Published online 25 May 2019 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* ***W151–W157***
doi: 10.1093/nar/gkz451

IAMBEE: a web-service for the identification of adaptive pathways from parallel evolved clonal populations

Camilo Andres Perez-Romero[1,2,*], Bram Weytjens[1,2], Dries Decap[2], Toon Swings[3,4,5], Jan Michiels[3,4], Dries De Maeyer[1,2] and Kathleen Marchal[1,2]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, [2]Department of Information Technology, IDLab, Ghent University, IMEC, Ghent, Belgium, [3]VIB Center for Microbiology, Flanders Institute for Biotechnology, Leuven, Belgium, [4]Centre of Microbial and Plant Genetics, KU Leuven, Leuven, Belgium and [5]VIB Technology Watch, Flanders Institute for Biotechnology, Ghent, Belgium

Received March 11, 2019; Revised May 02, 2019; Editorial Decision May 08, 2019; Accepted May 10, 2019

ABSTRACT

IAMBEE is a web server designed for the Identification of Adaptive Mutations in Bacterial Evolution Experiments (IAMBEE). Input data consist of genotype information obtained from independently evolved clonal populations or strains that show the same adapted behavior (phenotype). To distinguish adaptive from passenger mutations, IAMBEE searches for neighborhoods in an organism-specific interaction network that are recurrently mutated in the adapted populations. This search for recurrently mutated network neighborhoods, as proxies for pathways is driven by additional information on the functional impact of the observed genetic changes and their dynamics during adaptive evolution. In addition, the search explicitly accounts for the differences in mutation rate between the independently evolved populations. Using this approach, IAMBEE allows exploiting parallel evolution to identify adaptive pathways. The web-server is freely available at http://bioinformatics.intec.ugent.be/iambee/ with no login requirement.

INTRODUCTION

In clonal systems, genotype-phenotype mapping is a popular technique to study the molecular mechanisms underlying complex phenotypes (1–3) or evolutionary principles (e.g. epistasis (4–6), clonal interactions (7,8) etc). Clonal populations that independently acquired the same adaptive phenotype are genotyped in order to identify the alterations, causal to the commonly adapted phenotype (referred to as drivers or adaptive mutations). Such populations can be obtained through either natural or experimental evolution (2,9–11).

Clonal evolution starts from a single clone cultivated for prolonged periods of time in predefined selective conditions. During this period of time, natural selection favors genetic changes (SNPs/indels hereafter referred to as mutations) that confer a benefit in the chosen condition leading to improved phenotypes (11). Clones carrying these selected adaptive mutations will undergo a selective sweep: mutations causal to the adaptive phenotype increase in frequency and eventually become fixed in the population. However, not all high frequency variants fixed in the evolved population are causal: neutral or slightly deleterious mutations also hitchhike to fixation. Distinguishing the adaptive or driver mutations from the hitchhiking or passenger mutations is a non trivial problem. In addition, increased mutation rates in the population elicited by the presence of hypermutation phenotypes results in an increased ratio of passengers to adaptive mutations, further complicating the identification of adaptive mutations (12,13).

To facilitate the identification of driver mutations the information gained from multiple independently evolved populations is exploited: genes that are mutated in multiple parallel evolved populations are more likely to be adaptive (3,11,14). Relying on such recurrency analysis (14,15) is not trivial, because the relatively low number of parallel samples decreases the power of the analysis. That is why the 'recurrence' with which a gene is observed to be mutated in the independently evolved populations is leveraged with additional information e.g. on the functional impact of the mutations (3,16) or, on the dynamics of mutations during evolution (e.g. whether the frequency increase of a mutation in a population (selective sweep) can be associated with a concomittant increase in the adaptive phenotype) (17).

However, in clonal systems just relying on the identification of 'mutational recurrence' in a set of parallel evolved

*To whom correspondence should be addressed. Tel: +32 486909943; Email: kathleen.karchal@ugent.be

populations does not always allow identifying adaptive mutations. Indeed, complex phenotypes originate by interfering with one or more causal pathways. As the same pathway can become altered in different ways, independent populations that acquired the same adaptive phenotype might all affect the same pathways but not necessarily by interfering with the same genes (14,17,18). As a result, the recurrence of an adaptive mutation does not have to be high in independently evolved populations and it is often difficult to identify rarely mutated drivers based on mutational recurrence. Searching for recurrently mutated pathways rather than genes increases the power of the analysis: in a set of independently evolved populations the chance of finding a pathway being recurrently mutated is higher than finding an individual gene being recurrently mutated (18–21).

