1

Differential strengths of molecular determinants guide environment specific mutational fates

4

5   **Rohan Dandage[1, 2], Rajesh Pandey[3,4], Gopal Jayaraj[1, 2], Kausik Chakraborty[1, 2,*]**

6

7

8   [1]CSIR- Institute of Genomics and Integrative Biology, Mathura Road Campus, New Delhi, India.

9   [2]Academy of Scientific and Innovative Research (AcSIR), New Delhi, India.

10   [3]CSIR Ayurgenomics Unit - TRISUTRA, CSIR- Institute of Genomics and Integrative Biology,

11   Mathura Road Campus, New Delhi, India.

12   [4]Mammalian Genetics Unit, MRC Harwell Institute, Harwell Science and Innovation Campus,

13   Oxfordshire OX11 0RD, United Kingdom.

14

15   *Corresponding author:

16   Email: kausik@igib.in

17

18

Abstract

20 Under the influence of selection pressures imposed by natural environments, organisms

21 maintain competitive fitness through underlying molecular evolution of individual genes across

22 the genome. For molecular evolution, how multiple interdependent molecular constraints play a

23 role in determination of fitness under different environmental conditions is largely unknown.

24 Here, using Deep Mutational Scanning (DMS), we quantitated empirical fitness of ~2000 single

25 site mutants of Gentamicin-resistant gene (GmR). This enabled a systematic investigation of

26 effects of different physical and chemical environments on the fitness landscape of the gene.

27 Molecular constraints of the fitness landscapes seem to bear differential strengths in an

28 environment dependent manner. Among them, conformity of the identified directionalities of the

29 environmental selection pressures with known effects of the environments on protein folding

30 proves that along with substrate binding, protein stability is the common strong constraint of the

31 fitness landscape. Our study thus provides mechanistic insights into the molecular constraints

32 that allow accessibility of mutational fates in environment dependent manner.

33

34

Author Summary

36 Environmental conditions play a central role in both organismal adaptations and underlying

37 molecular evolution. Understanding of environmental effects on evolution of genotype is still

38 lacking a depth of mechanistic insights needed to assist much needed ability to forecast

39 mutational fates. Here, we address this issue by culminating high throughput mutational

40 scanning using deep sequencing. This approach allowed comprehensive mechanistic

41 investigation of environmental effects on molecular evolution. We monitored effects of various

42 physical and chemical environments onto single site mutants of model antibiotic resistant gene.

43 Alongside, to get mechanistic understanding, we identified multiple molecular constraints which

44 contribute to various degrees in determining the resulting survivabilities of mutants. Across all

45 tested environments, we find that along with substrate binding, protein stability stands out as the

46 common strong constraints. Remarkable direct dependence of the environmental fitness effects

47 on the type of environmental alteration of protein folding further proves that protein stability is

48 the major constraint of the gene. So, our findings reveal that under the influence of

49 environmental conditions, mutational fates are channeled by various degrees of strengths of

50 underlying molecular constraints.

51

52

53

Introduction

55 Continuously fluctuating natural environments are known to a play central role in natural
56 selection by constituting environment to phenotype interactions (E2P). Such interactions can
57 predispose a particular genotype to alternative fates through diverted evolutionary trajectories
58 (1–5). At lower scales, molecular evolution of cellular genes, underlie the emergent organismal
59 adaptations. Cellular responses to the environmental conditions are well known to be often
60 enabled by maintenance of normal proteostasis (6). Therefore, an investigation of whether and
61 to how much extent protein folding and stability play a role in guiding mutational fates would be
62 a key factor in the efforts to elucidate molecular level E2P interactions.

63 Considering low rates of spontaneous mutations across variety of organisms (7), single site
64 mutations provide a view of immediate next fates of a gene. DMS approach has enabled
65 scanning of large scale of mutations in a high throughput manner (8,9); allowing comprehensive
66 analysis of sequence-space of genes. Resultant Distributions of Fitness Effects (DFE) provide a
67 continuous series of fitness effects ranging from strongly deleterious to beneficial which is a
68 valuable resource for quantitative genetics (10). In recent years, exploration of environmental
69 effects with a large-scale genotype to phenotype (G2P) data (11,12) has resulted in the
70 identification of environment specific differential mutational sensitivity. However, qualitative and
71 quantitative identification of determinants of fitness effects has been a challenging task (13). In
72 addition to improve much needed mechanistic understanding (14) of E2P interactions, such
73 information can potentially increase the robustness in current approaches of prediction of
74 phenotypes from genotypic information(15,16).

75 On a fitness landscape of a gene, levels of fitness of mutations may depend on number of
76 molecular constraints which would shape the landscape. Here, to systematically investigate the

77 underlying mechanisms of E2P interactions, firstly, we monitored fitness landscapes of

78 Gentamicin (Gm) resistant gene GmR (aminoglycoside 3-N-acetyltransferase (aacC1)) under

79 different physical and chemical environmental conditions which can have differential fitness

80 effects. Among physical environments, we studied effects of lower (30^0C) and elevated (42^0C)

81 temperature than the normal growth temperature of *E. coli* i.e. 37^0C. Elevated temperature

82 condition is known to limit foldability of temperature-sensitive mutants (17), hence its influence

83 on the fitness landscape of GmR allows us to understand the implication of proteostasis on

84 limiting the sequence space for this protein. Among chemical environments, we studied two

85 model chemical chaperones - TMAO and glycerol which are known to assist cellular protein

86 folding in vitro and in vivo, assist mutational buffering (18) and thermodynamically act on early

87 refolding intermediate by non-identical mechanisms (19).

88 In this study, we show that mutations may provide selective advantage or disadvantage

89 depending on acting environmental conditions. Among multiple molecular constraints of fitness

90 landscape of GmR, perturbation of stability is identified to be the universal strong constraint

91 across all test environments. Except for the case of simultaneous action of multiple individual

92 environments, the relative selection pressures conferred by physical and chemical environments

93 are largely dependent on folding constraint and hence are predictable. For instance, elevated

94 temperature imposes a negative selection pressure while chemical chaperones are found to

95 exert mutational robustness (buffering effect). Collectively, through mutational scanning of an

96 essential gene, this study uncovers the largely unclear role of environments in protein evolution

97 and contributions of underlying molecular constraints in determining the mutational fates.

98

99

Results

Deep mutational scanning of GmR

102 We implemented DMS method (8) to monitor survivabilities and empirical competitive fitness

103 levels of individual mutants from a Single Site Mutation (SSM) library of GmR (18). For co-

104 culture bulk competitions, we used purifying selection i.e. Gm concentration of 12.5 μg/mL

105 which is ~4 fold lower than the minimal inhibitory concentration (MIC) of *E. coli* (K-12)

106 expressing wild type GmR while being higher than the MIC of the host system i.e. *E. coli* alone

107 (S1 Fig 1A). While survivability of the antibiotic resistance is dosage dependent (S1 Fig 1B), use

108 of a weaker Gm selection level allows assessment of mutants with moderate catalytic

109 advantage which would otherwise be lost because of higher likelihood of 'quick fix' outcomes

110 (20). Note that when not otherwise mentioned, for the bulk competitions, 12.5 μg/mL Gm is

111 used. Co-culture bulk competitions of mutants were carried out in two sets. One in presence of

112 Gm selection (selected pool) and other in absence of Gm selection (unselected pool). At the

113 end of the bulk competitions, ultra-deep sequencing of the amplicons of GmR allowed

114 quantitation of counts of mutants survived at the end of each competition (Fig 1A). To

115 understand E2P effects of different environments, co-culture bulk competitions were carried out

116 under different test environmental conditions (Materials and Methods). 37^0C being the optimal

117 growth temperature of *E. coli*, we simply refer it as the reference environment. When not

118 otherwise noted, test environments would be compared against the reference environment

119 (37^0C, 12.5 μg/mL Gm).

120 Strong correlations between counts of the mutants from independent biological replicates of

121 bulk competitions (S1 Fig 2) indicate low level of biological noise associated with experiments.

122 Preferential enrichments i.e. log fold changes of the counts of the mutants in selected pool with

123 respect to that in unselected pool serve as proxies for competitive fitness of mutants. Next, we

124 implement a strategy adopted in previous study on similar mutational scanning of antibiotic

125 resistant gene (21), to classify statistically neutral mutations by applying a cutoff of $\mu \pm 2\sigma$ (μ:

126 mean, σ: standard deviation) obtained from fitness of unselected pool (S1 Fig 3). Accordingly,

127 mutants were classified as relatively enriched ($F_i > \mu + 2\sigma$) and depleted ($F_i < \mu - 2\sigma$). Fitness levels

128 of synonymous mutations across different environments are found to strongly correlate

129 (Pearson's $r > 0.88$, S1 Fig 4) with that fitness of synonymous mutations under reference

130 environment (37^0C, 12.5 µg/mL Gm); thus diminishing the influence of codon bias associated

131 with G2P interactions of GmR. Therefore, unless specified, in the subsequent analysis of

132 mutational data, we primarily utilize non-synonymous mutants. This way, from a co-culture bulk

133 competition under reference conditions (37^0C, 12.5 µg/mL Gm), we obtained empirical fitness of

134 2004 non-synonymous mutants (n) (S1 Fig 5, S2 Table). Among them, 609 mutants are

135 enriched, 262 mutants are depleted and 1133 are classified as statistically neutral (S1 Table1).

