SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
SYSTEMS BIOLOGY
COMPUTATIONAL BIOLOGY
BIOINFORMATICS
APPLIED PHYSICS

Received
20 June 2012

Accepted
6 August 2012

Published
29 August 2012

Correspondence and requests for materials should be addressed to M.Á.S. (marian.serrano@ub.edu)

Predicting effects of structural stress in a genome-reduced model bacterial metabolism

Oriol Güell[1], Francesc Sagués[1] & M. Ángeles Serrano[2]

[1]Departament de Química Física, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain, [2]Departament de Física Fonamental, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain.

Mycoplasma pneumoniae **is a human pathogen recently proposed as a genome-reduced model for bacterial systems biology. Here, we study the response of its metabolic network to different forms of structural stress, including removal of individual and pairs of reactions and knockout of genes and clusters of co-expressed genes. Our results reveal a network architecture as robust as that of other model bacteria regarding multiple failures, although less robust against individual reaction inactivation. Interestingly, metabolite motifs associated to reactions can predict the propagation of inactivation cascades and damage amplification effects arising in double knockouts. We also detect a significant correlation between gene essentiality and damages produced by single gene knockouts, and find that genes controlling high-damage reactions tend to be expressed independently of each other, a functional switch mechanism that, simultaneously, acts as a genetic firewall to protect metabolism. Prediction of failure propagation is crucial for metabolic engineering or disease treatment.**

The architecture of complex networks is imprinted with universal features that affect their resilience and condition their behavior[1–3]. Most relevant, the scale-free connectivity of many natural and man-made networks explains their fragility in front of attacks to the most connected nodes, while they are able to deal with accidental failures of single components[4,5]. A manifestation of this fragile yet robust nature of complex networks is that the failure cascade triggered by a local shock rarely propagates to the whole system[6–9]. Yet, network studies have mainly focused on single node failures, and systemic responses to more globalized forms of structural and functional stress still remain to be explored.

In the biological context, networks of molecular interactions in the cell are among the best probed in terms of robustness in front of a variety of *in silico* perturbation experiments. They have been found to comply with the design principles of error-tolerant scale-free networks[10], and recent progress in network dynamics is also starting to portray the concept of stress-induced network rearrangements[11,12]. Interestingly, metabolic networks offer an excellent arena for network stress testing and prediction, due to the amount and quality of the experimental data underlying their genome-scale reconstructions[13] which enable reliable complex network representation and analysis[14–18]. In this context, the exploration of single biochemical reaction inactivations has shown that the structural organization of metabolic networks reduces the likelihood of large damaging cascades[19]. At the same time, many individual mutations that eliminate enzyme-coding genes seem to have very little effect on cell growth[20,21]. By contrast, the impact of multiple failures could go beyond the mere accumulation of individual effects, producing amplified damage due to peculiar biochemical interweaving or gene epistatic interactions[22].

In this work, we explore and predict the effects of different forms of structural stress on the robustness of the metabolic network of *Mycoplasma pneumoniae*, a human pathogen that has recently been proposed as a genome-reduced model organism for bacterial and archaeal systems biology[23–25]. Our analysis considers the removal of single and pairs of biochemical reactions and the knockout of individual genes and clusters of co-expressed genes. We found that this organism exhibits network responses similar to those of larger model organisms, like *Escherichia coli* or *Staphylococcus aureus*, although its increased linearity and reduced redundancy[23] threaten its robustness against individual reaction removals. For all three organisms, we show that the impact of failure cascades spreading through the metabolic network can be predicted in terms of local network motifs. In this way, targets prone to introduce structural vulnerability can be readily detected prior to experimental testing without expensive computations, even for larger and more complex organisms. For *M. pneumoniae*, we also explored the effects of single and multiple gene knockouts by coupling, through enzyme activity, its metabolic network to the

experimentally extracted gene co-expression network. We observed that genes related to high-damage reactions are essential for the organism and that their expression tends to be isolated from that of other genes. This hints at the interplay between metabolism and genome, apparently evolved to favor the robustness of this organism by avoiding the potentially catastrophic effect of coupling the co-expression of structurally vulnerable metabolic genes.

Results

Predicting damage spreading in metabolic structure. We modeled metabolic networks as bipartite graphs with two different types of nodes, metabolites and reactions, connected by directed links. For irreversible reactions, the directionality of the flux determines the directionality of the links, while reversible reactions are treated as two coupled reactions to account for the forward and reverse fluxes. The failure of a reaction may turn some of its metabolites inviable if they cannot be maintained anymore at steady non-zero concentration. This happens when they are left without producing or consuming reactions (except for metabolites that the organisms exchange with the environment) or, in topological terms, without incoming or outgoing connections. On its turn, an inviable metabolite makes non-operational all associated reactions. This mechanism[19] propagates a failure cascade that stops when all reactions remaining in the network are viable. At this point, the corresponding damage is quantified as the number of reactions turned non-operational (see Supporting Information).

Next, we present the distribution of damages for cascades triggered by individual and by pairs of reactions in the metabolic network of *M. pneumoniae*, as compared to *S. aureus* and *E. coli* (in Materials and Methods we provide generic characteristics extracted from the databases for the three considered organisms). We later identify network motifs responsible for the propagation of cascades, and propose a local predictor for damage.

