Detecting gene subnetworks under selection in biological pathways

Alexandre Gouy[1,2]*, Joséphine T. Daub[3] and Laurent Excoffier[1,2]

[1] Institute of Ecology and Evolution, University of Berne, Baltzerstrasse 6, 3012 Berne, Switzerland
[2] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[3] Institute of Evolutionary Biology, Universitat Pompeu Fabra – CSIC, 08003 Barcelona, Spain

* To whom correspondence should be addressed. Tel: +41 31 631 30 28; Email: alexandre.gouy@iee.unibe.ch

ABSTRACT

Advances in high throughput sequencing technologies have created a gap between data production and functional data analysis. Indeed, phenotypes result from interactions between numerous genes, but traditional methods treat loci independently, missing important knowledge brought by network-level emerging properties. Therefore, evidencing selection acting on multiple genes affecting the evolution of complex traits remains challenging. In this context, gene network analysis provides a powerful framework to study the evolution of adaptive traits and facilitates the interpretation of genome-wide data. To tackle this problem, we developed a method to analyse gene networks that is suitable to evidence polygenic selection. The general idea is to search biological pathways for subnetworks of genes that directly interact with each other and that present unusual evolutionary features. Subnetwork search is a typical combinatorial optimization problem that we solve using a simulated annealing approach. We have applied our methodology to find signals of adaptation to high-altitude in human populations. We show that this adaptation has a clear polygenic basis and is influenced by many genetic components. Our approach improves on classical tests for selection based on single genes by identifying both new candidate genes and new biological processes involved in adaptation to altitude.

INTRODUCTION

Understanding the genetic basis of adaptation remains a central theme of evolutionary biology. Adaptation is typically viewed as involving selective sweeps that drive beneficial alleles from low to high frequencies in a population, lowering genetic diversity and increasing linkage disequilibrium near the selected region (1-3). Numerous statistical tests have been developed to detect selection from genomic data based on a simple selective sweep model (reviewed in De Mita et al. (4)). Therefore, most work in humans and other species has focused on identifying signals of strong selection at individual loci (5). These methods have been quite successful in humans to identify loci involved in several adaptations such as diet, altitude, disease resistance, and pigmentation (reviewed in Vitti et al. (6)). However, examples of adaptation due to a selective sweep at a single locus remain relatively rare in human populations. Therefore, some authors have argued that adaptation events could occur by the evolution of polygenic traits rather than via the fixation of single beneficial mutations (7-9).

Recent Genome-Wide Association Studies (GWAS) in various model organisms have confirmed that variation at many important traits is controlled by a large number of loci scattered throughout the genome, e.g. human height (10,11). Selection acting additively on this kind of traits could therefore

lead to small shifts in allele frequencies (8). This verbal model has been studied analytically, showing that in some cases, polygenic selection may indeed lead to subtle shifts in allele frequencies (12-14). However, these small allele frequency changes may remain below the detection limit of most of outlier detection methods (15). Therefore, the generality of conclusions drawn from significant tests can be seriously challenged because phenotypic traits exhibiting clear-cut molecular signatures of selection may represent a biased subset of all adaptive traits (16). Another caveat of classical genome scans for selection is that lists of candidate genes are sometimes difficult to connect to a particular adaptive mechanism, since SNP-level results are unlikely to reveal complex mechanisms of adaptation, given the lack of signal of small-effect alleles. It seems therefore necessary to consider alternative approaches to study the genetic basis of adaptation of complex traits.

Current approaches to detect selection acting on polygenic traits rely mostly on quantitative genetics models. Classical quantitative genetics approaches are not based on genetic data, but on an explicit description of continuous phenotypes (e.g. height, body mass index, fertility, etc.). These methods have strong theoretical foundations, and allow one to disentangle the genetic from the environmental variance by taking into account the heritability of the traits, and therefore to detect shifts in the distribution of the phenotype under selection (17). But these methods do not permit to identify the genetic basis of adaptation, and other approaches must be considered to associate genetic data to quantitative traits responding to selection. Correlative approaches have emerged where associations between a genotype and various environmental variables are tested (e.g. (18-21)). Finally, recent approaches have tried to estimate selection coefficients from GWAS data (9,22,23), but all these methods need some phenotypic measures of the tested individuals or associated environmental data, which can be sometimes difficult to obtain.

In contrast to a gene-centric approach, some studies have considered testing if a set of genes as a whole is yielding signals of selection (7,24). Indeed, different genes within pathways (i.e. molecular networks leading to a given biological function) may interact to produce a given phenotype (25,26), and therefore be under the same selective pressure. Finding sets of outlier interacting genes can be achieved using gene-set enrichment methods (e.g., (27,28)). The idea is to assign a score (i.e. proxy for selection) to each gene within a biological pathway (i.e. gene-set) and to test if the distribution of scores within the pathway is significantly shifted towards extreme values (7). This approach has successfully identified candidate pathways involved in various human adaptations, such as response to pathogens (7), or adaptation to altitude (24). However, this gene-set enrichment approach mainly identifies pathways where all its members show a shift in the distribution of a given tested statistic. It might thus be underpowered to find more subtle signals, where only a subset of genes is under selection in a large pathway, which is a more likely situation than assuming that all the genes in a pathway have responded to selection.