Hence clonal genotype-phenotype mapping can benefit from approaches that exploit parallelism between independently evolved populations at pathway rather than at single gene level. Network-based methods are promising in this regard (22). By searching in a network scaffold for recurrently mutated network neighborhoods as proxies for molecular pathways, they obviate the need of using predefined pathways (22). The used network scaffolds, in which nodes represent genes and the edges the interactions between the genes to drive their analysis are derived from available interaction databases (KEGG, Reactome etc.)

Network-based driver identification has been successfully applied in larger cancer genomics studies (18,20,21,23). However, their applicability in the context of clonal microbial evolution is limited as they require a relatively large number of samples (number of independently evolved populations) and do not exploit the additional information on mutational dynamics during evolution that is typically available in the context of experimental evolution studies.

Hence, to facilitate the identification of adaptive mutations/pathways, we developed **IAMBEE-web**. **IAMBEE-web** is a generic tool that can in principle be applied to any clonal system, but it is designed to accommodate the specific information that is available in the context of experimental evolution studies. The algorithm underlying IAMBEE is described in Swings *et al.* (17).

METHODOLOGY

IAMBEE-web is compatible with any up to date internet browser. The web server's documentation provides detailed guidelines on how to perform the analysis, tune the parameters and interpret the results. The service is freely accessible. Based on the job name and description a unique space is created inside the server on which the user can upload data in real time and track the progress of the analysis.

IAMBEE starts from the genotype information obtained from populations that independently acquired the same adaptive phenotype. The algorithm underlying **IAMBEE-web** is based on a probabilistic pathfinding approach (19,24,25). It uses a topology weighted network prior for the organism of interest to search for network neighborhoods that are affected in multiple parallel evolved populations (Figure 1). These neighborhoods are proxies for adaptive pathways. Prior to running **IAMBEE** a topology weighted interaction network is derived from the prior interaction scaffold. Hereto a sigmoidal function is used which down-weights edges originating from large hubs while avoiding to penalize interactions involving nodes with low out-degrees (see help file for detailed information). Such correction is needed to avoid biasing the search for relevant subnetworks towards hubs and their neighboring nodes.

Algorithmically, **IAMBEE** proceeds in two steps. In a first step, called the pathfinding step all genes with at least one mutation in any of the independently evolved populations are mapped on a topology-weighted interaction network. The topology weighting accounts for the negative impact of hubs during the analysis (17). In this pathfinding step, all possible paths that originate from an aberrant gene in a population and end in any other gene mutated in another population are enumerated and given a probability which reflects the degree of belief that the path is associated with the adaptive phenotype (Figure 1). A path is defined as a series of consecutive edges in the interaction network. However, as enumerating all possible paths is computationally too expensive only the N-best paths with the highest probabilities are enumerated. The probability of a single path depends on the topology-based weights of the edges that define the path, combined with a weighting of the path based on the 'relevance' of the start and stop genes that make up the path. The latter is derived from additional information on the functional impact of the mutations occurring in these genes and their dynamics during evolution (see below). The total set of N-best paths (together with their nodes and edges) are used as input in the optimization step.

During the second optimization step, the algorithm searches for a collection of highly probable paths that connect as many as possible mutations occurring in different populations. It does this while selecting as few as possible edges. By imposing the latter constraint, the algorithm is forced to select paths with overlapping edges and hence focuses on neighborhoods in the interaction network that are recurrently mutated in the different populations (Figure 1).

IAMBEE defines the path probabilities in such a way that they can also reflect additional information that is relevant in prioritizing adaptive mutations. This includes the fact that mutations that increase in frequency in the population during a selective sweep are more likely to be adaptive. In addition, adaptive mutations are expected to have a larger predicted functional impact than neutral mutations. It also makes sense to assume that because of their relatively larger accumulation of passengers, populations with higher mutation rates contribute relatively less information to the identification of recurrently mutated network neighborhoods than populations with a lower mutation rate. Including this extra information through the path probabilities allows maximally exploiting all information contained in an experimental evolution set up to optimally steer the search for recurrently mutated network neighborhoods.

