Predicting an epistasis-rich genotype-phenotype map with a coarse-grained bottom-up model of budding yeast polarity

A proof-of-concept for phenotype prediction in model organisms

Werner Karl-Gustav Daalman[1], Liedewij Laan[1*]

[1] Department of Bionanoscience, TU Delft, Delft, the Netherlands

[*] Corresponding author

E-mail: L.Laan@tudelft.nl

Abstract

Accurate phenotype prediction based on genotypical information has numerous societal applications, such as design of useful crops of cellular factories. However, the prevalence of epistasis, a phenomenon that prevents many biological systems to perform in accordance with the sum of its parts, necessitates modelling the complex path between genotype and phenotype. Defining intermediate levels in this path reduces the complexity of prediction, and may also elucidate the phenotype coupling to other levels by evolution. Inconveniently, the latter requires definitions that maintain biophysical justification from the bottom-up, which conflicts with tractability. By means of a cell growth model, we exemplify a resolution for this conflict by polarization of Cdc42p in budding yeast, a process requiring clustering of active Cdc42p to one zone on the membrane and known to generate ample epistasis. Concretely, our model parsimoniously encompasses constant membrane area growth, stochastic Cdc42p turnover and a simple, justifiable polarity rule we define as the 'mesotype'. Through intuitively interpretable simulations, we describe previously documented, yet puzzling epistasis inside the polarity module. Moreover, we generate evolutionary relevant predictions e.g., on environmental perturbations, which are general enough to apply to other systems. We quantify how poor growth medium can equalize fitness differentials and enables, otherwise very distinct, evolutionary paths. For example, the fitness of the crippled Δ*bem1* relative to WT can easily be raised from 0.2 to above 0.95. Finally, we can extend our predictions on epistasis to other modules. We determine that modelled epistasis predictions only add predictive value when functional information of the involved modules is included. This inspires a road-map towards modelling the bidirectional genotype-phenotype map for other model systems with abundant interactions, where the intermediate levels reveal targets that evolution can optimize and facilitate a biophysical justifiable incorporation of epistasis.

Author summary

Efforts to understand how traits follow from genes facilitate a broad range of applications. For example, crops can be engineered faster to better resist drought, salt and heat stress, and medicines can be better tailored to individuals. Unfortunately, the path from genes to traits can generally involve a complex interplay of hundreds of genes and gene products whose individual contributions can be heavily context-dependent. In this work, we provide the proof-of-concept in a relatively simple system for a road-map towards elucidating this path. We have constructed a cell growth model for budding yeast, only involving simple rules on membrane growth, protein production and centrally, polarity, the process where yeast chooses the future division site. Despite the simplicity, the polarity rule is fully justifiable from underlying biophysics. Model simulations show good accordance with formerly puzzling traits, and also predict the ease with which the environment can change evolutionary paths. While lab conditions may prohibit the emergence of certain polarity mutations, this becomes much more feasible 'in the wild'. The tractable model nature allows us to extrapolate the context dependence of mutational effects beyond polarity, showing that this method for understanding trait generation also helps to elucidate protein evolution.

Introduction

Many fields, such as personalized medicine [1], agriculture [2], chemical production [3] and forensics [4], will greatly benefit from advances in understanding of the so-called genotype-phenotype (GP) map, the way that traits are connected to genes. However, this connection can be quite complex even for known heritable traits ("missing heritability") [5], limiting the power of genome-wide association studies [6]. On the one hand, one gene can be responsible for multiple traits, pleiotropy, although this may not always be very common [7]. On the other hand, multiple genes can contribute to one trait. Frequently, their individual effects are non-

additive in humans [8,9], but also in model systems as *Escherichia coli* [10] or *Saccharomyces cerevisiae* (budding yeast) [11], a phenomenon known as epistasis. Theoretically, epistasis is expected to surface very easily based on metabolic network analysis [12], and has some known molecular origins [13]. While epistasis can be inconsequential for fitness evolution [14], its presence complicates the predictions of phenotypes from genotypes and consequently gene evolution [15,16]. Therefore, predictions on epistasis constitute an important challenge for GP-map models.

As a modelling tool to more easily decompose the GP-map, intermediate levels can be defined as stepping stones [17], which can be brought under the general denominator of causally cohesive genotype-phenotype models [18]. An intermediate level may provide an entry point for additional observables that fine-tune predictions, but an abstract, unobservable entity as a definition is also possible. Most importantly, a level serves to break up and re-bundle the intertwined paths from individual genes to traits such that a more modular and hence more tractable picture arises. In that view, a suitable level definition acts as a tree which branches out to otherwise difficult to connect genotypes and phenotypes.

Multiple level examples exist, such as the biofunctional gene ontology level (ontotypes) [19], the network based trophic level [20], the diffuse endophenotypes [21] and the mathematical system design space [22]. Ideally, a one-level-fits-all approach exists, where the level definition facilitates understanding of the emergence of phenotypes from genotypes, while at the same time elucidating the handles for evolution, the reverse path in the GP-map. This requires steering away from phenomenological or statistical formulations to move towards biophysically sound versions, while at the same time maintaining tractability which often complicates bottom-up approaches. Consequently, the generation of a suitable level definition for the tractable

bidirectional path in the GP-map, if possible, involves coarse-graining of the underlying biophysics, specifically of molecular interactions and protein transport.

A promising attempt is the iMeGroCy model [23], where growth and cell cycle processes are simplified, and more details are kept for the module of interest, in this case the metabolism, which follows Michaelis-Menten kinetics. While effective in modularizing e.g., the pedigree phenotype emergence as a function of medium in *S. cerevisiae*, it is not straightforward to extrapolate this approach to other modules. The kinetics crucially assume well-mixed components, neglecting spatial heterogeneity arising from crowding in the cell [24]. Furthermore, coupling of diffusion and reactions involves conditions for pattern formation [25] that should be taken into account. Inclusion of spatio-temporal information is also known to be essential to understand the evolution of a network [26]. We therefore construct a cell growth model that encompasses the lessons derived from rigorous reaction-diffusion analysis, and maintains the simplicity across other growth features. The level definition associated with coarse-graining the molecular details will be called the 'mesotype'.

For the purpose of testing this bottom-up modelling approach, we require a model system known to exhibit ample epistasis (e.g., in doubling times) [27], namely polarity establishment in *S. cerevisiae*. Here, the unicellular organism budding yeast breaks its internal spherical symmetry to direct bud growth in one direction. This process is essential for the proliferation of the cell and relies on correct functioning of Cdc42p [28]. The mechanics behind this process is known to large detail (except the role of unexpected wildcard Nrp1p [27]), and involves clustering of the active form of small GTPase Cdc42p, which is bound to a GTP molecule, to one patch on the plasma membrane (Fig. 1A). In a wild-type (WT) background, rapid Cdc42p clustering is governed by a positive feedback involving Bem1p and Cdc24p [29], the relevant

guanine nucleotide exchange factor (GEF) for Cdc42p [30], which appropriately transport and activate Cdc42p. By contrast, its deactivation outside the membrane patch is ensured by GTPase activating proteins (GAPs), a protein class to which Bem2p, Bem3p [30], Rga1p [31] and Rga2p [32] pertain.