To address this problem, network analysis can provide new insights into the genetics basis of adaptation. In the last few years, network-based approaches have spread into a large number of research areas, and were successfully used to solve a wide range of biological problems; e.g. gene expression studies (29,30), GWAS (26) or evolutionary biology (31,32). Here, we present a new network-based method to detect polygenic selection in natural populations. The general idea is to

search for subnetworks of interacting genes within biological pathways that present unusual features. This search is a typical combinatorial optimization problem that can be solved using a heuristic approach like simulated annealing (29,33). We implemented such an algorithm to search for high-scoring subnetworks of genes in biological pathways, and we developed a testing procedure that explicitly takes into account the optimization process involved in this search. After studying the sensitivity and precision of our method with simulated data, we reanalysed data from a previous study looking for convergent adaptation to altitude in Tibetans an Andeans (24). As compared to the original study, we discover new genes and biochemical functions potentially related to adaptation in these human populations. Our method can thus complement classical genome scans by providing functional information and discovering new genes with weaker effects that are involved in complex selective processes. Finally, we discuss the limits and potential improvements, as well as other possible applications of our methodology.

MATERIAL AND METHODS

Pathway databases and conversion to gene networks

We considered biological pathways as gene networks. More formally, we define a gene network as a graph $G(V,E)$, where V is a set of nodes (i.e. genes), and E is a set of edges (i.e. interactions between genes). In this study we used three signalling and metabolic pathway databases that are considered as references in systems biology: (i) KEGG, the Kyoto Encyclopaedia of Genes and Genomes Pathway database (34); (ii) NCI, the National Cancer Institute / Nature Pathway Interaction Database (35); and (iii) Reactome (36,37). We then used the R/Bioconductor *graphite* package to convert biological pathways into graphs of interacting genes (see (38) for more details on this procedure).

Computation of summary statistics on gene networks

To characterise the structure of networks and check for potential differences between databases, we generated the distributions of three standard summary statistics for each of the three databases. We thus computed for each network i) the number of nodes, ii) the number of edges, and iii) the graph density. The graph density d is a measure of connectivity between the nodes of the network, and it is defined as the number of edges in a set E compared to the maximum number of possible edges between nodes in a set V, therefore $d = 2 * |E| / (|V| * (|V| - 1))$, where $|X|$ represents the number of members of a set X. We also analysed the overlap between pathway databases by computing the number of genes they share. Finally, we quantified the redundancy between pathways within a database by computing Jaccard's similarity index. For a pair of networks A and B with sets of nodes V_A and V_B, Jaccard's index is defined as $J_{AB} = (|V_A| \cap |V_B|) / (|V_A| \cup |V_B|)$.

Workflow to detect outlier subnetworks

As the detection of outlier subnetworks includes several distinct steps, we describe here our analysis pipeline. The goal of our approach is to search within each gene network the subnetwork with the largest signal of interest (e.g. evidence of selection) using a simulated annealing approach (33). Our algorithm is globally similar to that used by Ideker et al. (29), but our method differs in two important

ways, as described below. First, whereas Ideker et al.'s method aimed at finding the highest-scoring *subset* of nodes, we aim here at finding the highest-scoring *subnetwork* (i.e. a subset of genes that are directly connected by edges). Second, we consider a statistical testing procedure that explicitly considers the optimization procedure when computing the p-value of a given subnetwork. Indeed, the score of a given subnetwork identified by the simulated annealing algorithm cannot be compared to that of a random subnetwork, as simulated annealing would identify a high scoring subnetwork even in absence of any true signal (39).

Gene and subnetwork scores

We begin our testing procedure by assigning a score to each of the gene (node) in our network. In population genetics applications looking for subsets of selected genes, this score might be a measure of population differentiation between populations (e.g. F_{ST}), the result of a selection test, or the difference in some measure between cases and controls. If this score is available for different SNPs in a given gene, we need to summarize their scores in some way, as our method assumes that each gene has a single score. For instance, the SNP with maximum score can be selected to represent a gene, or the average of the SNP-specific scores can be computed over all SNPs assigned to a gene. We then use an aggregate score for a subnetwork of size k following (29) as $s = \sum(x_i) / \sqrt{(k)}$, where x_i is the score of the i-th node (gene).

Subnetwork score normalization

We then normalize the scores of subnetworks such as to be able to compare subnetworks of different sizes. Indeed, we expect to observe less variance in subnetwork aggregate scores in large than in small networks. The score of a given subnetwork of size k is thus normalized as $z'_k = (s_k - \mu_k) / \sigma_k$, where μ_k and σ_k are the mean and standard deviation of the score of a subnetwork of size k, computed empirically over 10,000 random subnetworks of size k, obtained for each data base separately. The means and standard deviations of subnetworks of sizes k_{min} to k_{max} are computed once and stored in a lookup table. Random subnetworks of size k are obtained by i) randomly selecting a network in the database with a probability depending on the network size, ii) randomly selecting a gene from this network as an initial member of the subnetwork, and iii) iteratively adding k-1 other randomly chosen genes that are connected to the growing subnetwork.