Genome-scale impact of individual reactions failures. Although close to 50% of all individual reaction failures in the three organisms considered did not propagate cascades, and most cascades were indeed small (59% of the cascades in *M. pneumoniae*, 38% in *S. aureus*, and 55% in *E. coli* propagate to only one or two reactions), the removal of some particular reactions may trigger relatively far reaching damages. This is shown in Figs. 1a–c, that display the cumulative probability distributions $P(d_r' \geq d_r)$ that the failure of a reaction r attains at least $d_r - 1$ other reactions in each metabolic network (see also Supporting Information). All species show similar broad distributions, although the crossover in the tail of the distribution from power-law-like to exponential-like is not evident in *M. pneumoniae* probably due to its limited redundancy (see Materials and Methods). In order to assess the significance of cascades, the computed distributions are compared with those corresponding to degree preserving randomized variants of the metabolic networks taken as null models (see Supporting Information). To check consistency, we perform Kolmogorov-Smirnov (K-S) tests[26] measuring the maximum absolute difference between the corresponding distributions (see caption of Figure 1 for specific values). This difference is transformed into a significance level directly compared to a chosen threshold, typically $\alpha = 0.05$. If the significance associated to the K-S test statistic is equal or smaller than α, the compared distributions cannot be considered consistent. Both *E. coli* and *S. aureus* display values much below the threshold, meaning that their metabolic organization appears to have evolved towards reducing the likelihood of large failure cascades (probably lethal for the organisms) or, equivalently, towards increased

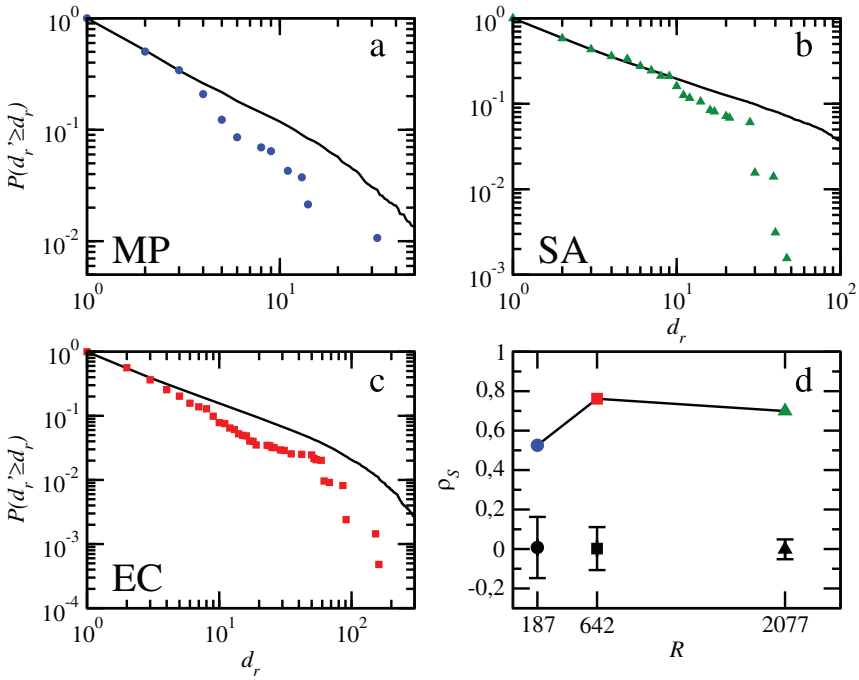

Figure 1 | Damage in cascades triggered by individual reactions. (a–c). Cumulative probability distribution functions of damages in *M. pneumoniae, E. coli*, and *S. aureus*. Results are compared with damages produced in degree-preserving randomized versions of the metabolic networks in order to discount structural effects. In each case, the solid black curve is the average over 100 realizations. Results for *S. aureus* and an older version of *E. coli* were already presented in Ref.(19). The results of the Kolmogorov-Smirnov tests are given in terms of the K-S statistic and its associated significance level: 0.095/0.07, 0.086/0.0002, and $0.079/1.4 \cdot 10^{-11}$ for *M. pneumoniae, S. aureus*, and *E. coli* respectively. With a significance value of $\alpha = 0.05$, distributions of damages can be considered not consistent with those for randomized variants, except for *M. pneumoniae*. (d) Spearman's rank correlation coefficient ρ_S between predictors and damages, plotted against metabolic network size (number of reactions R). Results are compared to random reshuffling of the predictor value associated to reactions (100 realizations for each organism). Average Spearman's rank correlation coefficients for the randomizations appear in black, and error bars delimit the maximum and the minimum values obtained.