136 We compared pairs of DFEs obtained for various test environments by a metric defined as a

137 difference in average fitness scores of mutants between test and reference condition (ΔF) which

138 indicates the relative increase or decrease in fitness of surviving mutants in test condition with

139 respect to reference environment. Additionally, number of mutants having significantly higher or

140 lower fitness in test environment are measured (n_{pos} and n_{neg} respectively). Here, to account for

141 the inherent experimental and biological noise, a cutoff is assigned for defining significant

142 increase or decrease in the fitness over the inherent dispersion within replicates (Materials and

143 Methods).

144 DFE of GmR obtained at 25 µg/mL Gm (37^0C) reveals a depletion of fitness with a skew

145 towards the deleterious fitness levels as compared to DFE obtained at 12.5 µg/mL Gm (Fig 1B).

146 The lowered average fitness ($\Delta F = -0.38$) is accompanied by lowered survival of mutants in this

147 condition ($\Delta n = -246$). Such dosage dependent deleterious fitness effects are expected from

7

148   increased stringency for *in vivo* catalytic functionality and survivability of mutants which is

149   consistent with previous reports on mutational scannings of other antibiotic resistant genes (21–

150   23). In terms of distribution of enriched and depleted mutants per position, mutations at sites

151   with low evolutionary rate, such as binding sites, show depleted levels of fitness while mutations

152   at sites with high evolutionary rate, specifically at N-terminal region, exhibit high fitness levels

153   (S1 Fig 6). Collectively, dosage dependence and consistent trends with conservation scores

154   support that the observed empirical fitness values are representatives of the *in vivo*

155   functionalities of GmR mutants. This further allows us to ascribe empirically quantified fitness

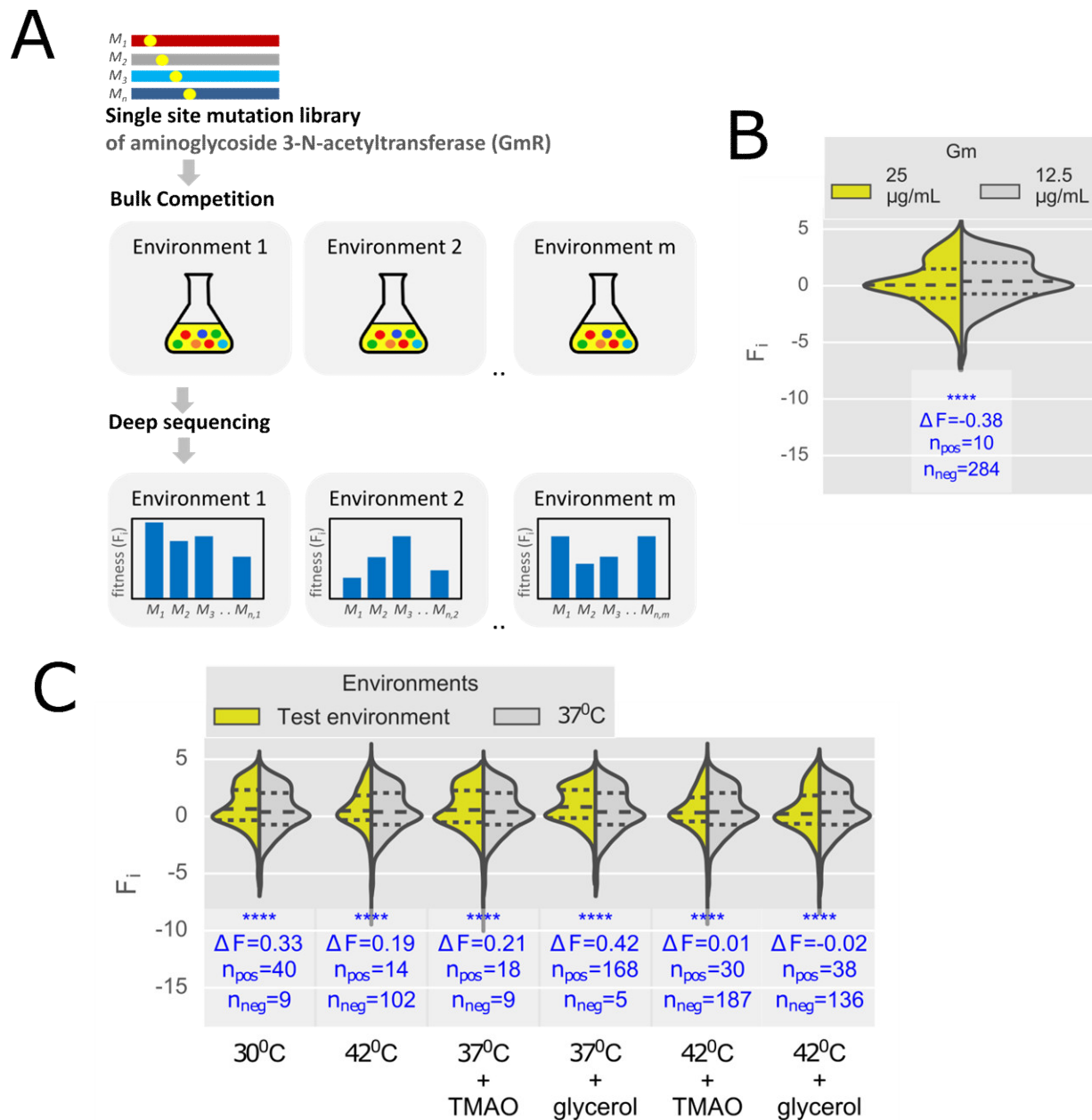156   values to the *in vivo* functionalities of the mutants.

157

Fig 1. Deep mutational scanning of GmR. (A) Experimental strategy to monitor survivabilities and competitive fitness of GmR mutants (Materials and Methods). SSM library of GmR was subjected to co-culture bulk competition in different environments in presence of Gm selection (selected pool). An independent co-culture bulk competition in the absence of Gm selection (unselected pool) acts as a reference condition. Preferential enrichments in counts of mutants - measured via deep sequencing - in selected with respect to that unselected pool provides a

9

164   measures of the survivabilities and proxies for competitive fitness levels of mutants. (B) DFE

165   obtained at stringent concentration of Gm i.e. 25μg/mL (37^0C) is compared to DFE obtained at

166   12.5 μg/mL (37^0C). (C) Similarly, DFEs obtained under physical and chemical test environments

167   and combinations of environments are compared to reference environmental condition (37^0C).

168   Here, all the bulk competitions were carried out at purifying selection of 12.5 μg/mL Gm. F_i

169   denotes fitness score. Histograms of DFEs are fitted by kernel density estimation. Heights of the

170   DFEs are in scale with the frequencies of mutants. Dashed lines represent quartile values of the

171   fitness per distribution. Level of significance is measured by two sided Wilcoxon signed-rank

172   test. Relative selection pressure imposed by a test environment is with respect to control

173   condition is measured in terms of difference between average fitness scores (ΔF) and number

174   of mutants with higher (n_{pos}) or lower (n_{neg}) survivability than control environment.

175

Environmental conditions induce variable fitness effects

177   Using our experimental system, to understand E2P interactions, we monitored effects of sets of

178   physical and chemical environments. In our study, low (30^0C) and elevated (42^0C) temperature

179   conditions comprising a set of physical environments, while treatments of chemical chaperones

180   TMAO and glycerol constitute a set of chemical environments. Physical environments especially

181   elevated temperatures are well known to have limiting influence on the mutational tolerance

182   (17). On the other hand, in case of chemical environments, solvent accessible surface area is

183   known to be one of the strong predictor of protein evolution rate (24–26), suggesting that

184   solvent-protein interactions are critical for the understanding of molecular evolution. In our

185   study, we monitor effects of TMAO and glycerol which are known to act as chemical chaperones

186   in assisting folding of proteins by non-identical mechanisms (19) as well as assisting buffering of

187   genetic mutations (18).

10

188   We utilize prior information gained from growth kinetics of the model system i.e. wild type GmR

189   harboring *E. coli* under different test environments to establish the strengths of selection

190   pressures relative to reference environment. As compared to normal growth temperature of *E.*

191   *coli* (37^0C), both lower (30^0C) and elevated temperatures (42^0C) show non-optimal growth

192   (shown in S1 Fig 7A). At elevated temperature, the amplitude of the growth kinetics decreases

193   by ~40%, while at low temperature there is more than two fold increase in the lag phase of the

194   growth kinetics. Both the chemical environments evidently show to suboptimal growth rates (S1

195   Fig 7B), especially decrease in the amplitude in presence of glycerol suggests a stronger

196   selection pressure.