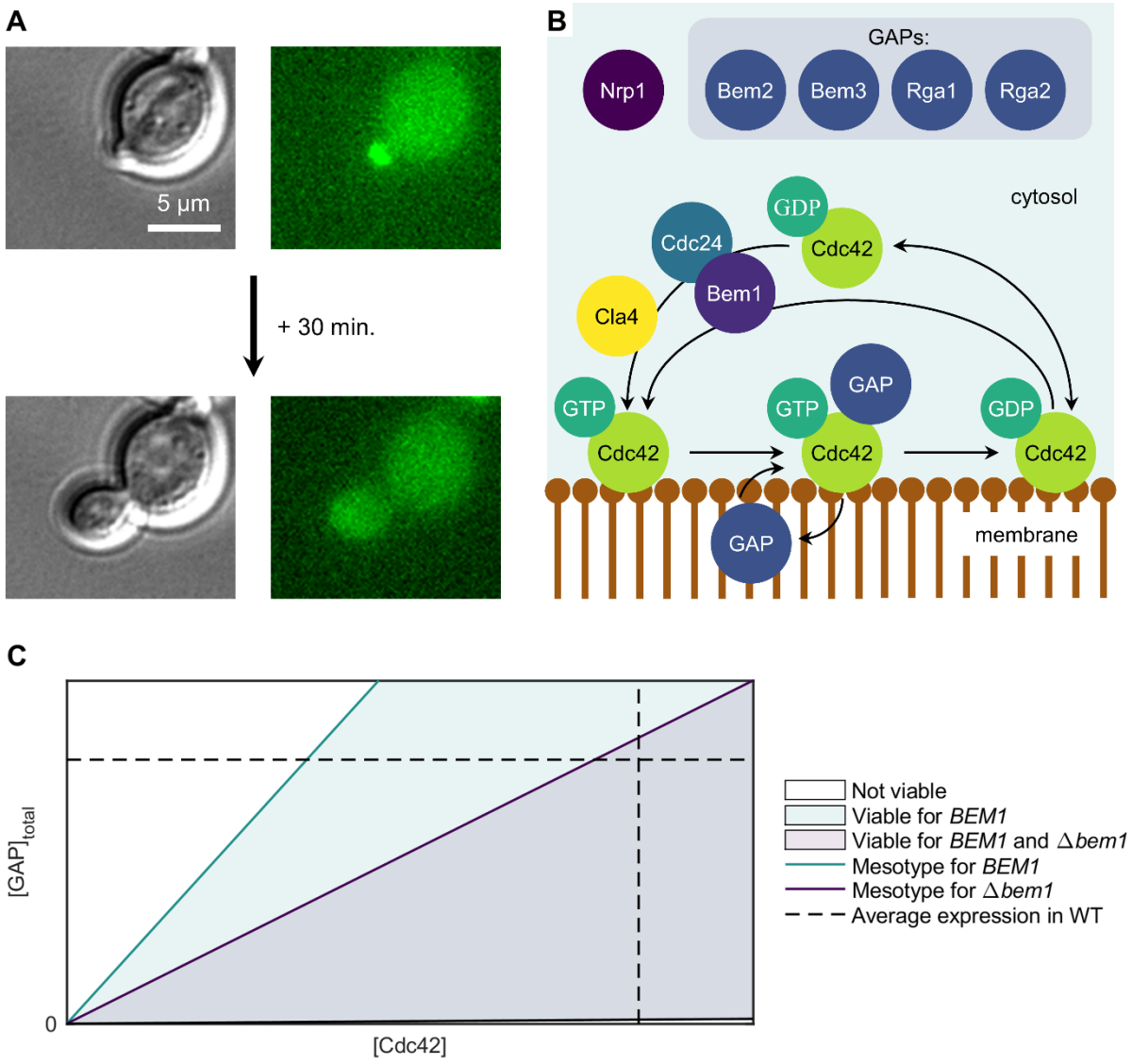


Fig 1. Yeast polarity as suitable genotype-phenotype map model for epistasis description. (A) Brightfield (left) and widefield fluorescence example images of a polarizing budding yeast cell (scale bar 5 µm). Key for polarity is clustering of active Cdc42p (of which a binding partner is fluorescently labelled in the images) to one zone on the membrane. This location marks the site of polarized growth. (B) Schematic overview of polarity protein core (proteins not to scale).

A positive feedback for (active) Cdc42p-GTP is mediated by either the Bem1p-Cdc24p complex, and likely to lesser extent by Cla4p. For Nrp1p it is unclear how it mechanistically links to other components. (C) Schematic diagram depicting phenotype (viability) as function of genotype through Cdc42p (active and inactive) and GAP concentration in the cell with or without Bem1p. An intermediate, the 'mesotype', is defined here as the limiting Cdc42p concentration. Epistasis is readily observed as the same increase in e.g., GAP concentration can yield inviability in the Δ*bem1* background but not in the *BEM1* background.

In absence of Bem1p, GAPs can more easily deactivate even the Cdc42p localized in the main patch that marks the future division site, which would generate a lethal situation for the cell. This can be circumvented if the abundance of Cdc42p is large enough to continuously sequester the GAPs found around the main patch, forming a rescue mechanism to establish polarity when combined with a generic positive feedback [33], such as through Cla4p [34] (Fig. 1B). Theoretical analysis of the underlying reaction-diffusion equations reveals a strong dependence of the ability to polarize success on the GAP/Cdc42p copy number ratio, where a broader range is viable in the presence of Bem1p [33] (Fig. 1C). This motivates a coarse-graining of the protein dynamics to a threshold for the protein concentration, which forms the mesotype level definition in this context.

Results

Coarse-grained bottom-up model design

Coarse description of cell expansion. We modeled the yeast cell cycle as a process involving three modules, namely coarse-grained cell growth, protein turnover and cell polarity (Fig. 2A). A parsimonious approach to cell membrane growth was chosen consisting of two stages of constant membrane area growth, as several alternative formulations proved immaterial for

phenotype description (S1 Fig.). The membrane expansion rates (C_1 for G1, C_2 elsewhere) were brought in decent agreement with literature [35,36] (see S2 Table for further justification of parameter choices.).

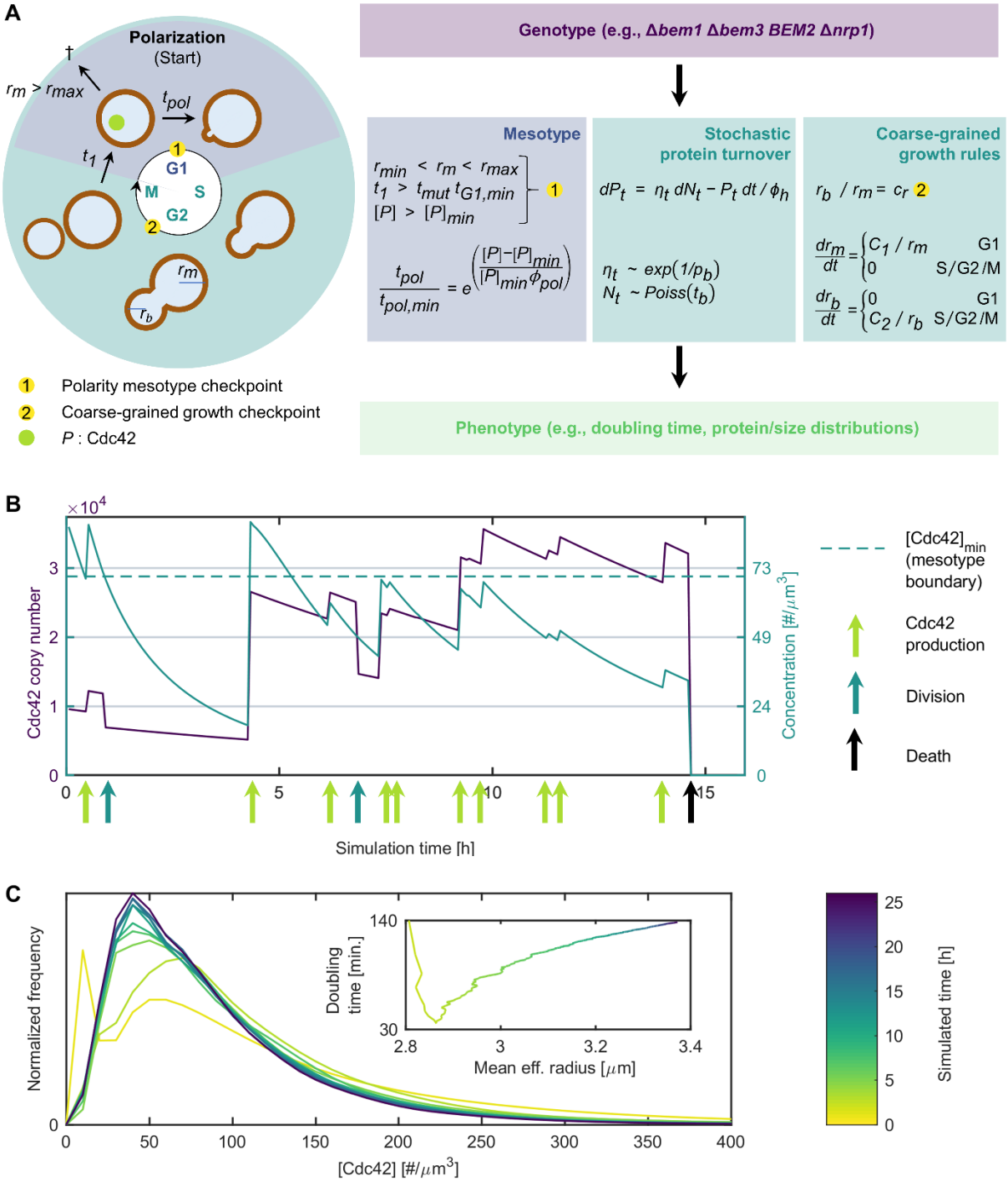


Fig 2. Coarse-grained, bottom-up growth model integrating the polarity mesotype to facilitate epistasis and phenotype prediction. (A) Schematic depiction of the translation of