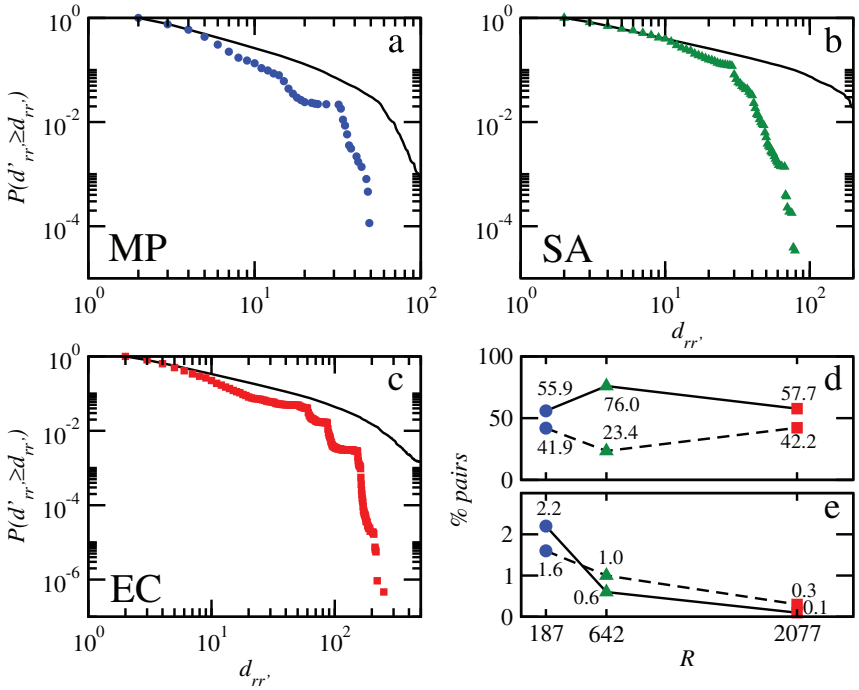

Figure 2 | Damage in cascades triggered by pairs of reactions. (a–c). Cumulative probability distribution functions of damages in *M. pneumoniae, E. coli*, and *S. aureus*. Results are compared with damages produced in randomized versions of the metabolic networks in order to discount structural effects. In each case, the solid black curve is the average over 100 realizations. Results of the Kolmogorov-Smirnov tests (K-S statistic/associated significance level): 0.15/0, 0.14/0, and 0.13/0 for *M. pneumoniae, S. aureus*, and *E. coli* respectively. Taking $\alpha = 0.05$, distributions of damages can be considered not consistent with those for randomized variants. (d) Most frequent double cascades output. Solid line: interference without amplification. It is related with cases b and c in Figure 3. Dashed line: no interference, which is related with case a in Figure 3. e) Non-linear effects in double cascades. Solid line: overlap. It is related with cases c and e in Figure 3. Dashed line: amplification. Amplification is related with cases d and e in Fig. 3.

structural robustness, as previously seen for *S. aureus* and for an older version of the metabolic network of *E. coli* in (19). In contrast, the value of the associated significance level for *M. pneumoniae* is very similar to the threshold. As a consequence, one cannot say that the difference between cascade size distributions in the original network and in the randomized counterparts is statistically significant, even though the probability for large cascades is still smaller in the original metabolic network. This can be explained by the increased linearity and limited redundancy of *M. pneumoniae* metabolic network structure, according to available data[23].

Along with structure, biochemical insight contributes to explain why some reactions trigger larger cascades. For *M. pneumoniae*, the most vulnerable reactions at the top of the damage ranking (see Supporting Information) can be classified into four groups related to vital functions. One group is associated to metabolites phosphoenolpyruvate and protein L-histidine, each solely produced by one generating reaction and both of them directly related to phosphorylation processes, vital for instance in the synthesis of ATP. The second group relates to formate, which has a prominent role in the energy metabolism on many bacteria. The third group involves reactions where the important metabolite is thioredoxin, an antioxidant protein essential to reduce oxidized metabolites, along $NADP^+$. Finally, the failure of reactions in the fourth group trigger large cascades that affect the synthesis of fatty acids by turning acyl carrier proteins inviable.

Prediction of the damage caused by the failure of individual reactions is possible on the basis of local information relative to the triggering reaction alone. We give the explicit mathematical expression of our predictor P_r and a detailed explanation in Materials and Methods. In simple terms, reactions need to be linked to propagator motifs in order to ignite damaging cascades. Basically, these motifs are represented by branched metabolites with just one in or out connection that happens to be attached to the triggering reaction. The higher the branching ratio of these metabolites, the higher the likelihood that the reaction propagates a large cascade, and thus to become a target for structural vulnerability of the network. To give an example, the two most vulnerable reactions in *M. pneumoniae* produce phosphoenolpyruvate, a compound involved in glycolysis and gluconeogenesis that acts as a source of energy. It happens to be a highly-branched cascade propagator motif connected to two reversible reactions and, as a product, to eight irreversible reactions. See Supporting Information for a categorization of cascade propagator motifs in bipartite semidirected networks.

To check the predictive power of our predictor P_r, we measured the Spearman's rank correlation coefficient ρ_S between predictors and damages for each organism. Basically, Spearman's correlation is the Pearson correlation coefficient between two ranks, here given by the positions in ordered lists of reactions according to predictor values P_r and damages d_r. A high ranking position by predictor value is expected to correlate with vulnerable reactions, at the top of the damage ranking. For all three organisms, we found very high values of the correlation coefficient, which are statistically significant (see Fig. 1d). This evidences the ability of our predictor, calculated on the basis of local information, to rank reactions by damage without directly computing the effect of the failure.

Non-linear effects in cascades triggered by pairs of reactions. As expected, the simultaneous failure of two reactions leads to higher damages as shown in Figs. 2a–c. The graphs display the cumulative probability distributions $P\left(d'_{rr'} \geq d_{rr'}\right)$ calculated from all possible pairs of reactions initiating the cascades. It is worth stressing that the order of initiation is irrelevant. Notice that the exponential cut-off is still present, and becomes more prominent even for *M.*

pneumoniae. Again, we assess metabolic robustness by comparing cascades in the original networks with those in degree-preserving randomized counterparts using K-S tests (see caption of Figure 2 for specific values). We find that, for all three organisms including *M. pneumoniae*, the probability for large cascades triggered by pairs of reactions is significantly smaller in the original metabolic networks as compared to those in the randomized variants, suggesting that the organization of metabolic networks has evolved towards protecting metabolism against multiple reaction failures.

We also observed that cascades caused by individual reactions combine in different ways when two reactions fail simultaneously (see the illustrative sketches in Fig. 3, panels 3a–e). The crucial concept here is that of the pattern of interference of the respective areas of influence of the two individually considered cascades. By that we refer to all metabolites and reactions altered (by reactions altered but not removed, we mean reversible reactions that become directed by