197   Through DMS, DFEs obtained for all test environments show significant disparities as compared

198   to reference environment i.e. 37^0C (Fig 1C, S1 Table1). Position-wise distributions of fitness

199   reveal differential enrichment and depletion along the length of the gene across all test

200   environments (S1 Fig8). Notably, survival of mutants under elevated temperature condition and

201   environments involving elevated temperature condition is greatly lowered (S1 Table1) indicating

202   deleterious effects of elevated temperature condition. Among environment specific DFEs,

203   compared to reference environment, at low temperature, survival of a subpopulation of mutants

204   is compromised (Δn = -141) while average fitness (ΔF = 0.33) of survived mutants in the

205   condition is increased. The increased average fitness is majorly due to compromised survival of

206   the fraction of depleted mutants. As the natural host of GmR gene i.e. *S. marscens* requires

207   lower growth temperature for optimal growth (27), the gene may have been evolutionary

208   optimized to prefer lower temperatures thus exhibiting partial mutational robustness (n_{pos} = 40,

209   n_{neg} = 9). Elevated temperature, imposes drastic deleterious effects on the survivabilities of

210   mutants (Δn = -637) accompanied by increased average fitness of survived mutants (ΔF = 0.19).

211   Lowered survivabilities are arguably due to protein misfolding (28) while because of prominent

212   elimination of low fitness temperature sensitive mutants the average fitness of mutants is

213 increased. The survival of high fitness mutants under elevated temperature condition may have

214 resulted due to stabilization to a particular trait such as higher thermal stability.

215 Among chemical environments, both the chemical chaperones are found to exert mutational

216 robustness as evident from retained survivabilities of mutants. In terms of fitness change, fitness

217 gain is stronger in case of glycerol ($\Delta F = 0.42$) than TMAO ($\Delta F = 0.21$) (Fig1C). The mutational

218 robustness supports the known mutational buffering as reported earlier (18). Notably, different

219 sets of mutations are buffered by the two chemical environments (S2 Table) supporting known

220 mechanistic difference in aiding protein folding (18,19). Further, at stringent Gm selection (25

221 µg/mL), compared to minimal Gm selection, TMAO treatment is found to exert mutational

222 robustness (S1 Fig 9A). Remarkably, glycerol treatment exerts mutational robustness compared

223 stringent (S1 Fig 9A) as well as minimal Gm selection (S1 Fig 9B). Collectively, among

224 individual treatment of environments, low temperature and chemical chaperones induce

225 mutational robustness while, elevated temperature exert deleterious fitness effects.

226 **Fitness effects of combinations of environments**

227 Phenotypic effects of simultaneous action of multiple environments have been historically

228 illusive to decipher (29). Using our experimental system, we set out to elucidate the effects of

229 combinations of different sets of environments having prominent but contrasting effects i.e.

230 elevated temperature and chemical chaperones. From growth kinetics of the model system,

231 under combination of environments, it is apparent that TMAO is able to counteract the effects of

232 elevated temperature while in case of treatment of glycerol, compared to growth kinetics at

233 elevated temperature, the kinetics drastically slows down (S1 Fig 7C). Co-culture bulk

234 competitions under simultaneous action of environmental conditions i.e. 42^0C + TMAO and 42^0C

235 + glycerol reveal drastic decrease in survivability of mutants. Interestingly, average fitness of

236 mutants is in case of 42^0C + TMAO condition is higher than that for 42^0C + glycerol condition.

237   Contrasting effects of the combinations of environments are consistent with the low

238   predictabilities fitness effects associate with simultaneous action of multiple environments.

239   In order to deduce relative influence of individual environments in the combinations, we

240   quantitated their selection metrics against individual constituent environments. Compared

241   against 42^0C alone (Fig 2A), both 42^0C + TMAO and 42^0C + glycerol conditions show weaker

242   differences (Fig 2A) than when compared against chemical environments (Fig 2B and C). This

243   suggests that the effects of combinations of environments closely resemble to treatment of

244   elevated temperature which has a strong influence in the combination. If considered as simple

245   additive effects, for both the cases of combination of environments, the resultant difference in

246   fitness do not seem to be a simple addition (Fig 2D and E). While addition of changes in fitness

247   scores for the constituents environments weakly correlate with the fitness changes of

248   combination of environments, among the two constituent chemical environments, glycerol

249   treatment produces lower predictability in terms of the additive model (r=0.54) than TMAO

250   (r=0.75). Collectively, while in terms of survivabilities of the mutants, elevated temperature has a

251   strong influence, explaining the non-additive fitness effects would need mechanistic insights.
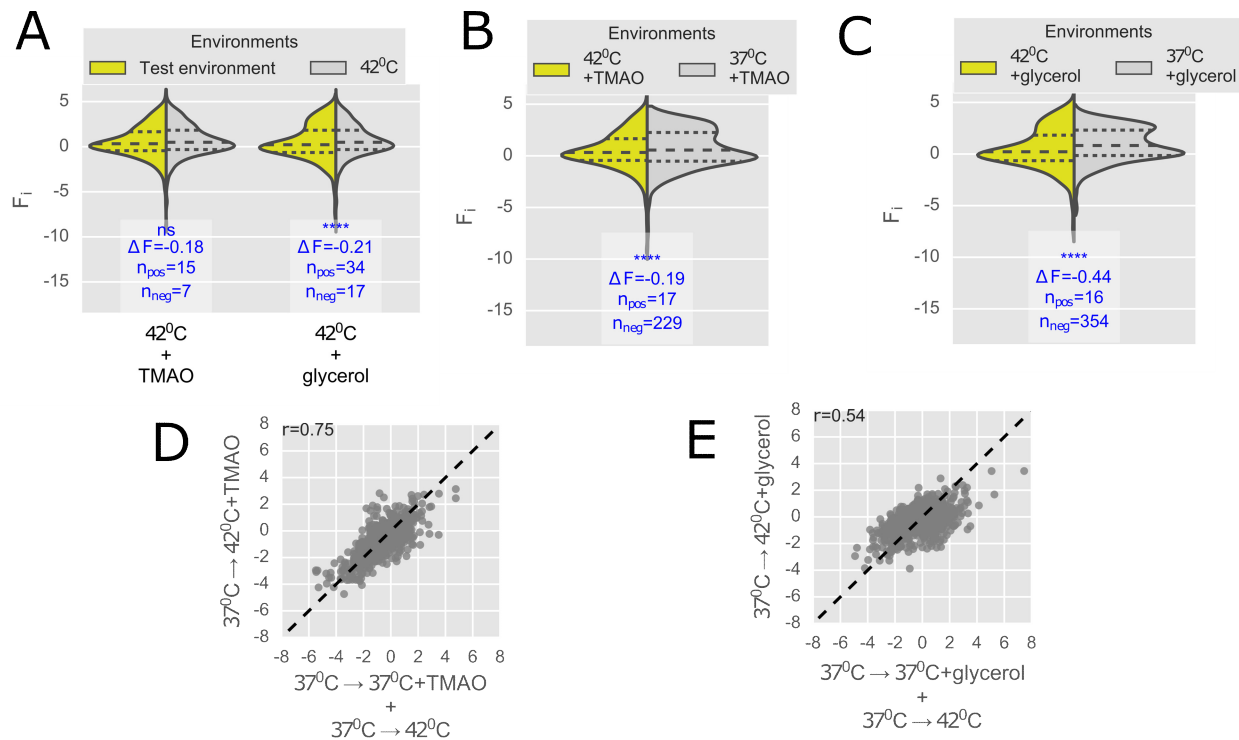
13

252

253

Fig 2. Selection strengths of combination of environments with respect to individual constituent environments. (A) In order to determine the relative selection strength of constituent elevated temperature condition in the combinations, DFEs of combination of environments i.e. 42^0C+TMAO and 42^0C+glycerol are compared against the DFE at 42^0C. Similarly, DFE obtained under 42^0C+TMAO condition is compared against TMAO condition (B) and DFE under 42^0C+ glycerol condition is compared against glycerol condition (C). F_i denotes fitness score. Histograms of DFEs are fitted by kernel density estimation. Heights of the DFEs are in scale with the frequencies of mutants. Dashed lines represent quartile values of the fitness per distribution. Level of significance is measured by two sided Wilcoxon signed-rank test. Relative selection pressure imposed by a test environment is with respect to control condition is measured in terms of difference between average fitness scores (ΔF) and number of mutants with higher (n_{pos}) or lower (n_{neg}) survivability than control environment. Correlations between change in fitness of mutants under combination of environments i.e. (D) 42^0C+TMAO

14

267   and (E) 42^0C+glycerol relative to reference environment with addition of fitness changes for

268   constituent environments relative to reference environment. Dashed line denotes equal values

269   of x and y variables. r is Pearson's correlation coefficient.

270

271   **Molecular constraints of underlying E2P effects**

272   Fitness effects would also depend upon constraints of multiple aspects such as structural,

273   folding and binding. Here, we used a set of molecular constraints describing multiple aspects of

274   gene fitness (S3 Table). Apart from measure of perturbation of stability by mutations ($\Delta\Delta$G), we

275   use residue depth as an additional feature representing folding component. Core residues are

276   known to play central role in protein stability (30); so mutations at the core of protein would be

277   less tolerated (lower fitness) than the ones at the surface of the protein. Other structural

278   features such as distances from dimer interface, residue flexibilities (B-factor obtained by PDB

279   structure) describe conformational constraints of the protein. Differences between physico-

280   chemical properties such as hydrophobicity (logP) and solvent accessibility of mutant and

281   reference amino acids account for mutational perturbations of non-covalent interactions.