INPUT

IAMBEE requires an interaction network to drive its analysis. Such network is a representation of all available knowledge on interactions between molecular entities in the organism of interest. For model organisms this interaction information is available in specialized databases (Reactome

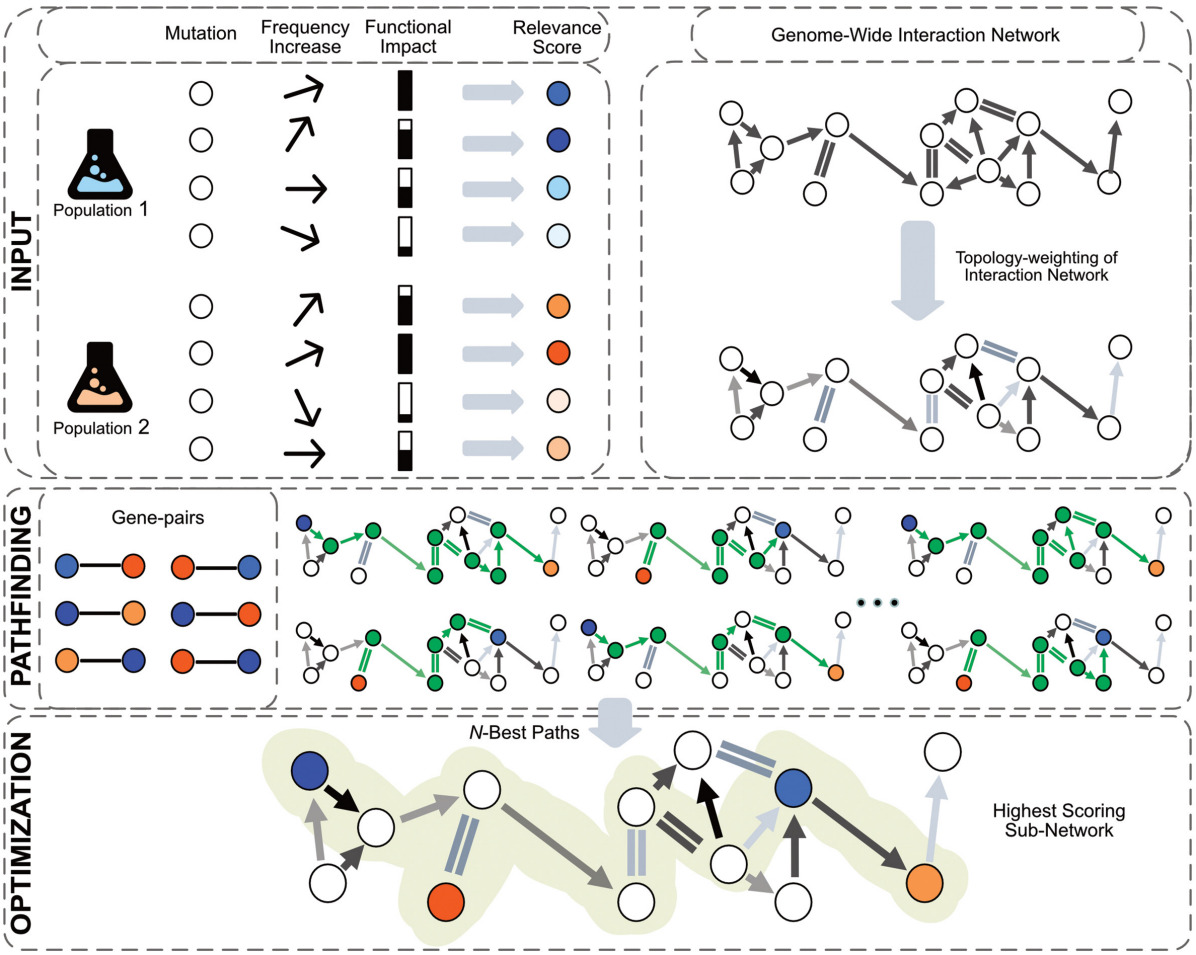

Figure 1. Overview of **IAMBEE**, a web-service for the identification of adaptive pathways from the sequence data of parallel evolved clonal populations. The input consists of a genome wide interaction network of the organism of interest and sequence data obtained from parallel evolved populations (each parallel population is indicated with a different color). Variant calling allows detecting for each population its variants (referred to as the mutant). Extra information on the 'functional impact' of each variant (larger functional impact is indicated with a darker coloring) and the frequency increase of the variants during the sweep are optional. The frequency increase together with the mutation rate of the different populations can also be estimated by **IAMBEE** from the VCF files. All genes with at least one mutation in any of the independently evolved populations are mapped on a topology-weighted interaction network. The functional impact and/or the frequency increase and/or the mutation rate of the population carrying the variant are used to assign to each gene (network nodes) a relevance score (reflecting the potential relevance of the node for the acquired phenotype). The degree of shading of the nodes is indicative of their relevance score. In this pathfinding step the *N*-best paths are enumerated that originate from an aberrant gene in a population and end in any other gene mutated in another population (indicated by the gene pairs). The probability of a path depends on the topology-based weights of the edges that define the path, combined with a weighting of the path that is based on the 'relevance scores' of the start and stop genes that make up the path. The subsequent optimization step operates on the collection of edges/nodes composing the *N*-Best paths selected during the pathfinding step. The optimization algorithm searches in this collection of preselected nodes/edges for highly probable paths that connect as many as possible mutations occurring in different populations using the least number of edges (referred to as the highest scoring subnetwork). This results in recurrently mutated neighborhoods that are a proxy of adaptive pathways (indicated by the shaded area).