Searching for optimal subnetworks with simulated annealing

We have used a simulated annealing algorithm to detect the Highest Scoring Subnetwork (HSS) of each gene network. The general idea is to start with a random subnetwork, and modify it progressively by adding or removing one node at a time until we reach a subnetwork with the highest possible normalized score. The algorithm takes as initial parameters the number of iterations N to perform and the annealing parameter alpha, which determines a temperature function $T(\alpha)$ that decreases geometrically over time.

In more details, our search algorithm is as follows:

1. Select a starting subnetwork of arbitrary size k_{min}, defined at random.

2. Calculate its normalized score z_i.
3. Modify the current subnetwork: First, select a node at random from the following list: i) nodes not belonging to the current subnetwork, but that are connected to it by a single edge, ii) terminal nodes of the current subnetwork, iii) internal nodes of the current subnetwork which are not articulation points (i.e. whose removal will not create two disjoint subnetworks). If the selected node is not part of the current subnetwork then add it, else remove it.
4. Calculate the new subnetwork's normalized score z_{i+1}
5. Accept the new subnetwork with probability min(1,p), where the annealing probability p = exp([z_{i+1} - z_i] / T_{i+1}), and T_{i+1} is the annealing temperature for the iteration i+1. This typical simulated annealing equation means that a new subnetwork is always accepted if its normalized score is larger than that of the previous subnetwork, and that less optimal subnetworks are more and more difficult to accept with more iterations of the algorithm.
6. Repeat steps 3-5 above for a given predefined number of iterations N.
7. Record the resulting subnetwork and its score.

This algorithm is expected to find the global optimum for a sufficient number of iterations (29), but as its performance could vary between networks, we have run it five times for each pathway, and recorded the subnetwork with the highest score.

Statistical testing procedure

To test if the score of the estimated HSS is significantly larger than what would be expected by chance, we need to generate the null distribution of HSS for subnetworks of a given size. To do this, we cannot simply randomly sample subnetworks and compute their scores in the original dataset, as we need to take into account the fact that the optimization procedure will bias the subnetwork scores towards high value. To take this effect into account, we have generated a null distribution of optimised scores. To do this, we first permute gene scores across all networks of a given database. Then a network is randomly chosen with a probability proportional to its size, and the optimisation algorithm is applied to obtain the HSS on the permuted dataset. The score of the resulting HSS is finally recorded. This procedure is repeated N times to generate the null distribution. The empirical p-value of a given observed HSS is then obtained as the proportion of random HSS of similar size that have a score larger or equal to the observed HSS score (unilateral test).

As many subnetworks are tested with our procedure, we have corrected the inferred p-values for multiple testing by computing q-values, which are false discovery rates (FDRs) that would be computed if the observed p-value was used as a threshold to declare significance. To do this, we used the FDR method (40) implemented in the R package *qvalue*.

Pipeline implementation

Our analysis pipeline has been implemented in R, and graphical representations of the networks and HSS were made using the software Cytoscape (41,42), called from R with the Bioconductor package RCytoscape (43).

Test of the method on simulated data

As simulated annealing is an approximate method, we studied its performance using a simulation-based approach. We simulated pseudo-observed data sets by building a random network of size N using a random edge model, i.e. where an edge is drawn with a given probability p. Then, a connected subnetwork of size k is randomly sampled within the network. The node scores from the subnetwork are drawn from a normal distribution $N(\mu_{HSS},1)$, where μ_{HSS} is the average score of this subnetwork. The score of the other nodes of the network are drawn from a standard normal distribution $N(0,1)$. We then apply our simulated annealing algorithm to find the highest-scoring subnetwork using with i iterations. Therefore, the outcome of our search depends on five parameters: the network size N, the HSS size k and its average expected score μ_{HSS}, the network connectivity p and the number of iterations i.

In order to characterize the accuracy of our network search and to better understand which parameters have an impact on our estimation, we computed, for each simulation, the number of true positives (*TP*, the number of nodes from the true HSS that are correctly identified), the number of true negatives (*TN*, the number of nodes that are not in the HSS and that are not identified), false positives (*FP*, the number of nodes wrongly identified as part of the true HSS) and false negatives (*FN*, the number of nodes from the true HSS that have not been identified). We then computed two measures of performance: precision or positive predictive value: $PPV = TP / (TP + FP)$; and sensitivity or true positive rate: $TPR = TP / (TP + FN)$.

To assess the impact of our five parameters on the precision (*PPV*) and on the sensitivity (*TPR*) of our estimation, we used a Generalised Linear Model (GLM) where the response variables are the counts of *TP* and *FP* for precision, and the counts of *TP* and *FN* for sensitivity. The predictor variables are the five above-mentioned parameters, and the error follows a binomial distribution.