the budding yeast cell cycle to model processes and parameters. Central is the moment of polarization which occurs when the cell has sufficient size, has been sufficiently long in G1 and has a Cdc42 concentration (abbreviated as [*P*]) exceeding a threshold, the latter defining the mesotype checkpoint for this system. Together with a coarse description of cell growth (constant membrane expansion of either mother or bud and a bud size checkpoint), stochastic protein production and deterministic degradation, this allows construction of the genotype-phenotype map. (B) Example trace of the Cdc42 copy number (blue) and concentration (green) of a single cell, which are subject to protein production and degradation (and dilution for concentrations). The cell must exceed the mesotype threshold ($[P]_{\min}$) before division can take place. When this is delayed for too long, the cells expands beyond r_{max} and the cell dies (after almost 15h). (C) Convergence of Cdc42 copy number distribution during simulations. Simulated time since ancestor is approximate as birth times of the cells in the starting population are distributed across an 83 min. bandwidth. The inset shows how the estimates of the population doubling time and the average effective cell size equilibrate as a function of time.

The first stage involves isotropic growth of a spherical cell of radius $r_m(t)$, which corresponds to the G1 phase including the Start transition, at the end of which a checkpoint must be passed, which is further explained in the following paragraph. Thereafter, the cell switches to the second stage of growth, where the membrane grows in a polarized manner defining a bud with radius $r_b(t)$, while the rest of the mother cell retains its size. The bud membrane growth area is constant for the modelled equivalent of the S, G2 and M phase but larger than for the mother in G1. Bud growth lasts until the second checkpoint, at which the bud proceeds as an independent cell when it has reached a sufficient size ($r_b = r_m\, c_r$).

193 **Biophysically justifiable mesotype inclusion.** The cells grow isotropically until three
194 conditions are met, defining the first checkpoint. Firstly, the radius r_m of the cell must exceed
195 the minimum size threshold r_{min}. Secondly, the time in this stage (t_1) exceeds a minimum time
196 ($t_{G1,min}$), which may be modified by a factor t_{mut} for certain mutations with respect to WT.
197 These two criteria result from key events in the timing pathway, particularly cell size dependent
198 control of Cln3p arrival to the nucleus by Ydj1p [37]. Finally, the concentration of Cdc42p, $[P]$,
199 must exceed a minimum concentration threshold $[P]_{\min}$. The existence of the Cdc42p
200 concentration threshold, which we define as the 'mesotype' for a particular mutant, follows
201 from rigorous theoretical and experimental analysis of reaction-diffusion equations of the
202 polarity network [33].
203
204 Once all three conditions are met, isotropic growth continues for a period of t_{pol}, which lasts at
205 least $t_{pol,min}$ and depends exponentially on the relative excess Cdc42 concentration above the
206 threshold (scaled by ϕ_{pol}). This is a simplified representation of the results in [33] and reflects
207 the period where Cdc42p clusters to one zone in the membrane. As it can occur that the Cdc42
208 threshold is never exceeded while growth continues and concentration are diluted, the cell is
209 considered dead when its radius exceeds the maximum size r_{max}.
210
211 **Noisy protein production.** Whether the Cdc42 concentration condition is met depends also
212 strongly on protein production, which is modelled as a stochastic process. Since mRNA lifetime
213 of Cdc42 is much smaller than its protein half-life t_h [38,39], Cdc42p production essentially
214 follows from instantaneous bursts, which are modelled as an compound Poisson process burst
215 process N_t with exponentially distributed size η_t (on average p_b) at exponentially distributed
216 intervals (on average t_b) [40]. Because it may be important for the precise crossing time of the
217 polarity threshold, we avoid absorbing protein degradation in an effective burst size, by

218 explicitly adding degradation to the Cdc42 copy number process P_t. Stochasticity in total GAP
219 copy number is not included as the cell-to-cell variability is much less than for Cdc42p (GAP
220 coefficient of variation < 0.15, only just above the smallest measured value of 0.10 [41],
221 compared to 0.83 for Cdc42p measured in this study).
222
223 **Coarse-grained bottom-up model verification and validation**
224 Firstly, the model design was verified by simulations of the computational model
225 implementation (see Materials and Methods) which allowed tracking the states of individual
226 cells or the population. Fig. 2B shows a Cdc42p copy number and concentration time trace of
227 a single cell. The Cdc42p production is burst-like and occurs as indicated on the time axis. The
228 copy number trace shows the proteins degrade between these bursts, and there is also dilution
229 due to cell volume growth for the concentration curve. In this example, divisions occur twice
230 shortly after checkpoint 1 has been passed, which also implies exceeding the mesotype
231 concentration threshold. Ultimately, this cell fails to exceed the threshold a third time and dies
232 after exceeding the maximum size r_{max}, as designed. Fig. 2C shows the rate of convergence of
233 relevant population phenotypes (without plotting the dilution step). After the population has
234 grown approximately 25 hours counted from the ancestor seed, the Cdc42p distribution has
235 largely converged. The size and doubling time change 0.6% and 0.9% respectively across the
236 last 200 minutes, well within the typical experimental error (see S2 Fig. for an example of
237 results including dilution).
238
239 Secondly, we turned to the model validation, where five parameters (four mesotype thresholds,
240 and one *nrp1*-dependent G1 time factor t_{mut}) are fitted. Non-trivial previously measured
241 observables are considered mostly from [27]; strong epistasis in growth rates between GAP
242 mutants only in the Δ*bem1* background, strong epistasis between *BEM1* and *NRP1*, and non-

243 monotonous optimization of G1 times for (reconstructed) experimentally evolved mutants
244 starting from Δ*bem1*. For the latter phenotype, the acceleration of G1 speed of the Δ*bem1* cells,
245 despite its poor fitness, compared to WT cells is particularly noteworthy. This combines to a
246 total of 20 phenotypes, well identifying the five free parameters.
247
248 **Description of the coarse GP-map exhibiting epistasis.** The simulated growth rates of
249 polarity mutants of [27] were calculated as a function of mesotype ($[Cdc42]_{min}$), which scales
250 linearly with total GAP concentration $[GAP]_{tot}$ due to the cone-like structure of Fig. 1C. These
251 can be converted to relative fitness values through division by the WT growth rate. Fitness
252 values in presence of *NRP1* were brought in accurate accordance with experiments of [27] (Fig.
253 3A) and consequently, the observed GAP epistasis is well (and robustly, see S1 Fig.) described.
254 The *nrp1* background was not always well fitted (5/8 correct within experimental error),
255 although these mutants suffered from relatively large experimental uncertainties. The four fitted
256 mesotype threshold concentrations are consistent with the *bem3* deletion effect that is twice as
257 large as for *bem2*. Given the GAP abundancies [42], this sets the *in vivo* Bem3p effective GAP
258 activity to be almost four times as large as for Bem2p, a difference much less pronounced than
259 measured *in vitro* [30].
260