(26), KEGG (27)). For less studied species STRING provides a useful resource. **IAMBEE** provides an automatic download for interaction networks available in STRING (28). The interaction network is provided by the network file. This file also allows specifying the molecular level of the interactions (transcriptional, signaling etc.) and their directionalities. To avoid excessive running times and spurious predictions, it is advisable to use a well curated, not too overconnected network.

Next to the interaction network **IAMBEE** also requires the genotypic information for each of the independently evolved populations. Genotypic information is provided in the mutation file, which minimally requires for each population the called variants with respect to the reference sequence, together with an indication of the position and ID of the gene to which the variants can be mapped. In the context of an evolution experiment the reference sequence ideally corresponds to the genomic sequence of the ancestral clone. One can choose to sequence the entire adapted population or individually adapted clones. The latter is suboptimal as it obviates deriving information on the 'frequency increase' of a called variant during evolution. When using population sequencing, the used variant caller should allow for calling the less frequent variants and for estimating their frequency in the population. Functional impact scores can be obtained from SIFT (29) as explained in the help file.

Users can choose to leave out synonymous mutations all together as they are unlikely to have a functional impact and might increase the signal to noise ratio in the data (ratio of adaptive versus passenger mutations).

The 'frequency increase' refers to the degree with which a mutation increases in the population during a selective sweep. To derive the frequency increase, sequence data should for each independently evolved population ideally be available for two time points during experimental evolution, one time point prior to the selective sweep and one after the sweep (i.e. the adapted population). If only the data of the adapted population are available, the increase can be estimated relative to the ancestral strain/population. The user can himself add the information on the frequency increase to the mutation file or alternatively upload the VCF files of the sequenced populations to enable **IAMBEE** deriving the frequency increase of each of the called variants.

In addition, the user can choose whether or not to account for differences in mutation rates between the studied populations when searching for adaptive pathways. If this option is switched on **IAMBEE** first identifies populations with significantly higher mutation rates using the modified Z-score for outlier detection based on the number of mutations present in each of the populations (see Swings *et al.* (17)). From this modified Z-score a population specific-correction factor is calculated. The correction factor intrinsically assigns a relatively lower value to outlier populations if a larger number of independent populations are available, hereby largely reducing the effects of populations with high mutation rates to reduce noise when a large number of independent populations is present. When only a limited number of independent populations is available, the correction factor will be relatively higher, as in that case also the populations with larger mutation rates are needed to exploit parallelism (as so few populations are available). The net effect of the correction is that mutated genes originating from a highly mutated population will receive a relatively lower relevance score and hence will less affect the outcome of the optimization.

All of the above mentioned additional information on the impact of mutations, their frequency increase and the mutation rate of the populations from which the variants are originating weight the impact variants will have on the final solution. Providing this additional Information is optional. However, the information will reduce the search space and steer the search towards a more biologically relevant solution, especially if only a low number of independent populations is available. In some cases the algorithm might not be able to converge without this extra information.