282   Mutations near or at active site are more likely to have deleterious effects than mutations far

283   from active site. Using this principle, mutational perturbations of ligand binding are accounted

284   using a proxy of distance of residues of the protein from active site residue (D147) of the

285   protein. Here, minimum distance between the atoms of the D147 and C-alpha atom of the

286   residue is used to ensure maximum sensitivity. Conservation scores of the residues were

287   included for accounting for culmination of structural, folding and binding constraints in the form

288   of evolutionary rate per site.

289   Correlations of fitness scores with multiple molecular features quantitated in terms of

290   Spearman's rank correlation coefficients (ρ) reveal distinct classification of molecular constraints

15

291   (Fig 3). Euclidian clustering along the axis of molecular features separates folding and binding

292   constraints. Conservation scores, in particular, are best correlated with the fitness scores across

293   different environments. Among negatively correlated molecular features, correlations with

294   measures of protein stability estimates ($\Delta\Delta G$) (31) and residue depth reveal that mutations with

295   high stability perturbations or high residue depth show lower fitness and vice versa.

296   Contrastingly, distances from sensitive active site and dimer interface of the protein are

297   positively correlated with fitness scores. So mutations near active site or near dimer interface

298   indeed show lower fitness. Other than top five molecular constraints, structural constraints such

299   as residue flexibility, Δ(logP) per substitution and Δ(Solvent Accessible Surface Area) per

300   substitution are mostly negatively but relatively weakly correlated to fitness scores across

301   different environments. Overall, apart from conservation scores, folding and binding

302   components stand out as the strongest structural predictors of the fitness effects across all

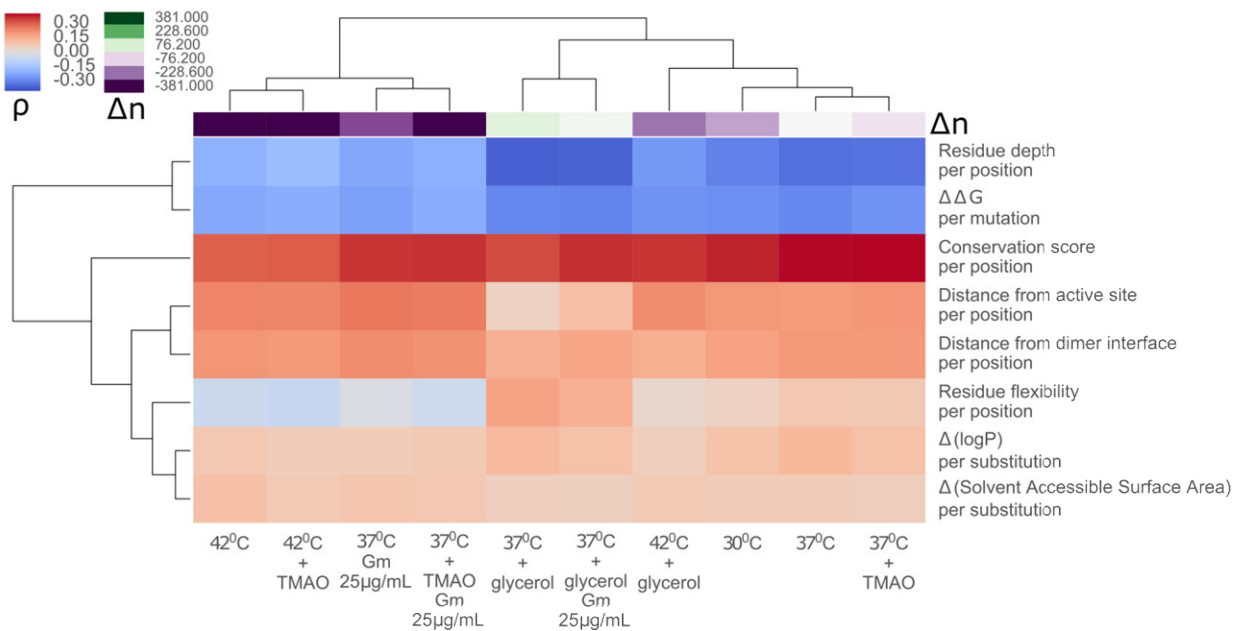303   environments; suggesting its major role in shaping the fitness landscape of GmR.

304

305 **Fig 3. Differential strengths of molecular constraints underlying environmental fitness**

306 **effects.** Fitness scores obtained under different environmental conditions are correlated with a

307 set of molecular features. Each locus in the heatmap indicates Spearman's rank correlation

308 coefficient between fitness scores of an environment (in columns) and a molecular feature (in

309 rows). ρ is Spearman's rank correlation coefficient and Δn is difference between number of

310 mutants survived in test environments with respect to reference environment (37^0C). Euclidean

311 clustering along rows and columns is based on the Spearman's rank correlation coefficients.

312

313 **Differential strengths of molecular determinants underlie environmental selection**

314 **pressures**

315 Alongside the classification of molecular constraints, Euclidean clustering (Fig 3) along the axis

316 of environments reveals a distinct classification which in turn aligns with their effects on

317 survivabilities i.e. Δn (Fig 3). Notably, correlation with conservation score is the highest in case

318 of reference condition (37^0C) than other environments; implying that the mutational tolerances in

319 this environment closely resemble to that existed in evolutionary history of the gene. This

320 supports our primary assumption to regard 37^0C as a reference environment. Molecular features

321 accounting for folding constraints i.e. $\Delta\Delta G$ and residue depth are best correlated in case of

322 environments with conserved (least changed) survivabilities suggesting that the environments

323 may induce allow survival of mutants with fitness changes that are aligned with the folding

324 constraints of the protein. In other words, folding constraint is likely to be stronger in such

325 environments than the environments with compromised survivabilities. Contrastingly, in terms of

326 distance from active site and dimer interface, fitness scores of mutants from environments with

327 lowered survivabilities are better correlated. This suggests that these factors particularly the

17

328   best correlated between the two i.e. distance from active site is potentially a major fitness

329   constraint.

330   Residue flexibility, interestingly, is negatively correlated with fitness scores of mutants from

331   environments with lowered survivabilities and positively correlated with majority of rest of the

332   environments. This implies that for environments that confer strongly deleterious fitness effects,

333   survived mutants with higher fitness scores are more likely to have lower residue flexibility at

334   mutation sites and vice versa. Elevated temperature conditions are known to promote

335   thermostability by reducing the conformational plasticity of the proteins (32,33); thus explaining

336   why residue flexibility emerges as a relatively strong constraint. Physico-chemical features of

337   the substitutions i.e. Δ(logP) per substitution and Δ(Solvent Accessible Surface Area) per

338   substitution, on the other hand, emerge as relatively weaker constraints; suggesting that

339   perturbations of non-covalent interactions have relatively minor role in guiding fitness of

340   mutants.

341   Although conservation score and molecular features describing folding and binding are seem to

342   be relatively strong constraints of fitness effects, individually, they are weakly correlated ($\rho<0.4$)

343   with the fitness scores. Assuming that the uncertainty in estimations of structural and predicted

344   features is considerably low and considering low experimental noise in determination of fitness

345   scores of GmR mutants, the weak correlations are consistent with known non-monotonic

346   relationship between fitness scores with biophysical constraints (13). Additionally, the weak

347   correlations can be a result of complex interplay among multiple factors such as well known

348   couplings between folding and binding components (34,35) or with other interdependent factors

349   (36). This exemplifies the underlying dependence of mutational tolerance on the inherent

350   structural constraints of the protein. Considering the possible complex interplay of set of

351   molecular constraints, we specifically focus on the strongest structural constraints i.e. folding

352   and binding for further contextualizing the environmental effects.

Dependence on folding and binding constraints

Protein folding and ligand binding are known to act as spandrels underlying fitness effects (34,35). Corroboratively, hereby, we demonstrated that the two factors potentially play central role in guiding the fitness of GmR mutants. Based on these molecular features, next, we made four subsets of mutants with distinct combinations of folding and binding states: proper folding and binding (FB), compromised folding and proper binding (cFB), proper folding and compromised binding (FcB) and compromised folding and binding (cFcB). Here, cF and cB indicates compromised folding (high $\Delta\Delta G$) and compromised binding (low distance from active site) respectively; while F and B denote non-compromised folding (high $\Delta\Delta G$) and non-compromised binding (high distance from active site) respectively. Median values of $\Delta\Delta G$ and distance from active site is used as threshold to bin the mutants in 4 equal sized subsets. For instance, subset cFcB represents mutants with $\Delta\Delta G$ values more than the median $\Delta\Delta G$ and distance from active site less than the median distances from active sites for all residues. Here, because the classification is based on values of molecular features inferred from crystal structure of GmR (37), to reduce the influence of the uncertainties involved, we exclude the mutations with values of molecular features falling between middle 20 percentiles from the classification. Notably, known important predictor of protein stability (30), residue depth is positively correlated with stability perturbation i.e. $\Delta\Delta G$ ($\rho=0.44$, S1 Fig 10). So mutants with high stability perturbation (cF) are highly likely to reside in the core regions of the protein while the ones low stability perturbation (F) are highly likely to lie on the surface of the protein.