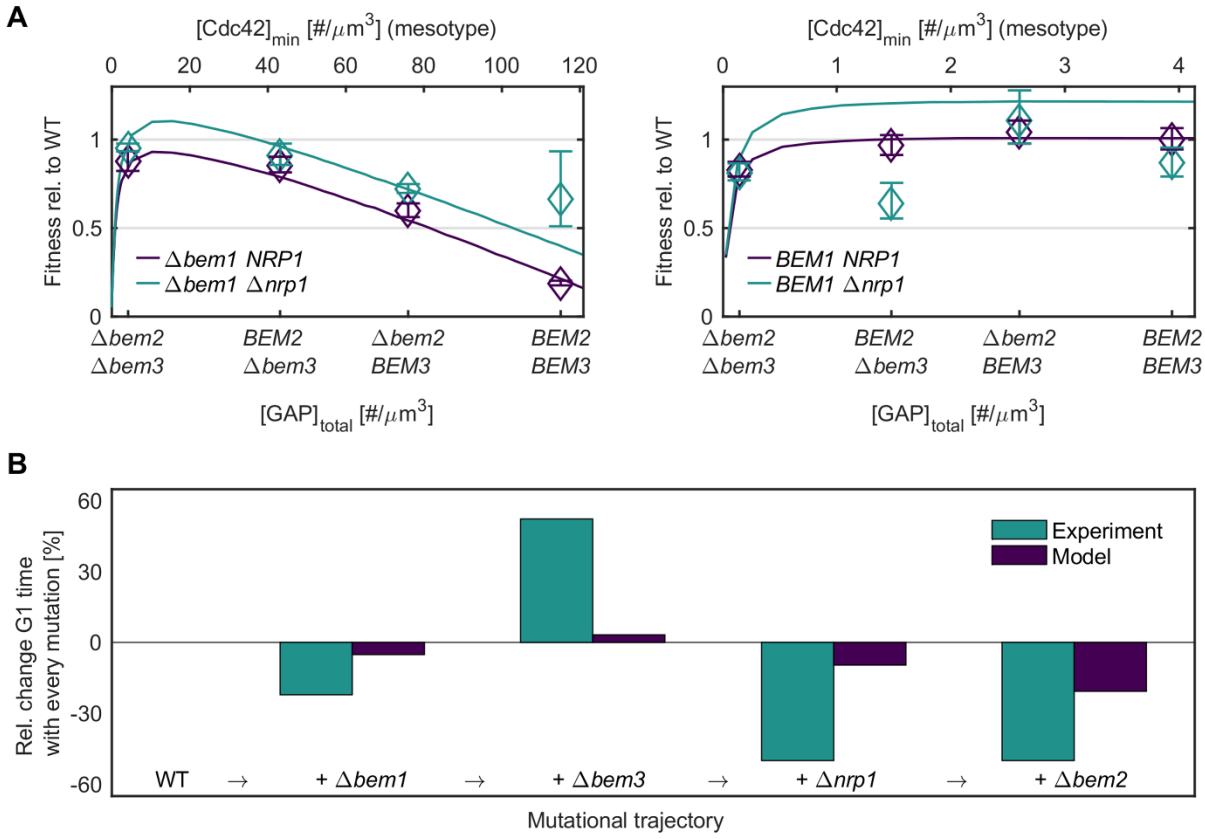


Fig 3. Comparison of the coarse-grained growth model to experimental data on coarse and subtler GP-maps. (A) Experimental fitness values relative to WT (phenotypes) for 16 different polarity genotypes [27] denoted by diamonds, which are fitted by the model as depicted by the dark purple and emerald lines for *nrp1* and *NRP1* background respectively. GAP genotypes can be linearly linked to the minimum Cdc42 concentration to polarize, the mesotype, as displayed on the top horizontal axis. Error of the *BEM2 NRP1* was not available and conservatively guessed. (B) The detailed phenotype of minimum G1 time as displayed by WT and four polarity mutants, comprising a single evolutionary trajectory. Experimental values are from [27] in emerald (defined there as tie to first polarity spot), model values are in dark purple (defined here as the time in G1 until both the size and time criteria are met). Both cases are normalized to their respective WT values, such that each column denoted the relative change in G1 time compared to the previous step in the trajectory.

Incorporation of a subtler GP-map. While doubling times represent a rather coarse phenotype, an example of the more detailed traits that can be modelled is time spent in G1. To this end, simulations were performed with half the normal membrane area rates C_1 and C_2, to mimic the poorer content of synthetic medium in which experiments from literature [27] were performed. The observed trends in G1 times along the evolutionary trajectory from WT to the fully evolved mutant in that paper were qualitatively matched, including the unusual inversion for the Δ*bem1* (Fig. 3B). The logic behind this inversion is that for WT cells in slower growth medium, the size requirement is the most important criterion for the first checkpoint of Fig. 2A, which can last longer than the minimum G1 time. By contrast, the on average less fit and larger Δ*bem1* cells are relatively more stalled by the minimum time criterion, and the long overall cycle times arise due to lengthy other phases.

A more realistic (and less coarse-grained) modification of the modelled cell cycle progression can improve the quantitative match. Suppose for example the *Δbem1* cells if the assumed minimum G1 time set is not a constant but a distribution (times for symmetry breaking in daughter cells can be quite stochastic [43]). Some Δ*bem1* cells have an early opportunity to fulfil the mesotype threshold concentration requirement, with which they usually struggle, while others are delayed more. This increases the cell-to-cell variation in fates in G1, since cells with fast G1 times are most likely to generate a first spot, while slow cells never generate this spot do not show up in the statistics. This is how less coarse-graining can lead to a larger decrease in G1 times than is the case with constant $t_{G1,min}$.

Genetic interaction predictions

Poorer medium quality reduces fitness differentials. As aforementioned, the effect of the environmental effects such as changes in growth media quality can be integrated in the model

through a change in membrane area growth rates C_1 and C_2. To assess the evolutionary consequences of poorer medium content, we considered a roughly three-fold area growth rate range that caused WT fitness to span between 0.5 and 1 (normalized to maximum growth). Fig. 4 shows the fitness ratio for various media within this range between the Δ*bem1* and *BEM1* background, as a function of GAP concentration, visualizing the trend of smaller fitness differentials for decreasing GAP concentrations and decreasing medium quality.

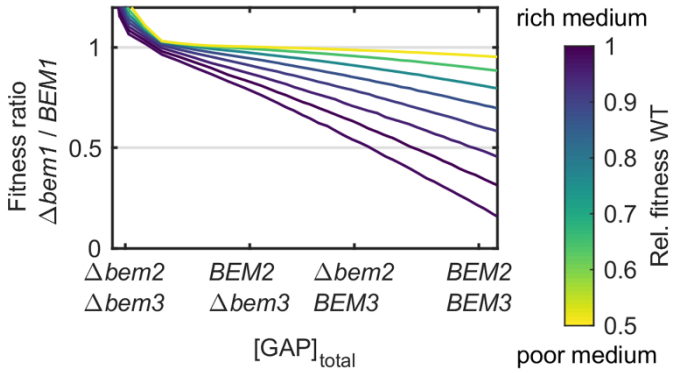


Fig 4. Growth model predictions of the environmental effect on polarity epistasis. Simulated fitness differences between *BEM1* and *bem1* backgrounds as a function of medium quality, which is integrated in the model through varying cell membrane area growth rates. The colors depict these rates through their associated WT doubling times. Generally, poorer medium reduces differences in fitness and genetic interactions between GAPs when comparing the *BEM1* and *bem1* backgrounds.