To test the performance of our significance testing procedure that explicitly takes the optimisation process into account, we computed p-values using the null distribution obtained from 10,000 runs of simulated annealing on data generated with $\mu_{HSS} = 0$ (i.e. the null hypothesis).

Application to real data: detection of convergent adaptation to altitude in humans

We analysed a dataset published by Foll et al. (24) on convergent adaptation to altitude in Tibetans and Andeans. This data set consists of 632,344 SNPs genotyped in four populations: two populations living at high altitude in the Andes (49 individuals) and in Tibet (49 individuals), as well as two lowland related populations from Central America (39 Mesoamericans) and East Asia (90 individuals). For each SNP, a probability of convergent adaptation has been computed under a hierarchical Bayesian model (24). To get a unique score per gene, as required in our methodology, we used the p-value of the highest scoring SNP mapped within a gene or less than 50 kb away.

We applied our methodology to detect subnetworks under selection on this dataset. The three pathway databases were analyzed separately (i.e. every step of the workflow has been done independently for each database). Pathways for which the largest connected subnetwork size was less than $k_{min} = 5$ nodes were removed from the analysis, since we wanted to avoid focusing on small subnetworks. Aggregate subnetwork score distributions have been generated by sampling 10,000

random subnetworks for each possible size *k*. The HSS search algorithm has been applied to every pathway with *N* = 10,000 simulated annealing iterations. The p-value of the obtained HSSs was inferred from the distribution of scores of 10,000 random HSS generated under the null hypothesis (i.e. permuted data).

RESULTS

Test of the method on simulated data

We first studied the performance of our approach in terms of precision (i.e. the fraction of selected genes in the estimated highest-scoring subnetworks (HSS)) and sensitivity (i.e. the proportion of selected genes that are identified as such) by analysing pseudo-observed data. We generated random networks and HSS based on five parameters (see Material and Methods). We ran our algorithm on the simulated data and compared the estimated HSS to the true HSS. Using logistic regressions, we show that out of the 5 parameters tested, 4 have a significant impact on both the precision and sensitivity of the method (Table 1). Most of the model deviance is explained by the mean score of the selected genes, the network size, and the subnetwork size. As expected, precision goes up with μ_{HSS} and the false positive rate is lower than 0.05 when $\mu_{HSS} > 4$ (Figure 1A). Furthermore, even if network (*N*) and subnetwork (*k*) sizes influence our ability to correctly identify HSS, *N* and *k* have a negligible impact on the precision of our estimations when the true subnetwork score is sufficiently large. Indeed, in this case precision remains high for a broad range of *N* and *k* values (Figure 1B and 1C). Even though one would have thought that the number of iterations of the simulated annealing algorithm was an important parameter for the success of the algorithm, it has a limited impact (Table 1) and 5,000 iterations appear enough to achieve high precision (Table 1). Finally, we find that network density has no real influence on the performance of our method.

Then, in order to verify that our statistical testing procedure behaves properly, we computed the p-value distribution under the null hypothesis of $\mu_{HSS} = 0$. In that case, p-values do not depart significantly from a uniform distribution (Kolmogorov-Smirnov test, $D = 0.03$, $p = 0.76$; Figure S1), which is the behaviour expected when the null hypothesis is true.

Pathways databases characteristics

We used pathways defined in three databases: KEGG, NCI and Reactome (including respectively 225, 189 and 1095 pathways). To see whether we should treat these databases separately or not, we first computed statistics summarizing the main properties of these databases. First, we characterized the overlap between these databases, i.e. the number of genes shared between databases. We show that even if they substantially overlap in their gene content, the three databases have a large number of private genes (Figure S2A). We also characterized the overlap between pathways within databases using Jaccard's index. We computed the redundancy within a database as the proportion of pathways pairs with an overlap higher than a given threshold as a function of this threshold (44). We find that pathways from the three databases have different levels of overlap, with Reactome having the largest fraction on non-overlapping pathways (Figure S2B). Finally, we computed summary statistics to understand the structures of the networks in the different databases. The distributions of the number

of nodes, the number of edges and the connectivity are also strikingly different between the three databases (Figure S3). Since pathways of these three databases had different properties, we have analysed them separately, using genes from each database to build separate null distributions and perform statistical tests.

Adaptation to altitude in humans

We analysed the data from Foll et al. (24), who studied adaptation in two human populations living at high altitude. For each SNP, they computed the probability of convergent adaptation to altitude in Andeans and Tibetans under a hierarchical Bayesian model, and we used this probability as our score. We define the gene score as the highest-scoring SNP within the gene or in a 50 kb surrounding window. The distribution of gene scores appears slightly different between databases (Figure S4), again justifying the separate analysis of the three databases.

