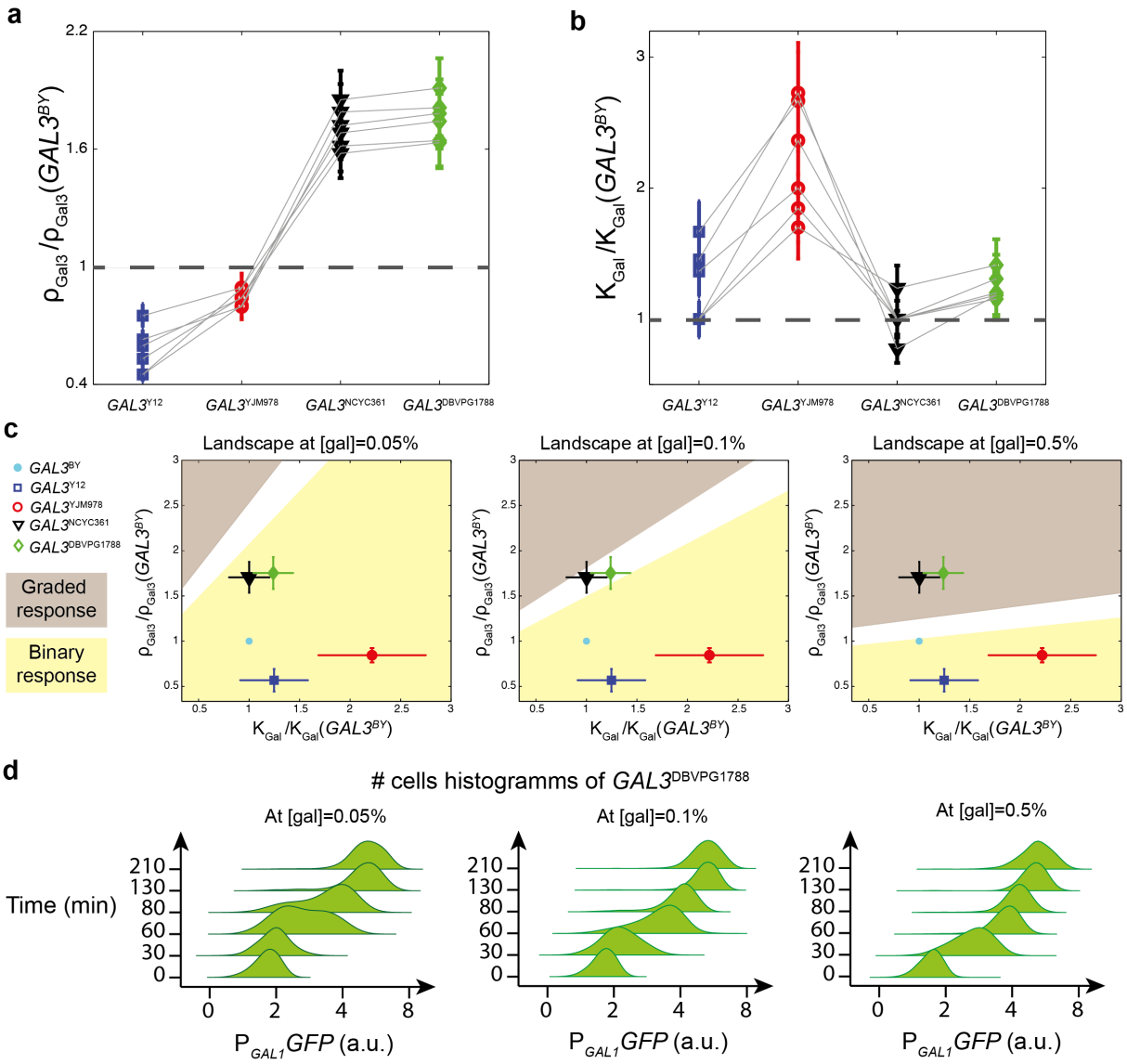

FIGURE 4

Figure 4. ***GAL3* alleles map to distinct locations of the parameter space. (a-b)** Parameter values obtained by fitting the model to experimental data collected on five strains at three concentrations of the inducer ([gal] = 0.05, 0.1 and 0.5%). Six independent fits were performed (one per grey line). For each one, different values of GAL3-independent parameters were chosen (see Supplementary Text 1), and parameters ρ_{Gal3} (**a**) and K_{gal} (**b**) were estimated for each strain. Dots represent their value for the indicated strain, relative to the value estimated for the $GAL3^{BY}$ strain. Error bars: uncertainty on parameter estimation for each inference (see Materials and Methods). (**c**) Phenotypic landscape predicted by the model. At defined concentrations of the inducer ([gal]), the values of ρ_{Gal3} and K_{gal} determine whether the response is gradual (brown) or binary (yellow). The white zone is an intermediate region where the distinction between gradual and binary is unclear. Using parameters inferred in (**a**) and (**b**), alleles are mapped to the landscape (colored