The intuition for this result is as follows. As seen in Fig. 3A, the Δ*bem1* background suffers from the high Cdc42p concentration threshold, relevant at checkpoint 1 (fig. 2A), and recover fitness when this threshold is lowered by successive GAP deletions. Fig. 2B had in turn shown the strong negative influence of dilution on the ability to exceed this threshold. Therefore, *Δbem1* cells benefit greatly from reducing the speed of membrane growth, while WT cells, for which the threshold is not a problem at all, only suffer from slowing down the membrane

321 growth. An unmodelled inhibitor of this effect would be a reduced Cdc42p production in
322 medium with lower quality. However, Cdc42p expression is at least known to remain stable
323 upon switching from dextrose to ethanol, an inferior carbon source [44].
324
325 **Information on biological function of mutated genes are a prerequisite for predicting**
326 **epistasis.** To assess whether we can extend the model predictions beyond polarity, we focus
327 on predictions of epistasis. This is the most generalizable quantity to assess cross-modular
328 interactions and is as mentioned in the introduction critical for constructing GP-maps. For this
329 purpose, we considered high-throughput data on numerous mutants, with varying levels of
330 detail regarding the mutant phenotypes, which we define as the information content. This
331 information will determine the precision with which the mutant can be incorporated into the
332 model.
333
334 Concretely, we restrict ourselves to epistasis between general mutants and Δ*bem1*, since we
335 suspected that fitness differences in this ill background are exaggerated and hence more likely
336 to have been picked up in literature. We used Bayesian analysis on the prevalence of epistasis
337 signs to determine what degree of information on the general mutants add value to sign
338 predictions. The general mutants were absorbed in the model in three different ways; either
339 using the coarse information on the single deletion phenotype (deleterious or beneficial), or the
340 mid-detail information on the single deletion phenotype (faster, slower, larger or smaller in G1),
341 or the functional information (proteasomal, phospholipid or ribosomal). Within these three
342 categories, there is a further subdivision into two sets, based on whether the model predicts
343 positive or negative epistasis with Δ*bem1* (Fig. 5A).
344

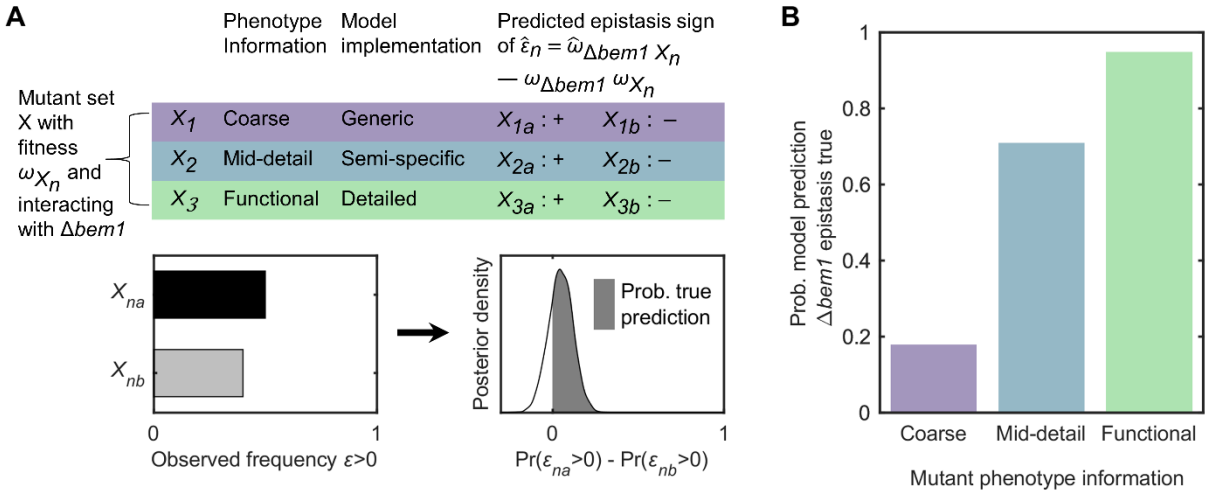


Fig 5. Adequate model predictions of epistasis rely on functional information concerning mutations. (A) Workflow for model prediction on epistasis ε of general mutants (fitness ω_X) with Δ*bem1*. Mutants are divided into three categories and two subsets, depending on the specificity of the mutant phenotype and model implementation and the subsequent model prediction. For each category *n* and subset *a*/*b*, the beta posterior density of the observed positive epistasis fraction can be constructed (from a binomial likelihood an uniform prior). The probability of a true prediction is then defined as the area below the posterior density of the difference of sets *a* (prediction ε>0) and *b* (prediction ε<0). (B) Bars reflecting Bayes factors for the model hypothesis; the ratio between the odds that the model prediction is true and false.

Firstly, mutants of which the coarse information are incorporated through modifying the membrane area growth rates, concretely smaller and larger rates for deleterious and beneficial mutants respectively. As seen in Fig. 4, smaller rates reduce the deleterious effect of the Δ*bem1*, prompting the prediction that negative epistasis with Δ*bem1* is generally more prevalent for deleterious mutants than for beneficial mutants. The analysis shows no evidence that this statement is correct (only a 20% chance, Fig. 5B).

Analogously, integrating the mutants on mid-detail information implies changing $t_{G1,min}$ (shorter when fast in G1, longer when slow) or r_{min} (lower when small in G1, higher when large). Mutants with shorter $t_{G1,min}$ and lower r_{min} disproportionally benefit the Δ*bem1* which suffers most from Cdc42p dilution before the mesotype checkpoint. Therefore, the model prediction is that mutants fast or small in G1 have more negative epistasis with Δ*bem1* than mutants that are slow or large in G1. Still, the experimental evidence is not compelling (70% chance).

Finally, when incorporating the mutants using functional information, we lower τ_h (proteasomal), membrane growth rates (phospholipid) and mean burst size p_b (ribosomal). The former two, which mitigate the problematic lack of Cdc42p in the Δ*bem1* to some extent, should exhibit more negative epistasis than the latter one, which deteriorates the Δ*bem1* situation. There is strong positive evidence for this statement (using the rules-of-thumb on Bayesian odds ratios [45]), which is true with around 95% certainty. This displays the benefit of integrating mutants based on functional information.

Discussion

We have constructed, verified, validated and applied a coarse-grained growth model encompassing the newly defined mesotype in order to describe phenotypes (subject to epistasis) from genotypes or predict these. When ample molecular information is present, as is the case for Bem1p and the GAPs, this strategy is quite successful to predict cell cycle times, given the largely good quantitative matches in Fig. 3A and C and qualitative match for the peculiar G1 time inversion for the Δ*bem1* compared to WT (Fig. 3B).

Additionally, the information content about the phenotypes, associated with mutated genes, required for predicting epistasis was assessed as it is a general hurdle for GP-map

388 models. As the mutant is encapsulated in our model through more detailed phenotypes, the
389 prediction quality increases accordingly. Typically, functional information is required to make
390 meaningful epistasis sign predictions (Fig. 5), similar to the ontotype strategy [19]. This
391 delimits the scope of this model.
392
393 Nevertheless, when only medium detail phenotypical information on the single deletion mutant
394 (such as in the case with Nrp1p) is used, predictions can still be of decent quality (Fig. 3A). The
395 efficacy of phenomenologically integrating Nrp1p into this model provided substance to the
396 claim that this protein is mechanistically involved in shortening G1. Since obtaining near-
397 complete information on the function of proteins is not within reach for most organisms, it is
398 comforting that mildly positive results may be achieved with phenomenological information
399 when building an otherwise biophysically justifiable bottom-up model.
400
401 Because the yeast polarity example shows the feasibility of our modelling strategy, we aim to
402 provide a road-map to apply these to general genotype-phenotype maps (Fig. 6). The core
403 functional component, in this case polarity, is modelled by justifiable coarse-graining, which
404 results in the mesotype of the system. This mesotype in turn emerges from functional subunits
405 [33], identifiable from the rigorous analysis of the underlying biophysics. Once multiple model
406 systems (such as the PAR protein system in *Caenorhabditis elegans* [46]) have been described
407 in this manner, it may be possible to construct a limited library of recurring subunits, making it
408 easier to recognize these in other systems and construct the corresponding mesotype. In
409 combination with a coarse-grained view of cell growth and noisy protein production, this
410 completes the bottom-up (population) phenotype prediction process.
411