Next, we calculated subset wise metrics of selection i.e. Δn and ΔF with respect to reference environment (37^0C, 12.5 µg/mL Gm). Firstly, in terms of survivability i.e. Δn, except for treatments of chemical chaperones at 12.5 µg/mL Gm and treatment of glycerol at 25 µg/mL Gm, for rest all the environments, survivability of the mutants is drastically decreased (Fig 4A). Notably, for environments with decrease in levels of survivabilities, across subsets, Δn seem to

378   show a pattern which is closely dependent on the binding and folding constraints. Survivability

379   of the mutants seem to decrease in a following order FB > cFB > FcB > cFcB. The greater

380   decrease in survivabilities with the compromise in folding and binding supports our primary

381   basis for classification of mutants and indicates that fitness scores of mutants are closely

382   dependent on folding and binding constraints.

383   In terms of change in average fitness (ΔF), with a clear exception of 37^0C Gm 25µg/mL, for

384   majority of environments average fitness of surviving mutants is greater than reference

385   environment (Fig 4B). Stringent Gm selection in addition to decreasing survivabilities of

386   mutants, seem to lower the fitness levels of mutants. Among subsets, surviving mutants from

387   subset FcB in particular have substantially lowered fitness scores potentially due increased

388   dependence on substrate binding and catalytic activity because of increased level of purifying

389   selection. Combination of environments too show lowered average fitness across subsets;

390   effect of which is particularly exemplified in FcB subset. For other environments, the average

391   fitness of surviving mutants seem to be increasing in the order FB < cFB < FcB < cFcB. This

392   suggests that the compromise of folding and binding allow exclusive survival of mutants with

393   fitness advantage. While mutants at the core of the protein and near to active site (cFcB) have

394   lowered survival, the survived mutants tend to possess relatively increased fitness advantage.

395   Elevated temperature conditions is an interesting example wherein the survivability of mutants is

396   drastically decreased (Fig 4A) and the environmental section only allowed survival of mutants

397   with fitness advantage (Fig 4B). Overall, FB subset is least affected while contrastingly cFcB is

398   the most affected subset by across all environments. Mutational robustness by TMAO and

399   glycerol is apparent by the retention survivability of mutants (Fig 4A) and increased fitness of

400   the mutants (Fig 4B). In combination with elevated temperature, both survivability and average

401   fitness of mutants is decreased. At stringent concentration of Gm, unlike TMAO treatment,

402   glycerol treatment confers mutational robustness. Collectively, distinct effects based on

403   underlying influence of folding and binding states reveal a unique pattern in survivabilities and

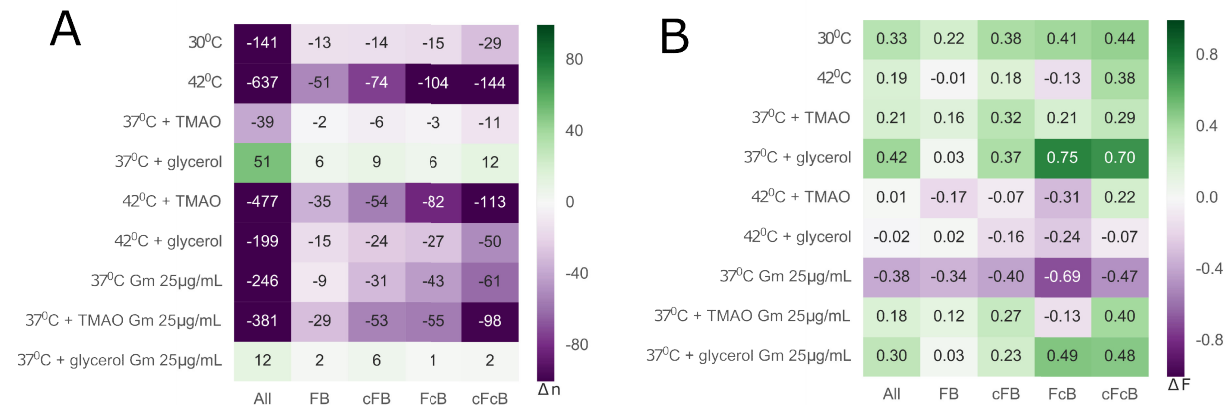404   relative changes in average fitness.

A

	All	FB	cFB	FcB	cFcB
30^{0}C	-141	-13	-14	-15	-29
42^{0}C	-637	-51	-74	-104	-144
37^{0}C + TMAO	-39	-2	-6	-3	-11
37^{0}C + glycerol	51	6	9	6	12
42^{0}C + TMAO	-477	-35	-54	-82	-113
42^{0}C + glycerol	-199	-15	-24	-27	-50
37^{0}C Gm 25µg/mL	-246	-9	-31	-43	-61
37^{0}C + TMAO Gm 25µg/mL	-381	-29	-53	-55	-98
37^{0}C + glycerol Gm 25µg/mL	12	2	6	1	2

Δn

B

	All	FB	cFB	FcB	cFcB
30^{0}C	0.33	0.22	0.38	0.41	0.44
42^{0}C	0.19	-0.01	0.18	-0.13	0.38
37^{0}C + TMAO	0.21	0.16	0.32	0.21	0.29
37^{0}C + glycerol	0.42	0.03	0.37	0.75	0.70
42^{0}C + TMAO	0.01	-0.17	-0.07	-0.31	0.22
42^{0}C + glycerol	-0.02	0.02	-0.16	-0.24	-0.07
37^{0}C Gm 25µg/mL	-0.38	-0.34	-0.40	-0.69	-0.47
37^{0}C + TMAO Gm 25µg/mL	0.18	0.12	0.27	-0.13	0.40
37^{0}C + glycerol Gm 25µg/mL	0.30	0.03	0.23	0.49	0.48

ΔF

405

406   **Fig 4: Survivability and fitness of mutants show dependence on perturbation of folding**

407   **and binding.** Subset-wise relative selections imposed by test environments with respect to

408   reference environment i.e. 37^{0}C at 12.5µg/mL Gm is measured in terms of (A) difference in

409   number of survived mutants (Δn) and (B) difference in average fitness of mutants (ΔF) with

410   respect to that of reference environment. 'All' denotes all mutations in a given the environmental

411   condition. Among subsets of mutants (FB, cFB, FcB and cFcB), F and B represent mutants with

412   low $\Delta\Delta$G and high distance from active site respectively, whereas cF and cB represent mutants

413   with high $\Delta\Delta$G and low distance from active site respectively.

414

Unique reshaping of environment specific fitness landscape

416   In order to better contextualize the contributions of coupling between folding and binding, we

417   visualized the fitness landscape of GmR by lining it with the folding and binding components

418   which evidently act as central constraints of mutational fitness (Fig 5A). So, regimes at the

21

419   corners of the fitness landscape represent the subsets of mutants with different states of folding

420   and binding i.e. FB, cFB, FcB and cFcB.

421   For reference environment, as expected from corresponding folding and binding states, we find

422   that majority of high fitness mutations occupy FB regime whereas cFcB regime is majorly

423   occupied by the deleterious mutants (Fig 5A). Notably, this pattern is conserved across all the

424   test environments (Fig 5B-J). For reference environment, folding constraint produces a

425   pronounced fitness cliff whereby mutants with above a threshold of $\Delta\Delta G$ (~2 kcal/mol) are highly

426   likely to show deleterious fitness. In case of selection at stringent Gm concentration,

427   corroborating with dosage dependent effects, mutations close to the active site show a

428   prominent decrease in fitness (Fig 5B). Here, the imposed higher load of Gm, seem to generate

429   additional pronounced fitness cliff along the binding axis. The substantial change in the fitness

430   landscape as a function of selection levels is consistent with known central role of purifying

431   selection in molecular evolution (21).

432   Among physical environments, fitness landscape corresponding to low temperature condition

433   show no peculiar regime wise distinction from that of reference environment (Fig 5C); capturing

434   earlier noted weaker selection pressure (Fig 1C). Elevated temperature conditions - in

435   accordance with imposed negative selection pressure - show selective survival of only mutants

436   with enriched fitness associated with lowered survival of mutants from cFB and cFcB subsets

437   (Fig 5D). Among chemical environments, mutational robustness imposed by TMAO and glycerol

438   is evident in the form of close similarity with the fitness landscape of reference environment (Fig

439   5E,G). At stringent Gm concentration, chemical chaperone seem to increase the fitness of

440   survived mutants lying at cFcB regime (Fig 5F,H) which are otherwise highly depleted (Fig 5B).

441   For combinations of environments, in alignment with in previous analyses, fitness landscapes

442   (Fig 5I,J) show substantially decreased survivabilities and contrasting average fitness levels i.e.