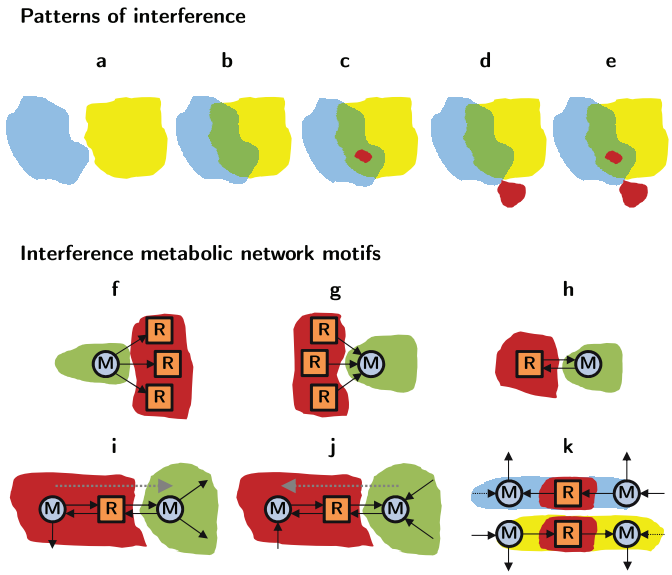

Figure 3 | Cascade propagator network motifs and typology of double cascades. (a–e) Illustration of possible interference patterns between individual cascades: additive, interference without overlap or amplification, interference with overlap and without amplification, interference without overlap and with amplification, interference with overlap and amplification, respectively. Blue and yellow stand for single cascades, green for interference, and red for overlap and amplification, depending on whether the red zone is in the interference zone (green) or not. (f–k) Metabolic network motifs in the interference of two individual cascades that induce amplification. Cases (f–g) Motif caused by a metabolite which loses its only generating reaction and at the same time it is the reactant of several reactions. These reactions are going to be become non-viable. Case g is equivalent to f but inverting the sense of the links. Case (h) Metabolite which has been left with one connection to a reversible reaction. This reversible reaction has zero net flux and becomes inviable. Cases (i–j) This motif appears when a modified metabolite is lead with only one incoming connection coming from a reversible direction. This fixes the reversible reaction towards the production of this metabolite. If this step turns a metabolite of the reversible reaction inviable, the reversible reaction becomes inviable. Therefore, this motif is a potential trigger of amplification. Case j is equivalent to case i when the senses of the reactions are inverted. Case k) The individual cascades fix the sense of a reversible reaction oppositely, one cascade forwards (k top) and the other backwards (k bottom) (note that the pictures illustrate the effects of both cascades individually). After superimposing the effects of the two cascades, one can see that this reversible reaction becomes inviable. Thus, metabolites surrounding the reaction may become inviable as well, depending on their degrees. It is also a potential trigger, as in cases i–j.

effect of the cascade), removed or not, by each single cascade. If there is no interference, the total damage $d_{rr'}$ is additive and equal to the sum of the two single damages d_r and $d_{r'}$. Otherwise, different situations are possible leading to a combined damage that can be equal, larger or smaller than the single added values. The latter case is a univocal signature of cascade overlapping $o_{rr'}$, pointing to the existence of a common subset of reactions that fail in both cascades (the most extreme realization is when one cascade is totally contained in the other). More interesting is the situation when, irrespectively of the presence or absence of overlap, we detect a non-linearly amplified damage involving a number $a_{rr'}$ of new reactions that break down under simultaneous black outs. For all cascades, $d_{rr'} = d_r + d_{r'} - o_{rr'} + a_{rr'}$. Interference without amplification is the most common situation, followed by the absence of interference (Fig. 2d). In contrast, overlap and amplification happen for a very small fraction of all double cascades, and their occurrence decreases with the size of the organism (Fig. 2e). In particular, the reduced incidence of amplification represents a new signature that organizational principles at play ensure the robustness of the organisms, despite increasing complexity and interweaving.

However, amplification may have a very large impact when it occurs. For instance, pyruvate (a product of glucose metabolism and a key intersection in several metabolic pathways) provides energy by fermentation. This process reduces pyruvate into lactate, a reaction that does not trigger any black out cascade when it fails, so $d_r = 1$. At the same time, pyruvate can also be decarboxylated to produce acetyl groups, the building blocks of a large number of molecules that are synthesized in cells. The failure of the first reaction in such pathway triggers a cascade of length $d_{r'} = 3$. In contrast, the simultaneous failure of both the fermentation and the reduction of pyruvate induces a large cascade of size $d_{rr'} = 36$, most likely lethal. As a biological explanation, we suggest that both processes are strongly interdependent to maintain the oxidation-reduction balance when fermentation is in action.

Collateral effects offer the clue to understand this amplification phenomenon. In parallel to rendering non-operational some reactions and their corresponding metabolites, a cascade can reduce the connectivity and increase the branching ratio of other viable metabolites in its influence area. When stricken by the propagation front of a second cascade, these metabolites are susceptible of becoming inviable, further spreading the failure wave. In this way, interference is a necessary but not a sufficient condition for amplification, and a large amplification can be possible even when there is no overlap and the interference between the individual cascades is small. To predict which pairs will trigger amplification, we look at metabolites in the interference of the influence areas of the two individual cascades. We confirmed that those metabolites that remain viable after each individual cascade but become inviable when the two effects are superposed will produce amplification, propagating the double cascade to new reactions. In Fig. 3, panels 3f–k, we provide the connectivity structure of all interference cascade propagator motifs responsible for amplification.

Impact of gene knockouts in metabolic structure. Reaction failures are usually associated to the disruption of an enzyme due to knockout, inhibition, or deleterious mutation of the corresponding gene. However, enzyme multi-functionality and gene essentiality are higher in *M. pneumoniae* as compared to other prokaryotic bacteria, so gene malfunctioning can potentially produce an acuter stress response at the level of metabolism. To address this issue, we coupled the metabolic network of *M. pneumoniae* to its gene co-expression network through the activity of enzymes and we considered knockouts of individual genes and clusters of co-expressed genes. Inherent to this analysis is the potential occurrence of individual, double, or multiple cascades simultaneously. We algorithmically handled multiple knockouts as an obvious extension of the previously considered situation of pair cascades.

Metabolic effects of individual mutations. Individual metabolic gene knockouts or mutations inhibit the production of catalytic enzymes and induced black outs of reactions propagating in the metabolic network as a failure cascade. From existing data, 71% of the 140 metabolic genes in *M. pneumoniae* have a one-to-one relation with reactions, and 21% of the genes regulate multiple reactions. Seldom the same reaction may be individually regulated by different enzymes produced by different genes, which happens for only four non-damaging reactions. More often, several genes are necessary to regulate the activity of a single reaction through an enzymatic complex. Twelve complexes codified by 26% of genes regulate the activity of 10% of metabolic reactions in *M. pneumoniae*. The removal of any of the genes involved in a complex is expected to cause the inactivation of the reaction controlled by the complex, which in principle may increase its vulnerability. However, we observe that almost all complexes are associated to low damage reactions, which indicates a certain degree of structural robustness.