PARAMETERS

Applying **IAMBEE** requires setting some running parameters: defaults are provided for all parameters. The 'N-best paths' parameter relates to the aforementioned pathfinding step. As enumerating all possible paths originating from an aberrant gene in a population and ending in any other gene mutated in another population is computationally too expensive, only the N-best paths with the highest probabilities that connect the respective aberrant genes in a pair will be considered. Increasing the number of best paths allows for a more accurate estimation of the probability that a path exists between two nodes of interest but takes longer. As **IAMBEE** uses a stochastic optimization procedure, repeating the algorithm with the same parameters will give slightly different results. The 'number of repeats' refers to the number of times the optimization step is repeated. Increasing the number of repeats increases the chance of finding the most optimal solution but comes at the expense of a higher computational cost. The optimization tries to connect as many mutated gene pairs as possible through paths over the interaction network using the least number of edges. This optimization is achieved by receiving a 'reward' for each pair of mutated genes that gets connected through a path and adding a penalty for each edge that is used to compose the path. The latter penalty is imposed by the 'cost parameter'. The larger the cost, the more the addition of edges in the connecting paths is penalized during optimization. Increasing the cost will favor a solution with less edges and decreases the size of the inferred subnetwork. As we observed that edges or nodes detected in a subnetwork obtained with a high cost are mostly also contained in solutions obtained at a lower cost, the cost parameter provides a way to balance between sensitivity and precision. Hence, performing a sweep over the cost parameter allows assigning a weight to the edges or nodes reflecting their signal strength in the data. Edges or nodes that are already detected at the higher cost represent the more pronounced and hence more reliable signals in the data and will be assigned respectively a higher weight (for edges) or a higher rank (for nodes). The user can either use the default values or tune the range of the sweep manually. If preferred the user can run the algorithm with just one value for the cost parameter. The network size is a mere visualization parameter which determines the maximal size of the network that will be visualized. This parameter does not affect the algorithm.

RESULTS

Using the input data, **IAMBEE** maps the mutational information from independently evolved populations on an interaction network and searches for network neighborhoods that are affected in multiple evolved populations (Figure 2). These recurrently affected network neighborhoods are proxies for adaptive pathways. **IAMBEE** outputs these neighborhoods in different formats (e.g. SIF, XGMML, TXT and JS/HTML) for download and further analysis in for example Cytoscape. The inferred subnetwork can also be visualized in **IAMBEE-web**. In this visualization genes are nodes and edges the interactions between the genes of the selected network neighborhoods. The edges in the network visualization are colored according to the information on the interaction types provided in the interaction file. The directionality of the edge, if provided is indicated by an arrowed edge. If a sweep is performed over the cost parameter, a single network will be visualized that merges the results obtained at each cost parameter. This merged network is the non-redundant union of the network neighborhoods recovered at each cost parameter. Edges with a higher weight are recovered at more stringent cost parameter values and hence are more reliable.