To search for high-scoring subnetworks in each pathway, we first generated the aggregate subnetwork score distributions for each database and for all possible subnetwork sizes. We then searched for the high-scoring subnetwork in each pathway using 10,000 simulated annealing iterations, and we assessed their significance from a null distribution of HSSs based on 10,000 permutated data sets (see Material and Methods). Interestingly, we find that subnetwork scores tend to be lower in denser pathways. Indeed, the estimated subnetwork score significantly decreases with the density of a pathway (linear regression, $F(1,1339) = 42.11$, $p = 1.2.10^{-10}$; $R = 0.17$; Figure S4). This result is unlikely to be an artefact, as our simulation study shows that our procedure is not affected by network density (Table 1). Therefore, genes potentially involved in adaptive processes seem to be preferentially found in pathways with less gene-gene interactions. These results are in agreement with other empirical studies that showed that deleterious mutations tend to accumulate at the periphery of gene networks (45). Even though positive selection can also act on genes with more interactions (31,46), this result suggests that adaptation to altitude has mainly targeted genes with less pleiotropic effects since the number of interactions of a gene is clearly correlated to its pleiotropy level (47).

We then considered a HSS as significant if it showed a p-value < 0.01 and a q-value < 0.20. None of the pathways tested in the Reactome database remained significant after multiple test correction. We identified four pathways with a significant HSS in the NCI database and six such subnetworks in KEGG (Table S1). The overall top-scoring pathway is the HIF-2-α transcription network (Figure 2), a pathway containing genes known to respond to hypoxia conditions. EPAS1 (HIF-2-α) is the top-scoring gene, it is a transcription factor active under hypoxic conditions. All the other significant genes within this pathway are directly interacting with EPAS1 and should thus play an important role in response to hypoxia. Some of these genes are inhibitors (CITED2) or cofactors (ARNT) of Hypoxia-Inducible Factors (HIF), others are regulated by HIF, such as VEGFA, a growth factor involved in angiogenesis.

When top-scoring HSSs were overlapping by one or more gene, we merged them in a single network (Figure 3). After this procedure, we observe four distinct clusters of genes. First, in the NCI database, we find a single cluster of genes within four pathways involved in vascular processes such

as angiogenesis, response to hypoxia or blood coagulation (Figure 3A). Among these, the top-scoring genes are Endothelial PAS domain-containing protein 1 (EPAS1), Interleukin-6 (IL6), Angiopoietin 1 (ANGPT1), Pleiotrophin (PTN), Tyrosine-protein phosphatase non-receptor type 11 (PTPN11) and Epidermal Growth Factor (EGFR). We also observe many genes in these HSS that present lower scores. Most of these are growth factors, such as genes in the Insulin Growth Factor (IGF), receptor tyrosine kinases (ErbB, EGFR), Neurotrophic Factors (NTF) or Interleukin (IL) families. We identified three other clusters of genes in the KEGG database that are involved in very different biological processes (Figure 3B). First, a large network of 32 genes involved in metabolic functions where the top-scoring genes are Alcohol Dehydrogenase (ADH) subunits, most of the other genes being other aerobic metabolism related enzymes such as the Glucuronosyltransferase (UGT), Glutathion S-tranferase (GST) or Glutamic-Oxaloacetic Transaminase (GOT) families. All of them present moderate probabilities of convergent adaptation (< 0.8). Second, an immunity-related cluster is observed, including 6 Human Leucocyte Antigen (HLA) genes. A last cluster consists in three genes related to neuronal cell growth, with Neuroligin 4 (NLGN4X) being the top-scoring gene in this database.

DISCUSSION

New insights into human adaptation to altitude

The challenges of living at high altitude impose a very strong selective pressure on individuals, mainly due to low oxygen levels leading to hypoxia (48). Physiological changes have been identified in Tibetans and Andeans living at high altitude (49), and many studies have unveiled the genetic bases of these physiological changes (reviewed in (48)). Adaptation to altitude thus offered us a good positive control to test our new method on real data, and therefore, the fact that our top subnetwork is found in the HIF-2-α transcription pathway is reassuring. This pathway is indeed a key component of the response to hypoxia, as it modulates or induces various physiological responses such as angiogenesis, haemoglobin concentration or erythropoiesis (50). Numerous genes within this pathway have already been proposed to be under selection in Tibetans and Andeans, such as EPAS1 and IL6 (50-53). In addition to these usual suspects, we identify many other genes with scores that remain below the detection threshold of the original genome scan (24), and which show a much more moderate signal of convergent adaptation. The identification of other candidate genes present in the HIF pathway is in line with the view that adaptation to altitude has a polygenic basis (50). For instance, we identified pleiotrophin (PTN), which acts as an angiogenic factor through multiple mechanisms (54), but which has to our knowledge never been identified as a major player in adaptation to altitude. Another gene, PTPN11, has a high score and a central position in one of the significant subnetworks. It encodes the protein tyrosine phosphatase SHP-2, which regulates heart and blood cells development during embryogenesis, as well as other tissues (55). The Cell Adhesion Molecules (CAM) pathway also presents interesting signals, as we have identified a small cluster of 3 genes coding for neuroligins, which are neuronal proteins involved in the modulation of synaptic transmission (56). However, it has recently been shown that these genes are also involved in vascular processes (57,58). The NLGN1 gene thus seems to be a strong candidate for adaptation to high-

altitude in Tibetans and Andeans and mechanisms linked to neuroligins action in angiogenesis at high altitude would deserve further investigation.