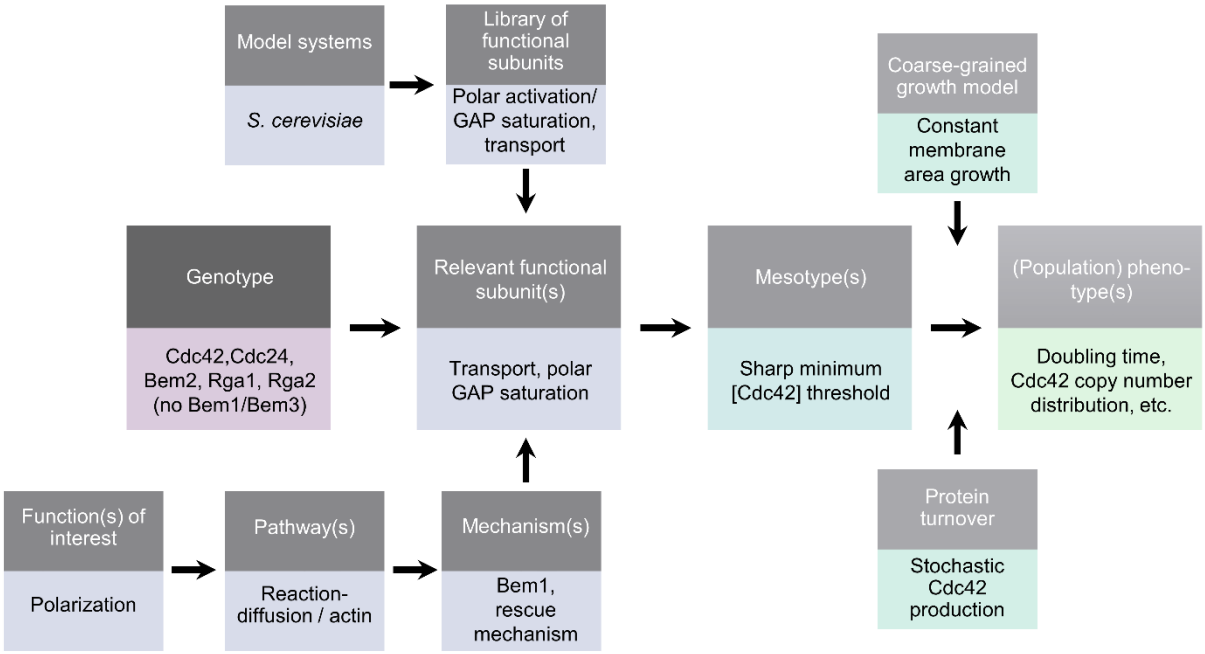


Fig 6. Proposed flow chart for phenotype predictions through intermediate levels. Bottom-up approach for phenotype prediction from genotype through mesotypes, which result from selecting the appropriate functional subunits. When more model systems are analyzed (e.g., polarization in *S. cerevisiae*, the min-system in *E. coli*, PAR-system in *C. elegans*), the toolbox from which to retrieve the relevant subunits expands. While it is possible to bypass the functional subunits and retrieve the mesotypes with rigorous numerical analysis of simulations of all protein components, the path displayed lends itself better to transfer knowledge of mesotypes to other systems. Bottom-half exemplifies the flow chart with the yeast polarity case.

Furthermore, the benefit of this approach is the tractable identification of evolutionary relevant quantities. For example, the GAP epistasis is accurately retrieved (Fig. 3A), and the prediction of the poor medium effect to reduce fitness differentials (Fig. 4) readily allows interpretation. The benefit of slower medium for the ill mutant Δ*bem1* fits the picture that haploinsufficiency in YPD is typically lifted in poorer medium [47], and opens up a distinct avenue for adaptation. Given that laboratory conditions are much more comfortable than the conditions under which historical evolution has taken and is taking place, the likelihood of fixation of a polarization

network optimized on Bem1p or an rescue mechanism (as experimentally occurring in [27]) becomes much more similar than naively expected. Moreover, this insight is quantifiable, we show e.g., that merely slowing WT down by a factor of 2 reduces the relative fitness differential to 0.05. Given that *BEM1* has comparable characteristics to an essential gene, the evolvability of essential genes may be greater than anticipated.

Materials and Methods

Model simulations were performed in MATLAB R2016a following a partial leap-like Gillespie algorithm [48] implementation (the G1 time until $r<r_{min}$ and $t_1<t_{G1,min}$, t_{pol} and the time through S/G2/M are one leap each). The core function and example script to demonstrate the functionality are found in S1 Code and S2 Code respectively.

Model parameters are summarized in S2 Table. An initial population asynchronized across a bandwidth of 83 minutes (all cells with equal radii of 2.2 μm and without protein) is grown until a population size of >5 million, after which a subsample of 1000 cells is regrown to the same condition. Doubling times are the average of the last hundred moving window (size 201 min.) linear regressions on the log number of cells.

Model calibration was done by supplying expression burst parameters for Cdc42p inferred from flow cytometry. These were fine-tuned, together with area growth rates C_1 and C_2, to yield a mean protein copy number of around 8700 [42] at the optimal growth doubling time of 83 minutes (WT in YPD [27]). Fluorescence measurements of the required C*DC42pr-GFP-CDC42* strain and a non-fluorescent strain (from [49,50]) were performed using a BD FACScan flow cytometer. Cells were pregrown in YNB (Sigma) + CSM -Met (Formedium) + 2% dextrose (Sigma-Aldrich), diluted to an OD_{600} of 0.1 and measured after 15h.

Doubling times of [27] in Fig. 3A were fitted using the native *fminsearch* on a normalized score objective for varying $[Cdc42]_{min}$ and manual inspection for setting t_{mut} (to 0.75) for the *nrp1* deletion. Interaction and phenotype data for Fig. 5 were obtained from BioGRID [51] and SGD [52] respectively (date of access March 2018).

Acknowledgements

We thank Marit Smeets for flow cytometry measurements of two strains and Leila Iñigo de la Cruz for growth data on one of these. Additionally, we thank Fridtjof Brauns for careful reading of the manuscript and Sophie Tschirpke for help in optimizing figure formatting.

References

1. Li C. Personalized medicine–the promised land: are we there yet? Clin Genet. 2011;79(5):403–12.
2. Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. Theor Appl Genet. 2013 Apr;126(4):867–87.
3. Oud B, Maris AJA, Daran J-M, Pronk JT. Genome-wide analytical approaches for reverse metabolic engineering of industrially relevant phenotypes in yeast. FEMS Yeast Res. 2012 Mar;12(2):183–96.
4. Kayser M, Schneider PM. DNA-based prediction of human externally visible characteristics in forensics: Motivations, scientific challenges, and ethical considerations. Forensic Sci Int Genet. 2009 Jun;3(3):154–61.
5. Maher B. The case of the missing heritability: when scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But

they were nowhere to be seen. Brendan Maher shines a light on six places where the missing loot could be stashed away. Nature. 2008;456(7218):18–22.

6. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017 Jul;101(1):5–22.
7. Wagner GP, Zhang J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nat Rev Genet. 2011 Mar;12(3):204–13.
8. Cordell HJ. Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet. 2009;10(6):392–404.
9. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci. 2012 Jan 24;109(4):1193–8.
10. Wünsche A, Dinh DM, Satterwhite RS, Arenas CD, Stoebel DM, Cooper TF. Diminishing-returns epistasis decreases adaptability along an evolutionary trajectory. Nat Ecol Evol. 2017;1(4):1–6.
11. van Leeuwen J, Pons C, Mellor JC, Yamaguchi TN, Friesen H, Koschwanez J, et al. Exploring genetic suppression interactions on a global scale. Science. 2016 Nov 4;354(6312):aag0839-aag0839.
12. Kryazhimskiy S. Emergence and Propagation of Epistasis in Metabolic Networks. bioRxiv. 2020 Jun 30;2020.05.24.113001.
13. Lehner B. Molecular mechanisms of epistasis within and between genes. Trends Genet. 2011 Aug;27(8):323–31.
14. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science. 2014;344(6191):1519–1522.
15. Papp B, Notebaart RA, Pál C. Systems-biology approaches for predicting genomic evolution. Nat Rev Genet. 2011 Sep;12(9):591–602.