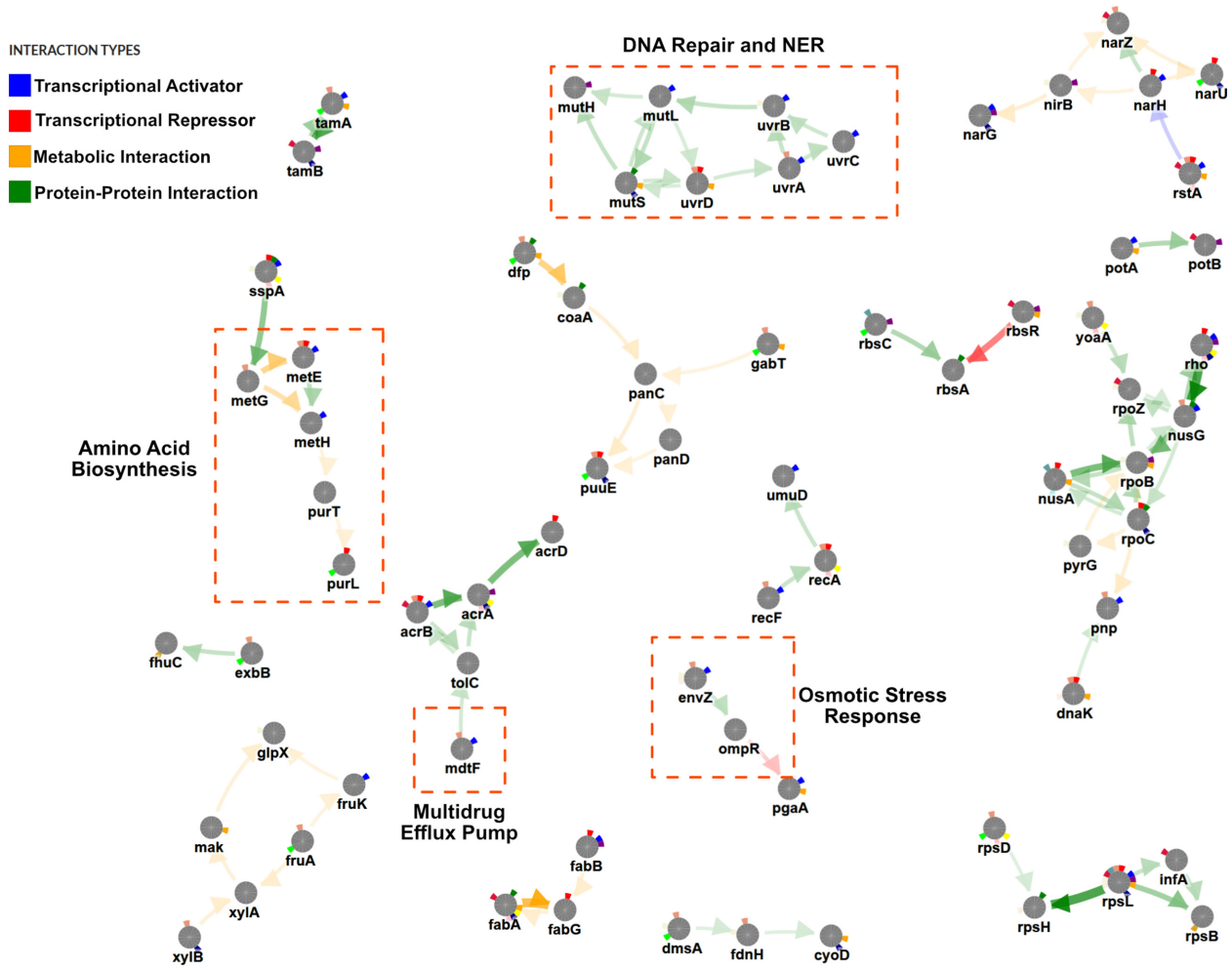

Figure 2. Adaptive pathways involved in ethanol tolerance. The colored segments surrounding each node indicate the populations in which the node (gene) was mutated. In total 16 parallel populations were analyzed, each indicated with a different color. If a gene was affected in multiple populations, it contains multiple colored segments. Genes involved in DNA repair, osmotic stress and amino acid biosynthesis are indicated in orange boxes. The edges in the network visualization are colored according to the interaction type they represent; each function of the interaction is explained in the legend. The edge width depicts the relevance of the edge to the phenotype (as determined by the sweep on the edge cost parameter). This weight is assigned to the edges based on the maximum edge cost for which they are still included in an optimal subnetwork. More reliable edges will have a smaller width.

CASE STUDY

To illustrate the workflow, a first example analysis was performed using the data obtained from Swings *et al.* (17): 16 independent *Escherichia coli* MG1655 populations were experimentally evolved under increasing ethanol concentrations. Their fitness assessed by measuring their growth at elevated ethanol concentrations was traced over time. The fitness trajectories for all 16 populations show selective sweeps between 6% and 6.5% ethanol tolerance. To identify which mutations were responsible for this sudden increase in ethanol tolerance, the populations sampled right before and after this selective sweep were sequenced. Read mapping against the reference genome (ASM584v2 – Ensembl) was performed using BWA V0.7.17 (30), variants were called using LoFreQ V2.1.3.1 (31). All mutations were mapped to the corresponding genes and the SIFT4G-annotator (29) was used to obtain their functional impact. **IAMBEE** was run with default parameters using for each mutation its functional impact score and its frequency increase during the sweep. The impact of mutated genes on the analysis was corrected for the mutation rate of the population in which they occurred.

The used network was constructed by compiling interactions from KEGG, RegulonDB and STRING (Swings *et al.* (17), network available in the tab Download Networks on the website). The retrieved subnetwork (or recurrently affected network neighborhood) is displayed in Figure 2. One of the prioritized network components consists of genes involved in DNA repair (*mutS, mutL* and *mutH*), Nucleotide Excision Repair (NER), (*uvrA, uvrB, uvrC* and *uvrD*). Finding mutations in DNA repair systems is in line with the increased mutation rates that were observed in this evolution experiment (32). In addition, part of the retrieved subnetwork could be associated with adaptation to higher ethanol concentrations e.g. the genes encoding the multidrug efflux pumps (*mdtF*), or the genes involved in amino acid biosynthesis (*metE, metG, metH, purT and purL*) and osmotic stress response (*envZ* and ompR) (for a full de-

scription see reference (17)). Figure 2 also illustrates that all strains that acquired the same tolerance phenotype display adaptive mutations in the same pathways, but not always through the same gene. This emphasizes the necessity of using network-based methods to enable the identification of adaptive mutation/pathways. This case study shows that, despite the increased mutation rate in these experiments and the concomitantly high ratio of passengers versus adaptive mutations **IAMBEE** was able to successfully identify pathways involved in the observed adapted phenotype. A second example in yeast based on the study of Jerison *et al*. (33) is provided in the help file.