dots). Error bars: standard deviation of the 6 distinct estimations. (**d**) Time-course flow cytometry data of the $GAL3^{DBVPG1788}$ strain, showing its transient binary response at low concentration of inducer (left) and its gradual response at higher concentration (right).

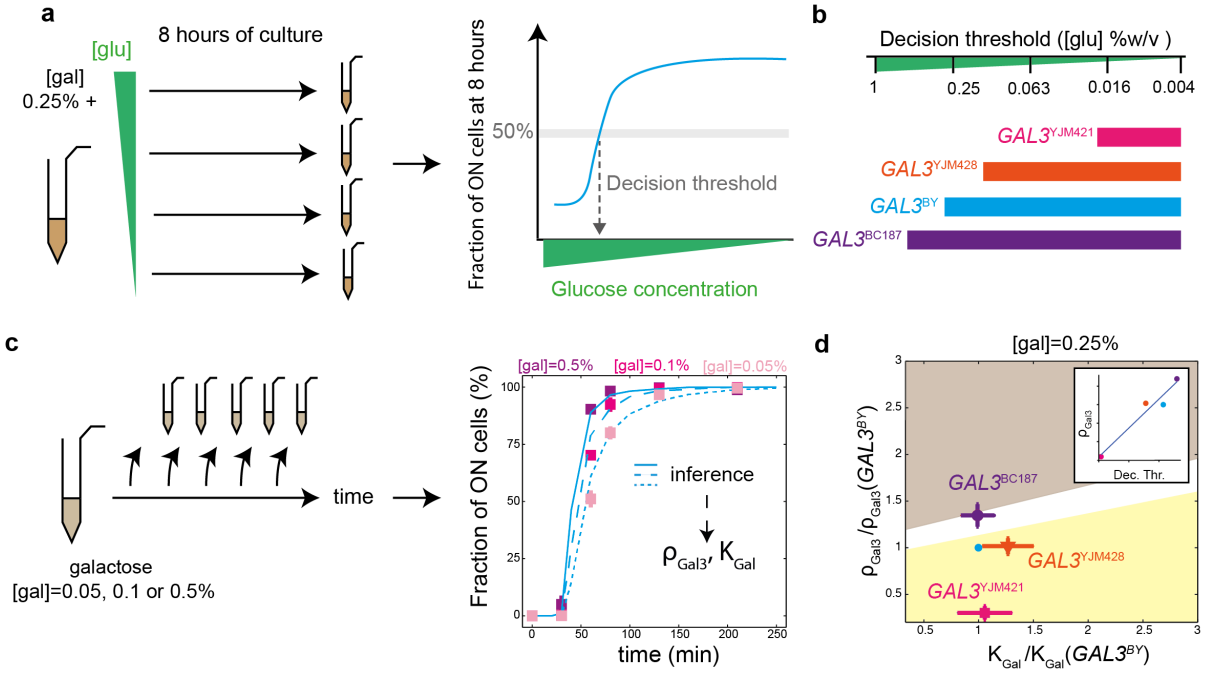

FIGURE 5

Figure 5. **Relationship between inducibility and diauxic shift decision threshold.** (**a**) Schematic representation of decision threshold measurement. The decision threshold corresponds to the concentration of glucose at which 50% of the cells are induced in the presence of 0.25% galactose. The blue curve is theoretical and shown to explain how the fraction of ON cells depends on glucose concentration. (**b**) Decision thresholds for strains $GAL3^{BY}$, $GAL3^{YJM421}$, $GAL3^{YJM428}$ and $GAL3^{BC187}$ at [gal] = 0.25%. (**c**) Schematic representation of *GAL3* induction parameters determination. (**d**) Location of the *GAL3* replacement strains in the phenotypic landscape of the model at [gal] = 0.25%. Inset: ρ_{Gal3} values as a function of the decision threshold, with dots corresponding to strains.