16. Miton CM, Tokuriki N. How mutational epistasis impairs predictability in protein evolution and design. Protein Sci. 2016;25(7):1260–72.
17. Blanco-Gómez A, Castillo-Lluva S, del Mar Sáez-Freire M, Hontecillas-Prieto L, Mao JH, Castellanos-Martín A, et al. Missing heritability of complex diseases: Enlightenment by genetic variants from intermediate phenotypes. BioEssays. 2016 Jul;38(7):664–73.
18. Rajasingh H, Gjuvsland AB, Våge DI, Omholt SW. When Parameters in Dynamic Models Become Phenotypes: A Case Study on Flesh Pigmentation in the Chinook Salmon (*Oncorhynchus tshawytscha*). Genetics. 2008 Jun;179(2):1113–8.
19. Yu MK, Kramer M, Dutkowski J, Srivas R, Licon K, Kreisberg JF, et al. Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems. Cell Syst. 2016 Feb;2(2):77–88.
20. Johnson S, Jones NS. Looplessness in networks is linked to trophic coherence. Proc Natl Acad Sci. 2017 May 30;114(22):5618–23.
21. John B, Lewis KR. Chromosome variability and geographic distribution in insects. Science. 1966;152(3723):711–21.
22. Lomnitz JG, Savageau MA. Elucidating the genotype–phenotype map by automatic enumeration and analysis of the phenotypic repertoire. NPJ Syst Biol Appl. 2015;1:15003.
23. Pelillo M, Poli I, Roli A, Serra R, Slanzi D, Villani M. Artificial Life and Evolutionary Computation. Springer; 2018.
24. Schnell S, Turner TE. Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. Prog Biophys Mol Biol. 2004 Jun;85(2–3):235–60.
25. Halatek J, Brauns F, Frey E. Self-organization principles of intracellular pattern formation. Philos Trans R Soc B Biol Sci. 2018;373(1747):20170107.
26. Yamada T, Bork P. Evolution of biomolecular networks — lessons from metabolic and protein interactions. Nat Rev Mol Cell Biol. 2009 Nov;10(11):791–803.

27. Laan L, Koschwanez JH, Murray AW. Evolutionary adaptation after crippling cell polarization follows reproducible trajectories. eLife. 2015;4:e09638.

28. Adams AE, Johnson DI, Longnecker RM, Sloat BF, Pringle JR. CDC42 and CDC43, two additional genes involved in budding and the establishment of cell polarity in the yeast Saccharomyces cerevisiae. J Cell Biol. 1990;111(1):131–142.

29. Klünder B, Freisinger T, Wedlich-Söldner R, Frey E. GDI-Mediated Cell Polarization in Yeast Provides Precise Spatial and Temporal Control of Cdc42 Signaling. Saucerman JJ, editor. PLoS Comput Biol. 2013 Dec 12;9(12):e1003396.

30. Zheng Y, Cerione R, Bender A. Control of the yeast bud-site assembly GTPase Cdc42. Catalysis of guanine nucleotide exchange by Cdc24 and stimulation of GTPase activity by Bem3. J Biol Chem. 1994;269(4):2369–2372.

31. Stevenson BJ, Ferguson B, De Virgilio C, Bi E, Pringle JR, Ammerer G, et al. Mutation of RGA1, which encodes a putative GTPase-activating protein for the polarity-establishment protein Cdc42p, activates the pheromone-response pathway in the yeast Saccharomyces cerevisiae. Genes Dev. 1995;9(23):2949–2963.

32. Smith GR, Givan SA, Cullen P, Sprague GF. GTPase-Activating Proteins for Cdc42. Eukaryot Cell. 2002 Jun 1;1(3):469–80.

33. Brauns F, de la Cruz LMI, Daalman WK-G, de Bruin I, Halatek J, Laan L, et al. Adaptability and evolution of the cell polarization machinery in budding yeast. bioRxiv. 2020;

34. Tiedje C, Sakwa I, Just U, Höfken T. The rho gdi rdi1 regulates rho gtpases by distinct mechanisms. Mol Biol Cell. 2008;19(7):2885–2896.

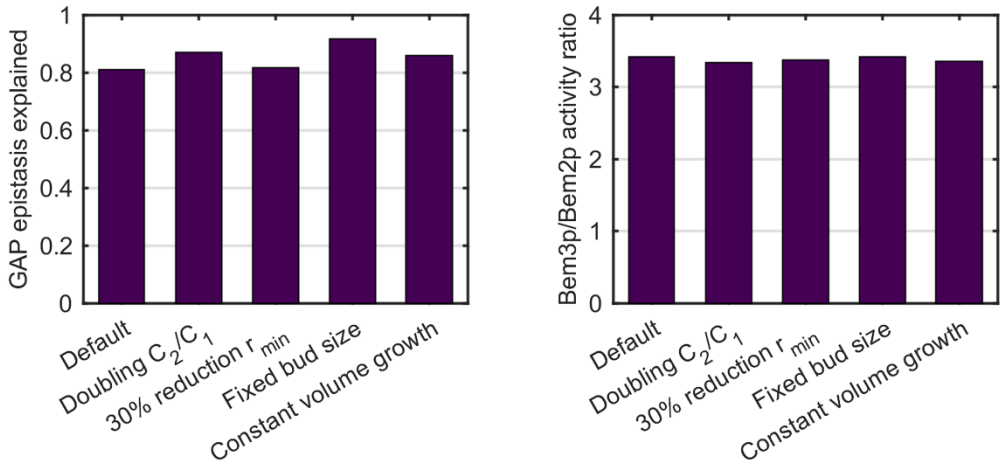
35. Goranov AI, Cook M, Ricicova M, Ben-Ari G, Gonzalez C, Hansen C, et al. The rate of cell growth is governed by cell cycle stage. Genes Dev. 2009 Jun 15;23(12):1408–22.

553 36. Ferrezuelo F, Colomina N, Palmisano A, Garí E, Gallego C, Csikász-Nagy A, et al. The
554 critical size is set at a single-cell level by growth rate to attain homeostasis and adaptation.
555 Nat Commun. 2012 Aug 21;3:1012.
556 37. Vergés E, Colomina N, Garí E, Gallego C, Aldea M. Cyclin Cln3 Is Retained at the ER and
557 Released by the J Chaperone Ydj1 in Late G1 to Trigger Cell Cycle Entry. Mol Cell. 2007
558 Jun;26(5):649–62.
559 38. Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR. Genome-Wide Analysis
560 of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals
561 Posttranscriptional Control of Ribosome Biogenesis Factors. Mol Cell Biol. 2004 Jun
562 15;24(12):5534–47.
563 39. Christiano R, Nagaraj N, Fröhlich F, Walther TC. Global Proteome Turnover Analyses of
564 the Yeasts S. cerevisiae and S. pombe. Cell Rep. 2014 Dec;9(5):1959–65.
565 40. Friedman N, Cai L, Xie XS. Linking stochastic dynamics to population distribution: an
566 analytical framework of gene expression. Phys Rev Lett. 2006;97(16):168302.
567 41. Chong YT, Koh JLY, Friesen H, Kaluarachchi Duffy S, Cox MJ, Moses A, et al. Yeast
568 Proteome Dynamics from Single Cell Imaging and Automated Analysis. Cell. 2015
569 Jun;161(6):1413–24.
570 42. Kulak NA, Pichler G, Paron I, Nagaraj N, Mann M. Minimal, encapsulated proteomic-
571 sample processing applied to copy-number estimation in eukaryotic cells. Nat Methods.
572 2014 Mar;11(3):319–24.
573 43. Moran KD, Kang H, Araujo AV, Zyla TR, Saito K, Tsygankov D, et al. Cell-cycle control
574 of cell polarity in yeast. J Cell Biol. 2019 Jan 7;218(1):171–89.
575 44. Smith EN, Kruglyak L. Gene–Environment Interaction in Yeast Gene Expression. Mackay
576 T, editor. PLoS Biol. 2008 Apr 15;6(4):e83.
577 45. Kass RE, Raftery AE. Bayes Factors. J Am Stat Assoc. 1995 Jun;90(430):773–95.