Note that we have also identified a cluster of genes involved in separate metabolic processes. Signals of adaptation at ADH and ALDH genes have been observed in the original study as well as in Ethiopian populations living at high altitude (59). As suggested in the original study, these genes could be involved in fatty-acid degradation and energy production in the mitochondrion: in case of hypoxia, alternative pathways such as omega-oxidation (including ADH genes) could be an alternative to beta-oxidation (24).

Advantages and limitations of the method

The search for high-scoring subnetworks is a combinatorial optimization problem for which several methods have been developed (29,39,60). Here, we describe a new method to detect selection in biological pathways based on a simulated annealing algorithm that extends a previous approach (29) by searching for the highest scoring *subnetwork* of interacting genes rather than for the highest scoring subset of nodes, i.e. we constrain the search to a single connected set of genes. Even though an exact algorithm has been developed to find the optimal subnetwork, it is not generally applicable, as it can only be applied to a list of p-values coming from a mixture of beta distribution (39). On the other hand, our simulated annealing method does not require any assumption on the distribution of gene scores, and it can therefore be applied to a wider range of problems. In addition, our statistical testing procedure explicitly takes into account the optimization procedure, by building a null distribution of high-scoring subnetworks in permuted data. The generation of this null distribution is a crucial step to prevent simulated annealing to identify subnetworks in the absence of any signal (39), and we show here by simulation that our statistical procedure behaves properly in terms of type I error (Figure 1 and S1).

An interesting feature of our approach is the integration of functional information into the analysis by directly testing biologically relevant gene sets. This procedure allows one to better interpret the output of a genome scan and to find the potential functions that are involved in the adaptive process. This is in clear contrast with traditional Gene Ontology (GO) analyses (61) that are typically performed on a list of top scoring candidate genes. For instance, a GO analysis performed in the original study (24) reveals only 2 significant GO terms: ‘‘ethanol oxidation’’ (GO:0006069) and ‘‘positive regulation of transmission of nerve impulse’’ (GO:0051971). This GO analysis thus missed some important biological processes involved in adaptation to altitude. Note that 14 of the 72 genes initially identified as candidates for adaptation to altitude are also present in our significant HSSs (Figure 3). 46 of the 72 top scoring genes are not included in our current analysis, either because they are absent from the pathways databases (n = 31), or because no SNP could be associated to them (e.g. because the closest SNP is more than 50 Kb away from them) (n = 15). Therefore, only 26 genes were identifiable with our method. The fact that a large fraction (46%) of these genes are not present in our significant HSSs and that only 4 out of 9 HSSs include a top-scoring gene shows that our method is not just agglomerating less significant genes around top scoring genes. In any case, our results seem biologically more relevant than a GO analysis output and in this case easier to interpret.

Our approach is conceptually close to the method developed by Daub et al. (7), which consisted in testing if a whole pathway presented a shift in the gene score distribution. The main difference with this previous approach is that we aim here at finding high-scoring subnetworks within pathways. Indeed, it is more likely that polygenic adaptive events have focused on only a subset of genes rather than on a whole pathway. In addition to be able to identify small subsets of genes even in large pathways, our approach allows one to identify outlier functions and genes at the same time, whereas under the previous whole-pathway approach, pathways had to be manually inspected in order to know which genes were driving adaptation (7). However, Daub et al.'s approach (7) has some advantages as it can be applied to any pathway, as it is not limited to pathways for which gene interaction networks are explicitly available. Therefore, the two approaches should be seen as complementary.

Whereas the present methodology overcomes some common problems associated to genome scans for selection, such as being able to identify genes with moderate selection score (Figure 2 and 3), and to explicitly associate candidate gene-sets to biological functions, it also presents some limitations as compared to other methods to detect selection. For instance, our approach is limited by the availability of pathway and network information. Therefore, some genes and biological functions cannot be tested, and the method is not easily applicable to non-model species for which no pathways databases are available. Then, one should be aware that only a certain type of biological functions are tested, i.e. biochemical phenotypes, and we thus have no information about higher-order phenotypes, e.g. height or weight. Finally, this method does not allow one to identify isolated top-scoring genes. However, such isolated outlier genes are easily identified with a classical genome scan. One can thus check if outliers are represented among significant subnetworks and therefore determine if selection has only targeted these single genes of if higher-order processes have been the target of selection.

Overall, our method allowed us to study an example of human adaptation from a gene network perspective. Based on information about gene interactions and a proxy for selection, we were able to identify potential undiscovered targets of selection, like pleiotrophin or neuroligins. This method has thus the potential to detect new genetic bases of adaptation in humans, as well as in other species for which gene interactions databases exist or could be inferred. In addition, even though we have applied this search algorithm to a case of human evolution, the same workflow can be used in other fields, such as for the study of differential gene expression, GWAS or any kind of analysis for which a score can be obtained for any given gene.

AVAILABILITY

Our approach has been implemented as a fully automated R package. The source code and documentation are available on Github (http://www.github.com/CMPG/signet).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We would like to thank Isabelle Dupanloup for bioinformatics support.