To study the metabolic effects of individual mutations we simulated the knockout of all reactions associated to the gene under consideration. As explained, most often this corresponds to one single reaction but sometimes multiple reactions are removed simultaneously. One first observation is that metabolic genes affecting vulnerable reactions will trigger large failure cascades. More interestingly, genes with large associated damages in metabolism turn out to be essential or conditionally essential for *M. pneumoniae* (see Table 1), with a unique exception discussed below. We use the classification in Ref. (23), where essentiality is defined according to the measured metabolic map and the definition of a minimal medium which allows *M. pneumoniae* to grow. Essential genes are those that are required for the survival of the organism, meaning that the products of the reactions that they control are essential for life and cannot be produced by alternative pathways, while conditional means that essentiality depends on the media composition available.

In fact, we have checked that all conditionally essential genes with the potential of producing high damage in the metabolism of *M. pneumoniae* were found to have an essential orthologue (essentiality determined by loss-of-function experiments) in *Mycoplasma genitalium*[23], a comparable genomereduced bacterium. The only exception to essentiality in Table 1 is gene *mpn062*, considered as nonessential in Ref. (23), while in our study it triggers a large failure cascade and so can be classified as a vulnerable target for metabolic structure. Its damaging potential can be explained by the fact that each of the four reactions controlled by the gene has a contribution that, although not extremely high individually, adds to the total damage and interferes to produce amplification. We propose *mpn062* as an essential gene for metabolic function in *M. pneumoniae*, a conjecture that is supported by the essentiality of its orthologue in *M. genitalium*.

Another interesting case is essential gene mpn429, whose knockout triggers the largest cascade in *M. pneumoniae*. Each of the four reactions it affects in the glycolysis pathway is not able to propagate a cascade individually. However, when they all are removed simultaneously, we observe the strongest amplification effect. The biochemical explanation is that the non-linear interaction of the cascades stops the production of phosphoenolpyruvate, which disrupts the synthesis of ATP, a circumstance particularly harming to the organism.

Metabolic effects of knocking out gene co-expression clusters. Groups of co-expressed genes in *M. pneumoniae* can be identified from gene expression data under different conditions, this revealing a complex gene regulatory machinery[25]. The functional deactivation of these clusters might be produced by the failure of common regulatory elements and important damage could be transmitted to metabolism.

Table 1 | Largest structural damages produced in metabolism by gene knockouts and correspondence with gene essentiality as given in Ref. (25). Damage in metabolic structure caused by gene knockout (third column) is measured in number of deleted reactions. In the fourth column, we give the number of reactions regulated by the corresponding gene, and in parentheses we give the damage associated to each of these reactions. Genes in monocomponent clusters are highlighted in boldface, and we used braces to denote genes that form complexes. Note that the complex at the end of the list is not detected by any of the three clustering procedures. Finally, gene *mpn062* is the only one in the table annotated as non-essential although it is associated to a large failure cascade

Gene	Essentiality	Damage	Reactions
mpn429	yes	49	4 (1,1,1,1)
mpn606	yes	32	1 (32)
mpn628	yes	32	1 (32)
mpn017	yes	25	3 (14,1,9)
mpn303	yes	18	8 (1,1,1,1,8,1,2,3)
mpn062	*no*	17	4 (6,3,2,3)
mpn576	cond	16	2 (13,2)
mpn005	yes	13	1 (13)
mpn336	yes	13	3 (4,3,6)
mpn354	yes	13	1 (13)
mpn627	yes	11	1(11)
mpn066	yes	9	4 (1,1,2,5)
mpn240	cond	9	1 (9)
mpn299	cond	9	1 (9)
mpn322	cond	9	4(1,1,2,1)
mpn323	cond	9	4 (1,1,2,1)
mpn324	cond	9	4 (1,1,2,1)
mpn034	yes	7	4 (1,1,2,3)
mpn378	yes	7	4 (1,1,2,3)

From this point of view, we investigated the effects on the metabolic structure of *M. pneumoniae* of suppressing gene co-expression clusters. To detect these functional clusters, we used three very different strategies applied to the correlation matrix of dependencies between the expression level of pairs of genes in *M. pneumoniae* (see Materials and Methods and Supporting Information). We first considered clusters as defined in (25), where a technique of average distance hierarchical clustering is used and clusters are defined as groups of nodes which are close to each other. Alternatively, we applied a random-walk-based algorithm called Infomap[27], where groups comprise nodes among which information flows quickly and easily. And finally, we used the method of Recursive Percolation, that identifies clusters as strongly interconnected groups of nodes. Notice that the correlation matrix involves all genes, including those non-metabolic. Although the latter do not affect directly metabolic function, they might act indirectly due to epistatic interactions.

The comparative analysis of the detected clusters of genes showed that, although the partitions found by each algorithm may differ in their composition and in the maximum size of the clusters (see Supporting Information), there are preserved commonalities independently of the method. One of them is that all methods are able to detect seven of the twelve complexes, since the related genes always appear classified in the same cluster. Another remark is that the three detection methods result in qualitatively similar power-law-like cluster size distributions (see Supporting Information), with most clusters having small size while some are relatively big. Interestingly, genes related to high damage spreading reactions are secluded into mono-component clusters. To be more precise, eight of the nineteen genes in Table 1 are recognized by all three methods as having an expression profile that is not correlated to other gene activity levels. This is surprising since, in principle, high-damage genes might be expected to be co-regulated with other genes, as influencing big parts

of metabolism usually requires coordinated gene activity. The fact that these genes appear isolated pinpoints them as potentially important metabolic regulator targets, since the alteration of only one gene may affect a large number of metabolic reactions. In any case, the lack of co-regulation of genes related to high damage spreading reactions is again an indication that the structural organization of the organism has evolved towards protecting the system against multiple failures.