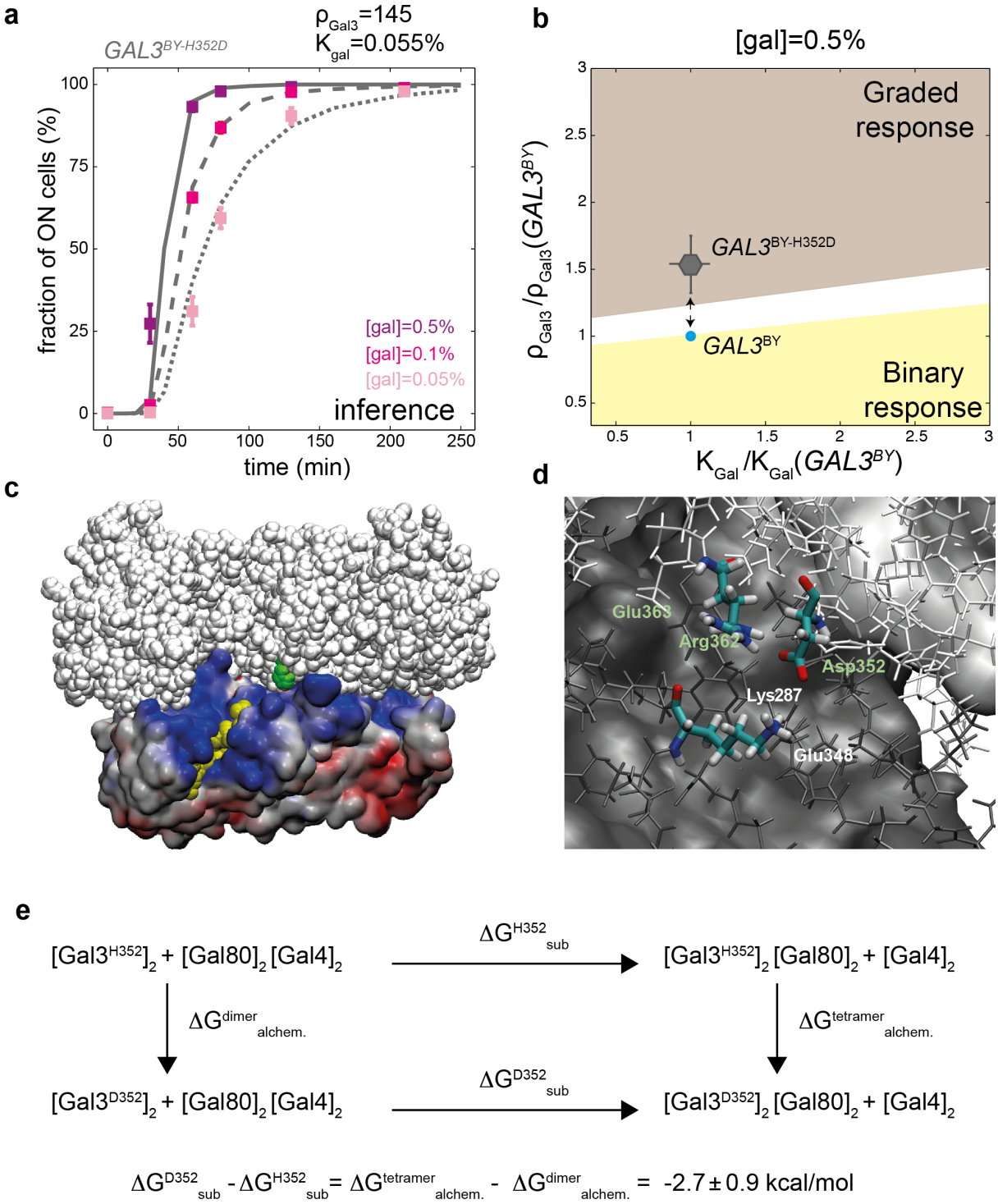

FIGURE 6

Figure 6. Functional inference of the H352D variant of *GAL3*. (**a**). Experimental acquisitions (dots) and model fitting (curves) of the induction dynamics of the $GAL3^{BY\text{-}H352D}$ strain. (**b**) $GAL3^{BY}$ (blue dot) and $GAL3^{BY\text{-}H352D}$ (grey dot with standard deviation bars) strains localisation in the phenotypic landscape of the model at [gal]=0.5%. Arrows: phenotypic trajectory between the two alleles. (**c**) Structure of the tetrameric complex

[Gal3p*]$_2$[Gal80p]$_2$ (PDB entry 3V2U). Residue His352 of one Gal3p unit is in the back side and not visible. The His352 residue of the other Gal3p unit is shown as green beads in the center; it is located at the binding interface of the Gal3p* dimer (white beads) and the Gal80p dimer (colored surface). Gal80p residues are colored according to their electrostatic surface potential from red (≤ -10 kT/e) to blue (≥+10 kT/e). Yellow beads: the acidic activation domain of Gal4p was inserted in the complex by superimposition with crystal structure 3BTS. A similar insertion in the other Gal80p unit is in the back side and not visible. Created with VMD software. (**d**) Local stabilization of Gal3p-Asp352 by residues Gal3p-Arg362 and Gal80p-Lys287 in the [Gal3]$_2$[Gal80]$_2$ complex. Green and white labels refer to residues from Gal3p and Gal80p units, respectively. The figure shows a snapshot from a molecular dynamics simulation of the mutation H352D carried out for a model system of the complex (see Supplementary Text 2). Atoms within 15 Å of residue 352 are shown as thin sticks in white (Gal3p) or dark grey (Gal80p). Remaining atoms are shown as a solid surface. Created with VMD software. (**e**) Thermodynamic cycle quantifying the energetic impact of the H352D mutation on the substitution of [Gal4p]$_2$ by [Gal3p]$_2$ as binding partner of [Gal80p]$_2$ (ΔG_{sub} , horizontal arrows). This impact is measured as $\Delta\Delta G = \Delta G_{sub}^{D352} - \Delta G_{sub}^{H352}$, which equals to $\Delta G_{alchem}^{tetramer} - \Delta G_{alchem}^{dimer}$ (vertical arrows) because free enthalpy is a state function. These latter quantities correspond to the free enthalpy change for the alchemical (double) mutation of His>Asp in the Gal3p-Gal80p tetramer and in the Gal3p dimer, respectively, which were computed by alchemical free energy calculations (see Supplementary Text 2).

Supplementary Figure 1. Sequences of natural *GAL3* locus used in this study. (**a**) Nucleotidic sequences of *GAL3* promoters. (**b**) Amino-acid sequences of Gal3p proteins. Alignment was performed using T-Coffee and visualized using Boxshade.

Supplementary Figure 2. **The lagging time in binary system depends on the initial activation force.**
For the same parameters as in Fig.2c, from each single-cell trajectory, we estimate the lagging time before single-cell reaches the threshold distinguishing ON from OFF cells (dashed line in Fig.2c). For 1000 simulated trajectories, we plot it as a function of the initial activation force defined as the value of the OFF to ON switching rate of GAL promoters k_i^{on} just at the moment of induction. This parameter depends on the initial number of Gal1p and Gal3p and on the effective constants K_1 and K_3 (see Materials and Methods and Supplementary Text 1). The Spearman correlation between lagging time and initial activation force is -0.75.

Supplementary Figure 3. Effect of ρ_{Gal1} , ρ_{Gal3} and ρ_{Gal80} values on network inducibility. Each panel shows the induction of the network as a function of time for different values of ρ_{Gal3} (colored lines) in a specific context of ρ_{Gal1} and ρ_{Gal80} values. Galactose concentration and K_{gal} were fixed to $[gal]$ = 0.5% and K_{gal} = 0.055%.