FUNDING

This work was partially supported by a Swiss National Science Foundation grant [310030B-166605 to L.E.].

REFERENCES

1. Maynard-Smith, J. and Haigh, J. (1974) Hitch-Hiking Effect of a Favorable Gene. *Genet. Res.*, **23**, 23-35.

2. Kaplan, N.L., Hudson, R.R. and Langley, C.H. (1989) The Hitchhiking Effect Revisited. *Genetics*, **123**, 887-899.

3. Stephan, W., Wiehe, T.H.E. and Lenz, M.W. (1992) The Effect of Strongly Selected Substitutions on Neutral Polymorphism - Analytical Results Based on Diffusion-Theory. *Theor. Popul. Biol.*, **41**, 237-254.

4. De Mita, S., Thuillet, A.C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J. and Vigouroux, Y. (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.*, **22**, 1383-1399.

5. Vitti, J.J., Grossman, S.R. and Sabeti, P.C. (2013) Detecting Natural Selection in Genomic Data. *Annu. Rev. Genet.*, **47**, 97-120.

6. Wollstein, A. and Stephan, W. (2015) Inferring positive selection in humans from genomic data. *Investig. Genet.*, **6**, 5.

7. Daub, J.T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M. and Excoffier, L. (2013) Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Mol. Biol. Evol.*, **30**, 1544-1558.

8. Pritchard, J.K. and Di Rienzo, A. (2010) Adaptation - not by sweeps alone. *Nat. Rev. Genet.*, **11**, 665-667.

9. Field, Y., Boyle, E.A., Telis, N., Gao, Z.Y., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M.I. *et al.* (2016) Detection of human adaptation during the past 2000 years. *Science*, **354**, 760-764.

10. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chun, A.Y., Estrada, K., Luan, J., Kutalik, Z. *et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, **46**, 1173-1186.

11. Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine, R.S., Lu, Y.C., Schurmann, C., Highland, H.M. *et al.* (2017) Rare and low-frequency coding variants alter human adult height. *Nature*, **542**, 186-190.

12. de Vladar, H.P. and Barton, N. (2014) Stability and Response of Polygenic Traits to Stabilizing Selection and Mutation. *Genetics*, **197**, 749-767.

13. Jain, K. and Stephan, W. (2015) Response of Polygenic Traits Under Stabilizing Selection and Mutation When Loci Have Unequal Effects. *G3 (Bethesda)*, **5**, 1065-1074.

14. Stephan, W. (2016) Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol. Ecol.*, **25**, 79-88.

15. Le Corre, V. and Kremer, A. (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Mol. Ecol.*, **21**, 1548-1566.

16. Chevin, L.M. and Hospital, F. (2008) Selective Sweep at a Quantitative Trait Locus in the Presence of Background Genetic Variation. *Genetics*, **180**, 1645-1660.

17. Visscher, P.M., Hill, W.G. and Wray, N.R. (2008) Heritability in the genomics era - concepts and misconceptions. *Nat. Rev. Genet.*, **9**, 255-266.

18. Coop, G., Witonsky, D., Di Rienzo, A. and Pritchard, J.K. (2010) Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, **185**, 1411-1423.

19. Gunther, T. and Coop, G. (2013) Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, **195**, 205-220.

20. Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F. and Bergelson, J. (2011) Adaptation to Climate Across the Arabidopsis thaliana Genome. *Science*, **334**, 83-86.

21. Hancock, A.M., Witonsky, D.B., Ehler, E., Alkorta-Aranburu, G., Beall, C., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J., Coop, G. *et al.* (2010) Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *P. Natl. Acad. Sci. USA*, **107**, 8924-8930.

22. Berg, J.J. and Coop, G. (2014) A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet.*, **10**, e1004412.

23. Racimo, F. and Schraiber, J.G. (2014) Approximation to the Distribution of Fitness Effects across Functional Categories in Human Segregating Polymorphisms. *PLoS Genet.*, **10**, e1004697.

24. Foll, M., Gaggiotti, O.E., Daub, J.T., Vatsiou, A. and Excoffier, L. (2014) Widespread Signals of Convergent Adaptation to High Altitude in Asia and America. *Am. J. Hum. Genet.*, **95**, 394-407.

25. McClellan, J. and King, M.C. (2010) Genetic Heterogeneity in Human Disease. *Cell*, **141**, 210-217.

26. Nakka, P., Raphael, B.J. and Ramachandran, S. (2016) Gene and Network Analysis of Common Variants Reveals Novel Associations in Multiple Complex Diseases. *Genetics*, **204**, 783-798.

27. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *P. Natl. Acad. Sci. USA*, **102**, 15545-15550.

28. Tintle, N.L. (2009) Gene Set analysis in Genome-wide Association Studies. *Genet. Epidemiol.*, **33**, 805-806.

29. Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18(suppl 1)**, S233-S240.

30. Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichita, C. and Draghici, S. (2013) Methods and approaches in the topology-based analysis of biological pathways. *Front. Physiol.*, **4**, 278.