Taking averages for equally sized clusters, we found that knockouts of co-expression clusters produce a damage on metabolic structure that increases with the number of affected metabolic genes, except when most metabolic genes in a cluster codify an enzymatic complex regulating one reaction (abrupt kink in Fig. 4a or Fig. 4d). The damage produced by the failure of the cluster also increases when plotted against the number of associated reactions (right panels in Fig. 4). In order to discount structural effects, we compare these results with those measured on randomized versions of the metabolic network of the genome-reduced bacterium. As evidenced in Fig. 4, all cluster detection methods identify clusters that produce lower damages in the real metabolic network of *M. pneumoniae* as compared to the randomized network. This supports the idea that the regulatory machinery that controls the coupled-to-metabolism co-expression of genes has evolved towards robustness.

Finally, since the three cluster detection methods propose different forms of aggregating metabolic genes, we checked whether cluster composition is relevant for failure propagation. As a null model, we considered randomization restricted not to the network itself but to the specific gene metabolic composition, while maintaining the total number of metabolic genes in each cluster. We observed that such a reshuffling of metabolic genes in clusters had no relevant effect on the damages measured on the metabolic network (see Supporting Information). This means that, surprisingly, the composition of the clusters is not as statistically relevant for metabolic vulnerability as the distribution of the cluster sizes itself. This feature, together with the large detected amount of mono-component clusters, point out to the existence of multiple levels of regulation, depending on experimental conditions, and, at the same time, explains why genes controlling high damage spreading reactions operate preferentially under functional isolation as a metabolism protection mechanism.

Discussion

Taken together, our results shed light on the genome-scale impact and prediction of failure propagation under structural stress in bacterial metabolism, with potential implications in areas like metabolic engineering or disease treatment. In this context, we demonstrate that *M. pneumoniae* exhibits network responses that are qualitatively comparable to those of other model bacterial organisms like *E. coli*, although we find it less robust against individual reaction removals with reactions more prone to trigger large metabolic failure cascades identified as key participants in the regulation of energy and fatty acid synthesis. Recently, it has been argued that the shape of the degree distribution in chemical reaction networks is a neutral feature which has no evolutionary implications in itself and that is dictated by the size of the network[28]. We find that the ability of the metabolic network of *M. pneumoniae* to spread cascades, as compared to that of larger bacteria, could not just be explained by its metabolite degree

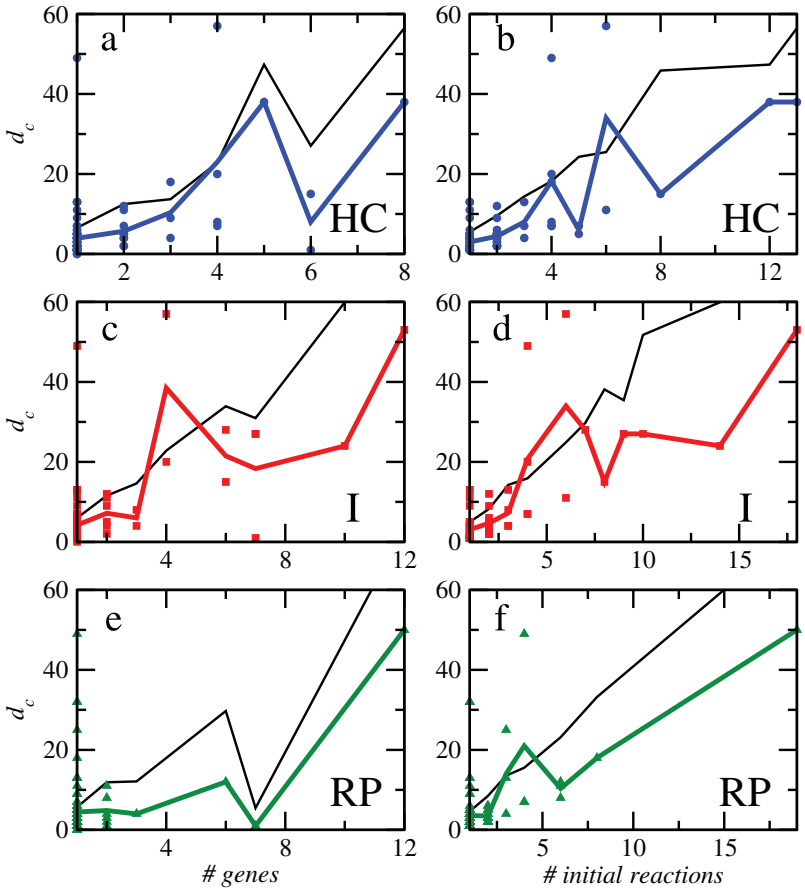

Figure 4 | Damages as a function of the number of metabolic genes and reaction failures in gene co-expression cluster knockouts. Clusters are defined according to three different methods: Hierarchical Clustering (HC), Infomap (I), and Recursive Percolation (RP), see Materials and Methods. Results are compared with damages produced in randomized versions of the metabolic networks in order to discount structural effects. In each case, the solid black curve is the average over 100 realizations.

distribution or small size effects. Different organizational levels of the network, which are surely influenced by evolutionary pressure, should come into play as becomes clear in the different behaviors found for single and multiple failures.

The concept of cascade amplification has been for the first time formulated and interpreted, as a signature of the subtle non-linearities underlying the structure of complex networks. Specific scenarios in *M. pneumoniae* have been discussed. In addition, we were singularly motivated to assess the predicting power of our formalism. In this sense, we have proposed both a predictor of damage propagation for single cascades and we have identified structural motifs underlying amplified failure patterns in situations of concurrent spreading.