443   particularly at cFcB regime glycerol treatment depletes the fitness of mutants more intensely

444 than TMAO treatment. Overall, the distinct reshapings of the fitness landscapes under different

445 environments reveal divergent fates and preferences for mutations of gene guided by differential

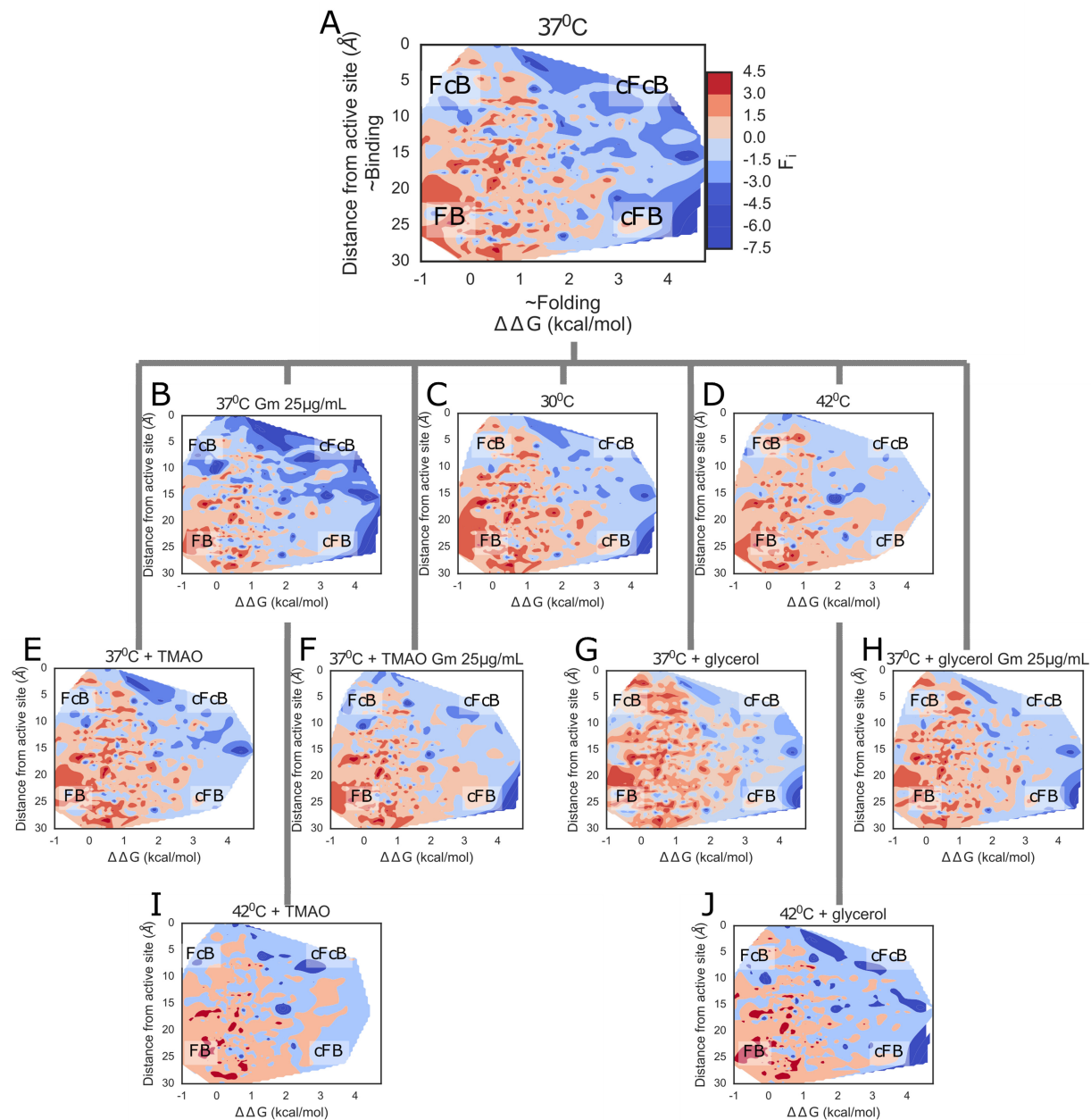446 strengths of key molecular constraints.

447

Fig 5. Unique reshapings of environment specific fitness landscapes. Environment specific

449 fitness effects are visualized in a form of a fitness landscapes (A-J) lined by folding ($\Delta\Delta G$) and

450 binding (distance from active site) components with Z-component of the plot scaled by fitness

23

451   levels (F_i). Contour surfaces are generated by nearest neighbor interpolation. Regimes at the

452   corners of the fitness landscapes thus represent subsets of mutants based on folding and

453   binding components i.e. FB, cFB, FcB and cFcB. Colors of all contour plots are scaled

454   according   to   the   colorbar   associated   with   panel   A.

24

455

Discussion

457 Large-scale elucidation of G2P interactions enabled by high throughput mutational scanning

458 (38) has opened up new possibilities to comprehensively assess fundamental questions in

459 molecular evolution. Here we investigate molecular level E2P interactions and their underlying

460 mechanistic insights. Upon monitoring empirical fitness of a library of single site mutants of

461 GmR, we characterized relative selection pressures imposed by sets of physical and chemical

462 environments (Fig 1C). Central role of maintenance of cellular proteostasis is exemplified by the

463 directionalities of the selection pressures which show a close dependence on the potential

464 alterations of protein folding. This also supports the identification of stability perturbation and

465 other protein folding related constraints as the one of the strongest constraints of fitness of GmR

466 mutants.

467 Through our data, as in case of mutational scanning's of other antibiotic resistant genes

468 (21,22,39), we find that the changes in fitness landscapes of GmR are dependent on purifying

469 selection level i.e. the dosage of antibiotic (Fig 1B). Corroborating with earlier studies (11,12),

470 we demonstrate that mutational landscapes monitored under different environmental alterations

471 of proteostasis are significantly different than that monitored under reference environmental

472 conditions (Fig 1C). Among such environments, elevated temperature (42^0C) exerts negative

473 selection pressure underscoring known protein misfolding effects (28) and temperature

474 sensitivity (17). Low temperature (30^0C) conditions impose weak mutational robustness

475 conforming to known non-deleterious effects on protein folding (40). Chemical chaperones, too,

476 conforming with their known favorable effects on protein folding (18,19), exert mutational

477 robustness. This shows that fitness effects in each environment on fitness landscape for GmR is

478 dependent on kind of environmental alteration of cellular protein folding.

25

479   Remarkably, simultaneous action of environments, specifically environments with oppositely

480   directed selection pressures i.e. elevated temperature and chemical chaperones are found to

481   lead to lower predictable outcomes which underscore similar observation reported previously

482   (11). In our study, estimation of selection pressures against constituent individual environments

483   revealed that strongest among the constituent environments guide the overall selection

484   pressure in the combination (Fig 2A). However in terms of average fitness, they have

485   contrasting (Fig 1C) and non-additive effects (Fig 2D and E). The contrasting effects of chemical

486   chaperones exemplified at the elevated temperature conditions may be a result of inherent

487   mechanistic differences in thermodynamic assistance to the folding of proteins i.e. entropic

488   assistance by TMAO and enthalpic assistance by glycerol (19). Additionally, effects of chemical

489   chaperones seem to be a function of growth temperature i.e. at 37^0C, supplementation of

490   chemical chaperones lead to increase in average fitness of mutants while that at elevated

491   temperature supplementation of chemical chaperones lead to decrease in average fitness of

492   mutants. Such effects could be due to simultaneous selection for multiple traits which

493   complicates the resultant selection pressure both qualitatively and quantitatively (29).

494   Next, in order to gain mechanistic insights into of the evident E2P interactions, we

495   contextualized them in terms of molecular constraints of the fitness landscapes. Correlations of

496   fitness scores with molecular constraints reveal their differential strengths for each environment

497   (Fig 3). Conservation score - accounting for evolutionary rate per site - is best correlated with

498   the fitness scores across all the test environments; thus substantiating the biological

499   significance of the obtained empirical fitness values. Next to conservation score, stability

500   perturbation ($\Delta\Delta$G), and perturbation in binding (distance from active site) are best correlated

501   with the fitness scores. Notably, however the strength of correlations are relatively weaker

502   (ρ<0.4); suggesting possibility of non-monotonic relationships (13) and a complex interplay of

503   interdependent constraints.

504   In our study, coupled factors, folding and binding constraints distinctly emerge as strong

505   determinants of fitness effects; and through their spandrel like property (34). To understand the

506   trade-off between the two, we visualized fitness of GmR mutants in a form of fitness landscape

507   lined by folding (ΔΔG) and binding axes (distance from active site) (Fig 5A). Across all

508   environments, corroborating with known trade-off in case of antibiotic resistant genes between

509   folding and binding (35), combinations of folding and binding states seem to underlie resulting

510   fitness effects. For instance, combination of weaker folding and binding constraints (FB) is

511   associated with largely enriched fitness levels while stringent folding and binding constraints

512   (cFcB) are associated with deleterious fitness levels of mutants (Fig 5A). In case of reference

513   environment, folding constraint introduces a prominent limiting fitness cliff (at ΔΔG=~2 kcal/mol).

514   Whereas for test environments, although at variable degrees, common existence of limiting

515   fitness cliffs along folding axis (Fig 5) underscores central role of protein folding and stability in

516   molecular evolution (41). The universal dependence on folding component especially at minimal

517   purifying selection also explains the conformity of evident fitness effects with known effects of

518   environments on protein folding and proteostasis. Alongside, this study therefore reveals a

519   unique possibility of controlling mutational fates based on environmental alteration of major

520   constraints of fitness landscape. Notably, binding constraint imposes a limiting fitness cliff which

521   is more pronounced at stringent concentration of Gm (25 μg/mL) (shown in Fig 5B) and from our

522   data, it seems to be a feature dependent on level of purifying selection levels.