578 46. Goldstein B, Macara IG. The PAR Proteins: Fundamental Players in Animal Cell
579 Polarization. Dev Cell. 2007 Nov;13(5):609–22.
580 47. Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, et al.
581 Mechanisms of Haploinsufficiency Revealed by Genome-Wide Profiling in Yeast.
582 Genetics. 2005 Apr;169(4):1915–25.
583 48. Gillespie DT. Approximate accelerated stochastic simulation of chemically reacting
584 systems. J Chem Phys. 2001 Jul 22;115(4):1716–33.
585 49. Gulli M-P, Jaquenoud M, Shimada Y, Niederhäuser G, Wiget P, Peter M. Phosphorylation
586 of the Cdc42 exchange factor Cdc24 by the PAK-like kinase Cla4 may regulate polarized
587 growth in yeast. Mol Cell. 2000;6(5):1155–1167.
588 50. Freisinger T, Klünder B, Johnson J, Müller N, Pichler G, Beck G, et al. Establishment of a
589 robust single axis of cell polarity by coupling multiple positive feedback loops. Nat
590 Commun. 2013 May 7;4:1807.
591 51. Stark C. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006
592 Jan 1;34(90001):D535–9.
593 52. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al.
594 Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids
595 Res. 2012 Jan 1;40(D1):D700–5.
596 53. Stewart T, Strijbosch LWG, Moors H, Batenburg P van. A simple approximation to the
597 convolution of gamma distributions. 2007;
598 54. da Silva J, Coetzer M, Nedellec R, Pastore C, Mosier DE. Fitness Epistasis and Constraints
599 on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region. Genetics. 2010
600 May;185(1):293–303.
601 55. Howell AS, Jin M, Wu C-F, Zyla TR, Elston TC, Lew DJ. Negative Feedback Enhances
602 Robustness in the Yeast Polarity Establishment Circuit. Cell. 2012 Apr;149(2):322–33.

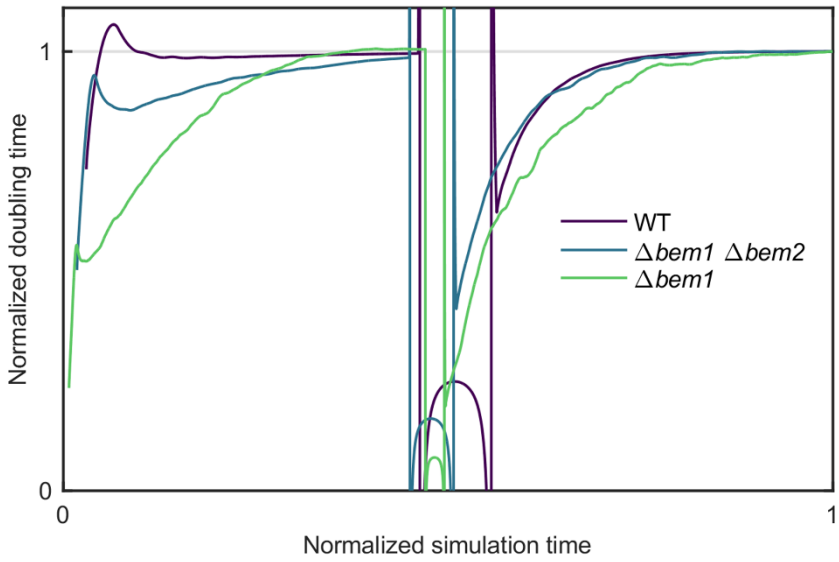
Supporting Information

The flow cytometry data (see S1 Dataset) was acquired using FlowJo CE software with a BD FACScan and later analyzed with home-written code in MATLAB. A single gamma distribution was maximum likelihood fitted on the fluorescence intensity counts of strains (from [49,50]) with simply endogenous *CDC42* ('background') or *CDC42p-GFP-CDC42* at 2% dextrose ('WT expression'). The WT expression distributions was analytically deconvolved for background counts using a gamma-sum approximation [53]. The average burst interval duration and average burst size result from these normalized distributions [40], a coarse doubling time estimate (200 min., processed as in [33]) for RWS1421 and the Cdc42p copy number estimate of 8700 from [42]. Calibration shows average burst sizes require a 20% reduction due to explicit inclusion of degradation in our model.



S1 Fig. Negligible influence of model details on GAP epistasis. Calculated GAP activity given model fits and abundancies from [42] (left) and relative multiplicative epistasis (definition of [54]) for the GAPs in the Δ*bem1 NRP1* background for the growth model and four variations; doubling of membrane area growth rate C_1, 30% reduction of r_{min}, change of

second checkpoint to a fixed bud size threshold of 1.8 µm, and constant cell volume instead of area expansion. WT membrane growth rates are recalibrated in each case to match 83 minutes.



S2 Fig. Convergence of simulated doubling times. Doubling times as function of simulation time for a fast (WT), medium (Δ*bem1* Δ*bem2*), and slow (Δ*bem1*) growing strain background. Doubling times and simulation times normalized to their respective final value. The dilution step midway temporarily causes unreliable estimates.

Strain name	**Genotype (W303 background)**	**Source**
RWS116	MATa *his3-11,15 ade2-1 can1-100 ura3 cln1::HisG* Δ*cln2 cln3::HisG YipLac204-MET-CLN2::TRP1*	[49]
RWS1421	MATa *his3-11,15 ade2-1 can1-100 ura3 cln1::HisG* Δ*cln2 cln3::HisG YipLac204-MET-CLN2::TRP1 CDC42p-GFP-CDC42::URA3*	[50]

S1 Table. Strain list. Strains used in this study.

Parameter	**Symbol**	**Value (background)**	**Source**
Cdc42 concentration threshold	$[Cdc42]_{min}$	115 # proteins/μm^3 (Δ*bem1*) 4 # proteins /μm^3 (*BEM1*) -62% (Δ*bem3*) -34% (Δ*bem2*)	This study, fitted
Minimum G1 time multiplier	t_{mut}	1 (*NRP1*) 0.75 (Δ*nrp1*)	This study, fitted
Minimum polarization time	$t_{pol,min}$	5 min.	[55]
Maximum polarization time	$t_{pol,max}$	600 min.	To truncate computations for cells with extremely low GAP content
Polarization time scaling parameter	ϕ_{pol}	25 (Δ*bem1*) 500 (*BEM1*)	$\phi_{pol,BEM1}$»1 and $\phi_{pol,BEM1}/\ \phi_{pol,\Delta bem1} \cong [Cdc42]_{min,\Delta bem1} / [Cdc42]_{min,BEM1}$ for observed small excess Cdc42 across backgrounds (this study)
Minimum G1 time	$t_{G1,min}$	15.6 min.	[36]
Minimum radius to polarize	r_{min}	2 μm	[36]

Maximum radius in G1	r_{max}	6 µm	To truncate computations for cells with very low Cdc42 content
Average Cdc42 expression burst interval time	$t_{b,WT}$	57 min.	This study, assuming theory from [40]
Average Cdc42 expression burst size	$p_{b,WT}$	4900	This study, assuming theory from [40] and with calibration
Cdc42p half-life	τ_h	474 min.	[39]
Bud/mother volume ratio checkpoint 2	c_r	0.89	Consistent with [36]
Ratio polarized/ isotropic membrane area growth rates (C_2/C_1)	c_p	2.13	Calibration
Isotropic membrane area growth rate	C_1	0.086 µm^2/min.	Analytical considerations for optimized WT (see S1 Text)

S2 Table. Growth model parameter list.

S1 Dataset. Flow cytometry and growth assay data. Flow cytometry data (raw, processed and fitted) of strains used in this study, and OD_{600} measurements of one strain (with fits).

S1 Code. Numerical implementation in MATLAB of the growth model in this study.

S2 Script. Test script in MATLAB calling the numerical model implementation of S1 Code.

S1 Text. Justification of membrane area growth rate value. We assume an optimized WT such that at checkpoint 1, the minimum size requirement is typically met at the same time that the minimum G1 time requirement is met, and that subsequent polarization time is minimal. After G1, the (squared) mother radius is then (integrating the radius equation from 0 to $t_{pol,min}$):

$$\frac{dr_m}{dt} = \frac{C_1}{r_m} \Longrightarrow r_m^2 = r_{min}^2 + \frac{C_1 t_{pol,min}}{2}$$

Thus, the bud radius is after next M-phase r_b=0.89 r_m. For self-consistentcy, this new cell must then expand to size r_{min} again at checkpoint 1, such that:

$$r_{min}^2 = r_b^2 + \frac{C_1 t_{G1,min}}{2} = 0.79 r_{min}^2 + 0.79 \frac{C_1 t_{pol,min}}{2} + \frac{C_1 t_{G1,min}}{2}$$

$$\Longrightarrow C_1 = \frac{0.42 r_{min}^2}{0.79\ t_{pol,min} + t_{G1,min}}$$

Using the values from S2 Table for r_{min}, $t_{pol,min}$, and $t_{G1,min}$, this leads to C_1=0.086 μm^2/min.