On what respects to the analysis of single gene knockouts, our analysis reveals its potentiality in capturing most of the scenarios of experimentally determined lethality for *M. pneumoniae*. Moreover, when clustered and knocked together new trends of the complex genomic regulation of the metabolism emerge. First, the distribution of cluster sizes seems to matter more than the actual composition of the clusters. This is connected to the fact that the regulation of high-damage genes tends to appear isolated from the that of other genes, a kind of functional switch in metabolic network that at the same time acts as a kind of genetic firewall.

The study of complex systems under stress poses a number of formidable challenges critical to understand their behavior as well as towards proposing successful strategies for prediction and control. In this framework, the study of human pathogens may help to develop more sophisticated forms of identifying new and more efficient drug targets. Finally, we emphasize that both our methodology as well as the formulated predictor index and detected propagator motifs proposed here are general tools for testing structural robustness of bipartite networks under stress, in whatever context.

Methods

Data and multilevel complex network representation. The three bipartite metabolic networks representation used in this work have been constructed from biochemical reaction data downloaded from publicly available databases. The general criteria applied in all representations are as follows:

1. All biochemical reactions are considered except *exchange*, *sink*, and other auxiliary reactions like biomass formation and ATP maintenance reactions included in the BiGG databases, sometimes introduced for consistency. Reversible reactions are treated as two coupled directed reactions in the forward and the reverse sense.
2. All metabolites involved in the reaction included in the network representation are considered. In particular, hubs are not excluded. Hubs stay neutral with respect to cascades and do not contribute to propagate them. Due to their large number of connections, they are highly unlikely to become nonviable as a consequence of single or double cascades reaching them. Since they are part of the system, we keep them into the bipartite network representation for completeness.
3. Metabolites in different compartments are treated as different metabolites (such as metabolites present in cytosol, periplasm or exterior).
4. The total degree of a reaction is the number of reactants and products involved. We characterize the connectivity of a metabolite by its incoming, outgoing, and bidirectional degree (k_i, k_o, and k_b), which count the number of reactions having the metabolite as a product, as a reactant, and the connections to reversible reactions, respectively.

The metabolic network of *M. pneumoniae* was obtained from Ref. (23), where the authors integrated biochemical and computational studies. Its metabolic reconstruction contains 187 reactions taking place in cytosol and exterior, and divided into four types (diffusion, transport, isomerization and transformation reactions). The number of metabolites is 228. The total average connectivity of metabolites is 3.5, and for reactions 4.3.

On the other hand, the genome of *M. pneumoniae*[25] comprises 688 genes, 140 of which having a metabolic function. Except for one spontaneous reaction and 20 reactions with unknown regulation, these metabolic genes codify 142 enzymes that catalize all reactions in the metabolic network of this organism. Correlations in the expression of genes were measured from tilling arrays under 62 different environmental conditions as provided in (25). The measured gene correlation matrix can be transformed into a network representation and coupled to the metabolic network of *M. pneumoniae* through the activity of enzymes to produce a multilevel network representation.

Data for the bipartite directed network reconstruction of the metabolism of *E. coli*[29] where downloaded from the BiGG database (http://bigg.ucsd.edu/). The BiGG database provides curated information which avoids unspecific reactions and complements the experimental data with flux balance analysis modeling. This makes the BiGG database of high quality for statistical inquiry and very appropriate for our investigations. For *E. coli*, we use the iAF1260 version of the K12 MG1655 strain. In this organism, reactions happen in three compartments: exterior, periplasm and cytosol. This version contains 2077 reactions and 1669 metabolites. The average connectivity of reactions is 4.3, whereas that of metabolites is 5.3.

The bipartite directed network reconstruction of *S. aureus* was obtained from Ref. (19), where they used an *in silico* version available in the BiGG database. Again, in this case there are only cytosol and exterior compartments. The number of reactions is 642, after excluding sink to cytosol reactions, and the number of metabolites is 644. The average connectivity of both metabolites and reactions is 4.8.

See Supporting Information for further analysis of the topological properties of these networks, including power-law goodness of fit tests for the degree distributions of metabolites.

Damage predictor. In order to predict the damage caused by the failure of individual reaction r, we look at its associated metabolites m and compute

$$P_r = \sum_{m \in r} \Big[(k_i + k_b)\delta^0_{k_o} \left(\delta^1_{k_b} + \delta^0_{k_b}\right)\left(\delta^1_{k'_o - k_o} + \delta^1_{k'_b - k_b}\right) + (k_o + k_b)\delta^0_{k_i} \left(\delta^1_{k_b} + \delta^0_{k_b}\right)\left(\delta^1_{k'_i - k_i} + \delta^1_{k'_b - k_b}\right) - \delta^0_{k_i}\delta^0_{k_o}\delta^1_{k_b}\delta^1_{k'_b - k_b}\Big]. \quad (1)$$

Degrees k_i, k_o and k_b respectively refer to the number of incoming, outgoing and bidirectional links of metabolite m (reactant or product) associated to the triggering reaction r after discounting the link used to propagate the cascade, and k'_i, k'_o and k'_b denote the original values before the cascade is triggered. We use δ^b_a for Kronecker's delta function. Basically, we identify metabolites susceptible to propagate the cascade, which are those having originally just one *in* or just one *out* link, which is the one connecting them to the triggering reaction, or those connected to the triggering reaction by a *bidirectional* link and lacking in or out connections. The contribution to the predictor of one of those metabolites counts the number of connections of this metabolite with the rest of reactions, which can then be considered susceptible to become non-viable and propagate the cascade (see Figure S4 in Supporting Information for an illustrations showing how the measure works for some particular cases). Bidirectional links introduce some subtleties. Some motifs involving them may propagate damage depending on the viability of the associated reversible reaction. We always take into account these motifs when we compute the value of the predictor for each reaction.