523   Collectively, from a simple experimental system consisting of a conditionally essential gene with

524   evidently weaker constraints imposed by events prior to protein expression, we identify

525   environment dependent differential fitness of mutations which are dependent on relative

526   strengths of underlying molecular constraints. Major implication of this information lies in the

527   improvement of our understanding of influence of environmental conditions on G2P interactions

528   at molecular level. Further, environmentally induced variability can potentially contribute a

529   requisite nuance in predictive models used for challenging task of inferring phenotypic outcomes

530   from genomic variants(15,16); making them more robust. In future, such study of

531   comprehensive environment specific fitness landscapes can be potentially extended to multiple

532   mutations to monitor combined effects of epistasis - arguably the major predictive factor in

533   molecular evolution (42), - more variety of environments as well as assessing fitness

534   landscapes of multiple proteins in gene regulatory networks.

Materials and Methods

Minimal inhibitory concentration (MIC) assays

Primary culture was prepared by inoculating (1% v/v) *E. coli* (K-12) in culture media (Luria-Bertani (LB) broth (HiMedia) containing 100μg/mL, ampicilin (Sigma) and 0.1% Arabinose (Sigma)) and incubating at 37^0C for 18 hrs. The primary culture was inoculated at OD_{600} of 0.025 in culture media containing a range of Gm (Sigma) concentrations from 6.25 to 400 μg/mL with 2 fold increase at each increment (in 96-well storage plates). The assay plates were incubated at 37^0C for 18 hrs before measuring growth (OD_{600}) in Tecan microwell plate reader.

Growth assays

E. coli (K-12) harboring pBAD-GmR is grown in culture media (LB media containing 100μg/mL and ampicilin 0.1% Arabinose) for ~18 hr. Primary culture was used as an inoculum (~0.01 OD) for the growth assays. Growth assays in different environments were carried out using Bioscreen C kinetic growth reader. The growth parameters were obtained by fitting absorbance data to five parameter Logistic equation.

Co-culture bulk competition assay

SSM library of GmR was constructed by PCR based site directed mutagenesis using primers with degenerate codons (NNK). For detailed information regarding the mutagenesis, please refer to Supporting methods described in Bandyopadhyay et al. (18). For co-culture bulk competition assays, the mutation library cloned in pBAD vector was transformed into *E. coli* (K-12). Primary culture was prepared by inoculating pool of SSM library (1% v/v) in culture media (LB media containing 100μg/mL ampicilin and 0.1% Arabinose) at 37^0C for 18 hrs. A competition was carried out at the secondary culture where primary culture in inoculated at OD_{600} of 0.025 and incubated for 18 hrs. Physical environmental conditions were created by carrying out the

29

558   bulk competitions at 30^0C (low temperature) or 42^0C (elevated temperature). Chemical

559   environmental conditions were created by supplementing either TMAO (250mM) or glycerol

560   (250mM) in the culture media of competition assay. Biological replicates were made by carrying

561   out independent co-culture bulk competitions of the mutant libraries. For measuring fitness of

562   mutants in a particular environmental condition, bulk competition under Gm selection (selected

563   pool) (as shown in Fig 1A) was carried out. An independent bulk competition was carried out at

564   37^0C in the absence of Gm (unselected pool) which serves as a reference for calculating

565   preferential enrichments.

Deep sequencing

567   At the end of bulk competition assays, cells are pelleted and plasmid is purified. Amplicons were

568   generated by a short PCR (initial denaturation: 95ºC for 3 min, denaturation: 95ºC for 1 min,

569   annealing: 60ºC for 15 sec, extension: 72ºC for 1 min, final extension: 72ºC for 10 min) using

570   high fidelity KAPA HiFi DNA polymerase (cat. no. KK2601). High template concentration (1

571   ng/µl) and 20 cycles were used to reduce potential PCR bias. Multiplexing was carried out using

572   flanking barcoded primers (4 forward, 4 reverse, sequences in S1 Table 2). Amplicons of

573   barcoded samples were grouped in equimolar concentration and gel purified. A dual index

574   library for each such set was prepared using Truseq PCR-free DNA HT kit (Illumina Inc. Cat no.

575   F-121-3003) and sequenced using paired end (300 X 2) chemistry on Illumina Miseq platform.

576   Raw sequencing data is available upon request.

Estimation of fitness scores from deep sequencing data

578   Analysis of sequencing data was carried out by using *dms2dfe* (43) - a comprehensive analysis

579   pipeline exclusively designed for analysis of deep mutational scanning data. Through *dms2dfe*

580   workflow, output files from the sequencer (.fastq) were demultiplexed using *ana0_fastq2dplx*

581   module of *dms2dfe*. Average read depth of each demultiplexed sample was ~1X10^5. Next,

30

582   though *dms2dfe*'s modules namely *ana0_fastq2sbam*, sequence alignment was carried out

583   using Bowtie2 (44) followed by variant calling through *ana1_sam2mutmat* module which utilizes

584   *pysam* libraries (45). A variant is called only if average Q-score of the read and that of the

585   mutated codon is more than 30. Additionally a cut off of 3 reads per variants is used to filter out

586   anomalous low counts. As a result a codon level mutation matrix of counts of mutations is

587   generated. Codon level mutation matrix is then translated to amino acid level (based on the

588   codon usage bias of the *E. coli*). For each experimental condition, counts of ~2000 individual

589   mutants were quantified (S2 Table). Raw sequencing data is available at Sequence Read

590   Archive (SRA) as a BioProject: PRJNA384918.

591   Through *ana2_mutmat2fit* module of *dms2dfe*, counts of mutants are first normalized by the

592   depth of sequencing at each position of the gene. Then preferential enrichments which are log

593   (base 2) fold change of counts of the mutants in pool selected in presence of Gm against

594   unselected (0 µg/mL Gm) reference pool are estimated. Here, preferential enrichment of a

595   mutant serves as a proxy for its relative fitness and hence we simply refer it as 'fitness'.

596   Alongside significance levels of the preferential enrichments are obtained by Wald test through

597   DeSeq2 workflow (46) (S2 Table).

598   **Classification of mutants based on comparison of DFEs**

599   In order to assess the difference in fitness scores in a given comparison of DFEs of two

600   environments, we assigned a statistical threshold to reduce the influence of experimental and

601   biological noise. The mutants with higher fitness scores in test environments than control

602   environment are classified as 'positive' and while mutants with lower fitness scores in test

603   environments than control environment are classified as 'negative'. We use replicates of test

604   and control environments to get thresholds of difference in fitness scores such that the

605   differences in fitness scores between the environments ('signal') is greater than the differences

606   within the replicates ('noise'). From difference between fitness scores of replicates, two

607   distributions corresponding to replicates of test and control environment are obtained. μ_{test} and

608   $\mu_{control}$ are means and σ_{test} and $\sigma_{control}$ are standard deviations of such distributions for test and

609   control environments respectively. So, the thresholds to segment the mutations into 'positive' or

610   'negative' is determined to be equal to $\frac{\mu_{test}+\mu_{control}}{2} \pm 2 * \sqrt{\frac{\sigma_{test}^2+\sigma_{control}^2}{2}}$. Number of mutants

611   undergoing positive effects is denoted as $n_{pos,}$ while number of mutants with negative effects is

612   denoted as n_{neg}.

Molecular features of GmR

614   Mutant stability perturbations ($\Delta\Delta G$) are predicted by PoPMusic (31) server. Conservation score

615   is acquired from ConSurf (7) server. *MSMS* libraries (47) were used for calculations of residue

616   depth from surface of protein. Distances between atoms of GmR are measured using various

617   modules of Biopython package (48). Physico-chemical properties of the amino acids such as

618   logP and pI were retrieved from PubChem (49) and ChemAxon (http://www.chemaxon.com).

619   Molecular features of mutations used in the study are included in S3 Table.

620

Supporting information

621

S1 Text. Supporting Figures and Tables.

622

S2 Table. Fitness scores of mutations in different environments.

623

S3 Table. Molecular features of mutations.

624

Abbreviations

625

G2P : Genotype to Phenotype

626

E2P : Environment to Phenotype

627

Gm : Gentamicin;

628

GmR : Gentamicin Resistant gene;

629

DMS : Deep Mutational Scanning;

630

DFE : Distribution of Fitness Effects;

631

Competing interests

632

The authors declare that they have no competing interests.

633

Authors' contributions

634

KC designed the study. RD carried out DMS including bulk competitions, DNA purification and

635

sequencing. RD and KC carried out the analysis of sequencing data. RD and KC wrote the

636

manuscript. R.P and G.J. helped in the amplicon sequencing. All authors read and approved the

637

final manuscript.

638

Acknowledgments

639

643

References

1. Harris DR, Pollock S V, Wood EA, Goiffon RJ, Klingele AJ, Cabot EL, et al. Directed evolution of ionizing radiation resistance in Escherichia coli. J Bacteriol. 2009;191(16):5240–52.