DISCUSSION

IAMBEE is a web-service that allows performing network-based identification of adaptive pathways in clonal systems. Despite being applicable to the analysis of any type of clonal system, our web service contains unique features that specifically facilitate the analysis of microbial evolution experiments.

It exploits parallel evolution to search in an interaction network for network neighborhoods recurrently mutated in different independently evolved samples. 'Rare' causal mutations that cannot be prioritized based on observed 'recurrence' can indirectly be recovered because they are a member of the prioritized network neighborhoods. In addition, the identified network neighborhoods are proxies for adaptive pathways. Hence, network-based methods differ from recurrence-based methods in prioritizing entire pathways rather than individual genes. The pathway view provides insight in the molecular mechanism underlying the adaptive phenotype. In addition, after having identified different adaptive pathways with **IAMBEE** one could trace back through the population specific mutation data whether the pathways are hit across the different populations in a conserved order or whether the presence of a mutation in a certain adaptive pathway excludes mutations in another pathway (mutually exclusivity (17)). Such analysis allows studying epistasis, not only at the gene but also at the pathway-level.

DATA AVAILABILITY

IAMBEE-web is available by using the link: http://bioinformatics.intec.ugent.be/iambee/

ACKNOWLEDGEMENTS

We would like to thank the anonymous reveiwers for their very useful comments and remarks which helped us improving the website and including additional case studies.

FUNDING

Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) [3G046318, G.0371.06]; Agentschap voor Innovatie door Wetenschap en Technologie (IWT) [NEMOA]; Katholieke Universiteit Leuven [PF/10/010] (NATAR). Funding for open access charge: Fonds Wetenschappelijk Onderzoek [G.0371.06].

Conflict of interest statement. None declared.

REFERENCES

1. Voordeckers,K., Kominek,J., Das,A., Espinosa-Cantú,A., De Maeyer,D., Arslan,A., Van Pee,M., van der Zande,E., Meert,W., Yang,Y. *et al.* (2015) Adaptation to high ethanol reveals complex evolutionary pathways. *PLoS Genet.*, **11**, e1005635.
2. Steenackers,H.P., Parijs,I., Foster,K.R. and Vanderleyden,J. (2016) Experimental evolution in biofilm populations. *FEMS Microbiol. Rev.*, **40**, 373–397.
3. Bailey,M.H., Tokheim,C., Porta-Pardo,E., Sengupta,S., Bertrand,D., Weerasinghe,A., Colaprico,A., Wendl,M.C., Kim,J., Reardon,B. *et al.* (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell*, **173**, 371–385.
4. Woods,R.J., Barrick,J.E., Cooper,T.F., Shrestha,U., Kauth,M.R. and Lenski,R.E. (2011) Second-order selection for evolvability in a large Escherichia coli population. *Science*, **331**, 1433–1436.
5. Khan,A.I., Dinh,D.M., Schneider,D., Lenski,R.E. and Cooper,T.F. (2011) Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, **332**, 1193–1196.
6. Kryazhimskiy,S., Draghi,J.A. and Plotkin,J.B. (2011) In evolution, the sum is less than its parts. *Science*, **332**, 1160–1161.
7. Diard,M., Garcia,V., Maier,L., Remus-Emsermann,M.N.P., Regoes,R.R., Ackermann,M. and Hardt,W.D. (2013) Stabilization of cooperative virulence by the expression of an avirulent phenotype. *Nature*, **494**, 353–356.
8. Plucain,J., Hindré,T., Le Gac,M., Tenaillon,O., Cruveiller,S., Médigue,C., Leiby,N., Harcombe,W.R., Marx,C.J., Lenski,R.E. *et al.* (2014) Epistasis and allele specificity in the emergence of a stable polymorphism in Escherichia coli. *Science*, **343**, 1366–1369.
9. Van den Bergh,B., Swings,T., Fauvart,M. and Michiels,J. (2018) Experimental design, population dynamics, and diversity in microbial experimental evolution. *Microbiol. Mol. Biol. Rev.*, **82**, e00008-18.
10. Dragosits,M. and Mattanovich,D. (2013) Adaptive laboratory evolution - principles and applications for biotechnology. *Microb. Cell Fact.*, **12**, 64.
11. Barrick,J.E. and Lenski,R.E. (2013) Genome dynamics during experimental evolution. *Nat. Rev. Genet.*, **14**, 827–839.
12. Sniegowski,P.D., Gerrish,P.J. and Lenski,R.E. (1997) Evolution of high mutation rates in experimental populations of E. coli. *Nature*, **387**, 703–705.
13. Wielgoss,S., Barrick,J.E., Tenaillon,O., Wiser,M.J., Dittmar,W.J., Cruveiller,S., Chane-Woon-Ming,B., Meédigue,C., Lenski,R.E. and Schneider,D. (2013) Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 222–227.
14. Tenaillon,O., Rodríguez-Verdugo,A., Gaut,R.L., McDonald,P., Bennett,A.F., Long,A.D. and Gaut,B.S. (2012) The molecular diversity of adaptive convergence. *Science*, **335**, 457–461.
15. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G. V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
16. Tokheim,C.J., Vogelstein,B., Papadopoulos,N., Kinzler,K.W. and Karchin,R. (2016) Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 14330–14335.
17. Swings,T., Weytjens,B., Schalck,T., Bonte,C., Verstraeten,N., Michiels,J. and Marchal,K. (2017) Network-based identification of adaptive pathways in evolved ethanol-tolerant bacterial populations. *Mol. Biol. Evol.*, **34**, 2927–2943.
18. Leiserson,M.D.M., Vandin,F., Wu,H.-T., Dobson,J.R., Eldridge,J. V, Thomas,J.L., Papoutsaki,A., Kim,Y., Niu,B., McLellan,M. *et al.* (2014) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.
19. De Maeyer,D., Weytjens,B., De Raedt,L. and Marchal,K. (2016) Network-based analysis of eQTL data to prioritize driver mutations. *Genome Biol. Evol.*, **8**, 481–494.
20. Reyna,M.A., Haan,D., Paczkowska,M., Verbeke,L.P.C., Valencia,A., Reimand,J., Stuart,J.M., Raphael,B.J. *et al.* (2018) Pathway and network analysis of more than 2,500 whole cancer genomes. bioRxiv doi: https://doi.org/10.1101/385294, 07 August 2018, preprint: not peer reviewed.