Methods for detecting gene co-expression clusters. The matrix of correlations between the expression levels of pairs of genes gives a fully connected network where the link between two genes carries a weight equivalent to the expression correlation between them. We used three different methods to detect gene co-expression clusters. First, we used the results in[25], where a distance hierarchical clustering technique was applied to the matrix of correlations after applying a threshold of 0.65 to it (which reduces the density of links to 0.007) and transforming the Pearson correlations of expression levels into a distance. Second, we applied an existing algorithm to find communities called Infomap (Ref. (27)) to the matrix of correlations after applying the threshold of 0.65 to the weights. This algorithm detects communities using a random walk with jumps. A community is a set of nodes where the random walker flows quickly and easily. Third, we consider the clusters in which the co-expression network is fragmented just below the percolation threshold, where the connected network disaggregates into smaller components (other fully connected networks have been analyzed at the percolation point, see for instance[30]). To find them, we removed links sequentially from lower to higher weight until we detected the percolation transition (by computing magnitudes which have a singularity in this point, like the size of the second largest cluster and the average size of the clusters excluding the largest one). At this point we measured the clusters using a burning algorithm. Typically, the gene co-expression network fragmented into two large clusters and several small clusters. We applied the same procedure to the two largest components and so on until the distribution of sizes was similar to those for the hierarchical clustering technique and Infomap (see Supporting Information). We called this procedure Recursive Percolation.

1. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
2. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. Critical phenomena in complex networks. *Rev. Mod. Phys.* **80**, 12751335 (2008).
3. Barrat, A., Barthélemy, M. & Vespignani, A. *Dynamical Processes on Complex Networks.* Cambridge University Press, Cambridge, (2008).
4. Cohen, R., Erez, K., ben Avraham, D. & Havlin, S. Resilience of the internet to random breakdown. *Phys. Rev. Lett.* **85**(21), 4626 (2000).
5. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378 (2000).
6. Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5766–5771 (2002).

7. Moreno, Y., Gómez, J. B. & Pacheco, A. F. Instability of scale-free networks under nodebreaking avalanches. *Europhys. Lett.* **58**, 630–636 (2002).
8. Motter, A. Error and attack tolerance of complex networks. *Phys. Rev. E* **66**, 065102(R) (2002).
9. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. Catastrophic cascade of failures in interdependent networks. *Nature* **464**, 1025–1028 (2010).
10. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cells functional organization. *Nature Reviews Genetics* **5**, 101–113 (2004).
11. Szalay, M. S., Kovacs, I. A., Korcsmaros, T., Bode, C. & Csermely, P. Stress-induced rearrangements of cellular networks: Consequences for protection and drug design. *FEBS Leters* **581**, 3675–3680 (2007).
12. Motter, A. E., Gulbahce, N., Almaas, E. & Barabási, A.-L. Predicting synthetic rescues in metabolic networks. *Molecular Systems Biology* **4**, 168 (2008).
13. Palsson, B. O. *Systems Biology: Properties of Reconstructed Networks* (Cambridge University Press, Cambridge, 2006).
14. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
15. Ma, H. & Zeng, A.-P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**, 270–277 (2003).
16. Guimerà, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).
17. Serrano, M. A. & Sagués, F. Network-based confidence scoring system for genome-scale metabolic reconstructions. *BMC Systems Biology* **5**, 76 (2011).
18. Serrano, M. A., Boguñá, M. & Sagués, F. Uncovering the hidden geometry behind metabolic networks. *Molecular BioSystems* **8**, 843–850 (2012).
19. Smart, A. G., Amaral, L. A. N. & Ottino, J. Cascading failure and robustness in metabolic networks. *Proc. Natl. Acad. Sci. USA* **105**, 13223–13228 (2008).
20. Edwards, J. S. &d Palsson, B. O. Robustness analysis of the escherichia coli metabolic network. *Biotechnol. Prog.* **16**, 927–939 (2000).
21. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* **99**, 15112–15117 (2002).
22. Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E. & Shlomi, T. Predicting selective drug targets in cancer through metabolic networks. *Molecular Systems Biology* **7**, 501 (2011).
23. Yus, E. *et al.* Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326**, 1263–1268 (2009).
24. Kühner, S. *et al.* Proteome organization in a genome-reduced bacterium. *Science* **326**, 1235–1240 (2009).
25. Güell, M. *et al.* Transcriptome complexity in a genome-reduced bacterium. *Science* **326**, 1268–1271 (2009).
26. Smirnov, N. V. Tables for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, **19**, 279 (1948).
27. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* **105**, 1118–1123 (2008).
28. Lee, S. H. *et al.* Neutral theory of chemical reaction networks. *New Journal of Physics* **14**, 033032 (2012).
29. Feist, A. *et al.* A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Molecular Systems Biology* **3**, 121 (2007).
30. Rozenfeld, A. F., Arnaud-Haond, S., Hernández-García, E., Eguíluz, V. M., Serrão, E. A. S. & Duarte, C. M. Network analysis identifies weak and strong links in a metapopulation system. *Proc. Natl. Acad. Sci. USA* **105**, 18824–18829 (2008).

Acknowledgments

We thank Luis Serrano, Eva Yus, Marc Güell and Ashley Smart for kindly providing the empirical data. We also thank Georg Basler for helpful comments. This work was supported by MICINN Projects No. FIS2006-03525, FIS2010-21924-C02-01 and BFU2010-21847-C02-02; Generalitat de Catalunya grant No. 2009SGR1055; the Ramón y Cajal program of the Spanish Ministry of Science, and the FPU grant of the Spanish Ministry of Science.

Author contributions

O.G., F.S., and M.A.S. contributed equally to the design and implementation of the research, and to the writing of the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Güell, O., Sagués, F. & Serrano, M.Á. Predicting effects of structural stress in a genome-reduced model bacterial metabolism. *Sci. Rep.* **2**, 621; DOI:10.1038/srep00621 (2012).