2. Zhang Q, Lambert G, Liao D, Kim H, Robin K, Tung C, et al. Acceleration of emergence of bacterial antibiotic resistance in connected microenvironments. Science [Internet]. 2011 Sep;333(6050):1764–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21940899

3. Minty JJ, Lesnefsky A a, Lin F, Chen Y, Zaroff T a, Veloso AB, et al. Evolution combined with genomic study elucidates genetic bases of isobutanol tolerance in Escherichia coli. Microb Cell Fact [Internet]. 2011;10(1):18. Available from: http://www.microbialcellfactories.com/content/10/1/18

4. Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, et al. The molecular divergence of adaptive convergence. Science (80-). 2012;457.

5. Blaby IK, Lyons BJ, Wroclawska-Hughes E, Phillips GCF, Pyle TP, Chamberlin SG, et al. Experimental evolution of a facultative thermophile from a mesophilic ancestor. Appl Environ Microbiol. 2012;78(1):144–55.

6. Powers ET, Morimoto RI, Dillin A, Kelly JW, Balch WE. Biological and chemical approaches to diseases of proteostasis deficiency. Annu Rev Biochem. 2009;78(August):959–91.

7. Lynch M. Evolution of the mutation rate. Trends Genet [Internet]. 2010;26(8):345–52. Available from: http://dx.doi.org/10.1016/j.tig.2010.05.003

8. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods [Internet]. 2010;7(9):741–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2938879

9. Araya CL, Fowler DM. Deep mutational scanning: Assessing protein function on a massive scale. Trends Biotechnol [Internet]. 2011;29(9):435–42. Available from: http://dx.doi.org/10.1016/j.tibtech.2011.04.003

10. Eyre-Walker a, Keightley P. The distribution of fitness effects of new mutations. Nat Rev Genet [Internet]. 2007;8(8):610–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17637733

11. Hietpas RT, Bank C, Jensen JD, Bolon DNA. Shifting fitness landscapes in response to altered environments. Evolution (N Y). 2013;67(12):3512–22.

12. Mavor D, Barlow K, Thompson S, Barad BA, Bonny AR, Cario CL, et al. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. Elife [Internet]. 2016;5:1–23. Available from: http://elifesciences.org/lookup/doi/10.7554/eLife.15802

13. Boucher JI, Bolon DNA, Tawfik DS. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. Protein Sci. 2016;25:1219–26.

684   14.   Liberles DA, Teufel AI, Liu L, Stadler T. On the need for mechanistic models in
685   computational genomics and metagenomics. Genome Biol Evol. 2013;5(10):2008–18.

686   15.   Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence
687   variants. BMC Genomics [Internet]. 2015;16(Suppl 8):S1. Available from:
688   http://www.biomedcentral.com/1471-2164/16/S8/S1

689   16.   Shendure J, Fields S. Massively Parallel Genetics. Genetics. 2016;203(June):617–9.

690   17.   Varadarajan R, Nagarajaram H a, Ramakrishnan C. A procedure for the prediction of
691   temperature-sensitive mutants of a globular protein based solely on the amino acid
692   sequence. Proc Natl Acad Sci U S A. 1996;93(24):13908–13.

693   18.   Bandyopadhyay A, Saxena K, Kasturia N, Dalal V, Bhatt N, Rajkumar A, et al. Chemical
694   chaperones assist intracellular folding to buffer mutational variations. Nat Chem Biol
695   [Internet]. 2012;8(3):238–45. Available from:
696   http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3527004

697   19.   Dandage R, Bandyopadhyay A, Jayaraj GG, Saxena K, Dalal V, Das A, et al.
698   Classification of Chemical Chaperones Based on Their Effect on Protein Folding
699   Landscapes. ACS Chem Biol [Internet]. 2015 Mar;10(3):813–20. Available from:
700   http://pubs.acs.org/doi/abs/10.1021/cb500798y

701   20.   Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet
702   [Internet]. 2013;14(12):827–39. Available from:
703   http://www.ncbi.nlm.nih.gov/pubmed/24166031

704   21.   Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a Function of Purifying Selection
705   in TEM-1 β-Lactamase. Cell [Internet]. 2015;160(5):882–92. Available from:
706   http://dx.doi.org/10.1016/j.cell.2015.01.035

707   22.   Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A comprehensive, high-resolution map
708   of a Gene's fitness landscape. Mol Biol Evol. 2014;31(6):1581–92.

709   23.   Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational
710   scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic
711   Acids Res. 2014;42(14):1–8.

712   24.   Ramsey DC, Scherrer MP, Zhou T, Wilke CO. The relationship between relative solvent
713   accessibility and evolutionary rate in protein evolution. Genetics. 2011;188(2):479–88.

714   25.   Scherrer MP, Scherrer MP, Meyer AG, Meyer AG, Wilke CO, Wilke CO. Modeling coding-
715   sequence evolution within the context of residue solvent accessibility. BMC Evol Biol
716   [Internet]. 2012;12(1):179. Available from: http://www.biomedcentral.com/1471-
717   2148/12/179,

718   26.   Franzosa EA, Xia Y. Independent Effects of Protein Core Size and Expression on
719   Residue-Level Structure-Evolution Relationships. PLoS One. 2012;7(10).

720   27.   Grimont F, Grimont PAD. The Genus Serratia. In: The Prokaryotes: Volume 6:
721   Proteobacteria: Gamma Subclass [Internet]. 2006. p. 219–44. Available from:
722   http://dx.doi.org/10.1007/0-387-30746-X

723   28.   Lindquist S. THE HEAT-SHOCK RESPONSE. Ann Rev Biochem. 1986;55:1151–91.

724 29. Haldane JBS. The cost of natural selection. J Genet. 1957;55(3):511–24.

725 30. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of
726   protein structure and stability. Structure [Internet]. 1999;7(7):723–32. Available from:
727   http://www.ncbi.nlm.nih.gov/pubmed/10425675

728 31. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the
729   estimation of protein stability changes upon mutation and sequence optimality. BMC
730   Bioinformatics [Internet]. 2011;12(1):151. Available from:
731   http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3113940

732 32. Vihinen M. Relationship of protein flexibility to thermostability. Protein Eng.
733   1987;1(6):477–80.

734 33. Reetz MT, Carballeira JD, Vogel A. Iterative saturation mutagenesis on the basis of b
735   factors as a strategy for increasing protein thermostability. Angew Chemie - Int Ed.
736   2006;45(46):7745–51.

737 34. Manhart M, Morozov A V. Protein folding and binding can emerge as evolutionary
738   spandrels through structural coupling. Proc Natl Acad Sci U S A. 2015;112(6):1797–802.

739 35. Wang X, Minasov G, Shoichet BK. Evolution of an antibiotic resistance enzyme
740   constrained by stability and activity trade-offs. J Mol Biol. 2002;320(1):85–95.

741 36. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Science. 2016. p. 1204–18.

742 37. Wolf E, Vassilev a, Makino Y, Sali a, Nakatani Y, Burley SK. Crystal structure of a
743   GCN5-related N-acetyltransferase: Serratia marcescens aminoglycoside 3-N-
744   acetyltransferase. Cell [Internet]. 1998;94(4):439–49. Available from:
745   http://www.ncbi.nlm.nih.gov/pubmed/9727487

746 38. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat
747   Methods [Internet]. 2014;11(8):801–7. Available from:
748   http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4410700

749 39. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational
750   scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic
751   Acids Res. 2014;42(14).

752 40. Baneyx F, Mujacic M. Recombinant protein folding and misfolding in Escherichia coli. Nat
753   Biotechnol [Internet]. 2004;22(11):1399–408. Available from:
754   http://www.nature.com/nbt/journal/v22/n11/full/nbt1029.html,

755 41. Zeldovich KB, Chen P, Shakhnovich EI. Protein stability imposes limits on organism
756   complexity and speed of molecular evolution. Proc Natl Acad Sci U S A [Internet].
757   2007;104(41):16152–7. Available from:
758   http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2042177

759 42. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the
760   primary factor in molecular evolution. Nature [Internet]. 2012;490(7421):535–8. Available
761   from: http://dx.doi.org/10.1038/nature11510

762 43. Dandage R, Chakraborty K. dms2dfe : Comprehensive Workflow for Analysis of Deep
763   Mutational Scanning Data. bioRxiv [Internet]. 2016;72645. Available from:

764   http://biorxiv.org/content/early/2016/08/31/072645.abstract

765   44.   Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods.
766   2012;9(4):357–9.

767   45.   Heger A. Pysam [Internet]. github.com. 2009. Available from: https://github.com/pysam-
768   developers/pysam

769   46.   Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
770   RNA-seq data with DESeq2. Genome Biol [Internet]. 2014;15(12):550. Available from:
771   http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8

772   47.   Sanner MF, Olson a J, Spehner JC. Reduced surface: an efficient way to compute
773   molecular surfaces. Biopolymers. 1996;38(3):305–20.

774   48.   Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely
775   available Python tools for computational molecular biology and bioinformatics.
776   Bioinformatics. 2009;25(11):1422–3.

777   49.   Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance
778   and compound databases. Nucleic Acids Res. 2016;44(D1):D1202--D1213.

779