Supplementary Figure 4. Inducibility predictions depend on ρ_{Gal1} , ρ_{Gal3} and ρ_{Gal80} meta-parameters rather than on their constituent parameters. Each panel represents model predictions of inducibility as a function of time after induction at the indicated galactose concentration. Colors correspond to different sets of parameter values in the model, blue (reference) referring to values of Table S1 completed with K_1=0.35, K_3=1.26, K_{80}=1.03 and K_{gal}=0.055%. (**a**) Parameters constituting ρ_{Gal1} (formula $\rho_{Gal1}=\alpha_1\gamma_1/(\beta_1\mu_1K_1)$) were changed in a way that kept ρ_{Gal1} invariant. For example, K_1 was divided by 2 and β_1 was doubled (red). (**b-c**) Same analysis but where constituents of ρ_{Gal3} (b) or ρ_{Gal80} (c) were changed (similar formula). (**d**) Same analysis as in b but at lower galactose concentration. All simulations were run with ρ_{Gal1}=100, ρ_{Gal3}=100 and ρ_{Ga80}=250.

Supplementary Figure 5. The network behaviour depends on galactose concentration and on two model parameters. Predictions of the model for the inducibility as a function of time at 5 different galactose concentrations for different values of the GAL3-dependent parameters ρ_{Gal3} and K_{Gal}. GAL3-independent parameters were fixed (Table S1) with ρ_{Gal1}=100 and ρ_{Gal80}=250 (see main text and Supplementary Text 1 for parameter definitions).

Supplementary Figure 6. **Inference of GAL3-dependent model parameters for *GAL3*$^{\mathbf{NCYC361}}$ and *GAL3*$^{\mathbf{DBVPG1788}}$.** Experimentally-measured inducibility of *GAL3*$^{\mathrm{NCYC361}}$ and *GAL3*$^{\mathrm{DBVPG1788}}$ strains, as a function of time, at 3 different galactose concentrations (symbols coloured according to the concentration). These data were used to fit the GAL3-dependent parameters ρ_{Gal3} and K_{Gal}. Full lines (resp. dashed and dotted lines) represent the behaviours predicted by the model at [gal]=0.5% (resp. 0.1% and 0.05%).

Supplementary Figure 7. **Inference of GAL3-dependent model parameters for alleles tested in diauxic shift experiments.** Experimentally-measured inducibility of *GAL3*$^{\mathrm{BY}}$, *GAL3*$^{\mathrm{YJM421}}$, *GAL3*$^{\mathrm{YJM428}}$ and *GAL3*$^{\mathrm{BC187}}$ strains, as a function of time, at 3 different galactose concentrations (symbols coloured according to the concentration). These data were used to fit the GAL3-dependent parameters ρ_{Gal3} and K_{Gal}. Full lines (resp. dashed and dotted lines) represent the behaviours predicted by the model at [gal]=0.5% (resp. 0.1% and 0.05%).

Supplementary Figure 8. **Alchemical free energy calculations. A**) Model system for the Gal3p*-Gal80p tetramer used for the alchemical free energy calculations. Residue 352 (of Gal3p*) in the center is shown in colored beads (only one of two possible residues is seen in the chosen orientation). Residues within 15 Å of the two residues 352 were allowed to move freely; they are shown as thin sticks in white (Gal3p*) or grey (Gal80p). Residues that were harmonically restrained are shown as solid surface. This protein substructure was solvated in a cubic box (side length = 90 Å) of water molecules (transparent blueish cube) and salt (blue and yellow dots) with an ionic strength of 0.15 M. **B**) Block analysis of the alchemical free energy calculations. $\Delta G_{alchemical}$ corresponds to the free energy change for transforming 2 x Gal3p*-Asp352 to 2 x Gal3p*-His352. (*i.e.*, changing the coupling parameter λ from 1 to 0, see Supplementary Text 2). $\Delta G_{alchemical}$ for the Gal3p* dimer (grey squares) and the Gal3p*-Gal80p tetramer (black circles) is plotted for consecutive blocks of 100 ps of sampling. The horizontal dashed lines indicate the mean values.

Table S1. Description and values of model parameters used in this study (in Supplementary Text 1)

Table S2. Growth rates of *GAL3* allele-replacement strains (in Supplementary Text 1)

Table S3. Strains used in this study.

Supplementary Text 1. Model description and analysis

Supplementary Text 2. Methods for molecular dynamics simulations