21. Verbeke,L.P.C., Van Den Eynden,J., Fierro,A.C., Demeester,P., Fostier,J. and Marchal,K. (2015) Pathway relevance ranking for tumor samples through network-based data integration. *PLoS One*, **10**, e0133503.
22. Dimitrakopoulos,C.M. and Beerenwinkel,N. (2017) Computational approaches for the identification of cancer genes and pathways. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **9**, e1364.
23. Le Van,T., Van Leeuwen,M., Carolina Fierro,A., De Maeyer,D., Van Den Eynden,J., Verbeke,L., De Raedt,L., Marchal,K. and Nijssen,S. (2016) Simultaneous discovery of cancer subtypes and subtype features by molecular data integration. *Bioinformatics*, **32**, i445–i454.
24. De Maeyer,D., Renkens,J., Cloots,L., De Raedt,L. and Marchal,K. (2015) PheNetic: network-based interpretation of unstructured gene lists in E. coli. *Mol. BioSyst.*, **9**, 1594–1603.
25. De Maeyer,D., Weytjens,B., Renkens,J., De Raedt,L. and Marchal,K. (2015) PheNetic: network-based interpretation of molecular profiling data. *Nucleic Acids Res.*, **43**, W244–W250.
26. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
27. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
28. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
29. Vaser,R., Adusumalli,S., Leng,S.N., Sikic,M. and Ng,P.C. (2016) SIFT missense predictions for genomes. *Nat. Protoc.*, **11**, 1–9.
30. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
31. Wilm,A., Aw,P.P.K., Bertrand,D., Yeo,G.H.T., Ong,S.H., Wong,C.H., Khor,C.C., Petric,R., Hibberd,M.L. and Nagarajan,N. (2012) LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.*, **40**, 11189–11201.
32. Swings,T., van Den Bergh,B., Wuyts,S., Oeyen,E., Voordeckers,K., Verstrepen,K.J., Fauvart,M., Verstraeten,N. and Michiels,J. (2017) Adaptive tuning of mutation rates allows fast response to lethal stress in escherichia coli. *Elife*, **6**, e22939.
33. Jerison,E.R., Kryazhimskiy,S., Mitchell,J.K., Bloom,J.S., Kruglyak,L. and Desai,M.M. (2017) Genetic variation in adaptability and pleiotropy in budding yeast. *Elife*, **6**, e27167.