31. Luisi, P., Alvarez-Ponce, D., Pybus, M., Fares, M.A., Bertranpetit, J. and Laayouni, H. (2015) Recent Positive Selection Has Acted on Genes Encoding Proteins with More Interactions within the Whole Human Interactome. *Genome Biol. Evol.*, **7**, 1141-1154.

32. Chakraborty, S. and Alvarez-Ponce, D. (2016) Positive Selection and Centrality in the Yeast and Fly Protein-Protein Interaction Networks. *BioMed Res. Int.*, **vol. 2016**.

33. Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983) Optimization by Simulated Annealing. *Science*, **220**, 671-680.

34. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353-D361.

35. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. and Buetow, K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674-D679.

36. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472-D477.

37. Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S. *et al.* (2016) The Reactome pathway Knowledgebase. *Nucleic Acids Res.*, **44**, D481-D487.

38. Sales, G., Calura, E., Cavalieri, D. and Romualdi, C. (2012) graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, **13**, 20.

39. Dittrich, M.T., Klau, G.W., Rosenwald, A., Dandekar, T. and Muller, T. (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, **24**, I223-I231.

40. Storey, J.D. and Tibshirani, R. (2003) Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol. Biol.*, **224**, 149-157.

41. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498-2504.

42. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. and Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431-432.

43. Shannon, P.T., Grimes, M., Kutlu, B., Bot, J.J. and Galas, D.J. (2013) RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics*, **14**, 217.

44. Vivar, J.C., Pemu, P., McPherson, R. and Ghosh, S. (2013) Redundancy Control in Pathway Databases (ReCiPa): An Application for Improving Gene-Set Enrichment Analysis in Omics Studies and "Big Data" Biology. *OMICS*, **17**, 414-422.

45. Garcia-Alonso, L., Jimenez-Almazan, J., Carbonell-Caballero, J., Vela-Boza, A., Santoyo-Lopez, J., Antinolo, G. and Dopazo, J. (2014) The role of the interactome in the maintenance of deleterious variability in human populations. *Mol. Syst. Biol.*, **10**, 752.

46. Luisi, P., Alvarez-Ponce, D., Dall'Olio, G.M., Sikora, M., Bertranpetit, J. and Laayouni, H. (2012) Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. *Mol. Biol. Evol.*, **29**, 1379-1392.

47. Tyler, A.L., Asselbergs, F.W., Williams, S.M. and Moore, J.H. (2009) Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays*, **31**, 220-227.

48. Bigham, A.W. (2016) Genetics of human origin and evolution: high-altitude adaptations. *Curr. Opin. Genet. Dev.*, **41**, 8-13.

49. Beall, C.M. (2007) Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *P. Natl. Acad. Sci. USA*, **104**, 8655-8660.

50. Bigham, A.W. and Lee, F.S. (2014) Human high-altitude adaptation: forward genetics meets the HIF pathway. *Genes Dev.*, **28**, 2189-2204.

51. Beall, C.M., Cavalleri, G.L., Deng, L.B., Elston, R.C., Gao, Y., Knight, J., Li, C.H., Li, J.C., Liang, Y., McCormack, M. *et al.* (2010) Natural selection on EPAS1 (HIF2 alpha) associated with low hemoglobin concentration in Tibetan highlanders. *P. Natl. Acad. Sci. USA*, **107**, 11459-11464.

52. Huerta-Sanchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M.Z., Somel, M. *et al.* (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, **512**, 194-197.

53. Peng, Y., Yang, Z.H., Zhang, H., Cui, C.Y., Qi, X.B., Luo, X.J., Tao, X.A., Wu, T.Y., Ouzhuluobu, Basang *et al.* (2011) Genetic Variations in Tibetan Populations and High-Altitude Adaptation at the Himalayas. *Mol. Biol. Evol.*, **28**, 1075-1081.

54. Perez-Pinera, P., Berenson, J.R. and Deuel, T.F. (2008) Pleiotrophin, a multifunctional angiogenic factor: mechanisms and pathways in normal and pathological angiogenesis. *Curr. Opin. Hematol.*, **15**, 210-214.

55. Chan, G., Cheung, L.S., Yang, W.T., Milyavsky, M., Sanders, A.D., Gu, S.Q., Hong, W.X., Liu, A.X., Wang, X.N., Barbara, M. *et al.* (2011) Essential role for Ptpn11 in survival of hematopoietic stem and progenitor cells. *Blood*, **117**, 4253-4261.

56. Craig, A.M. and Kang, Y. (2007) Neurexin-neuroligin signaling in synapse development. *Curr. Opin. Neurobiol.*, **17**, 43-52.

57. Bottos, A., Destro, E., Rissone, A., Graziano, S., Cordara, G., Assenzio, B., Cera, M.R., Mascia, L., Bussolino, F. and Arese, M. (2009) The synaptic proteins neurexins and neuroligins are widely expressed in the vascular system and contribute to its functions. *P. Natl. Acad. Sci. USA*, **106**, 20782-20787.

58. Samarelli, A.V., Riccitelli, E., Bizzozero, L., Silveira, T.N., Seano, G., Pergolizzi, M., Vitagliano, G., Cascone, I., Carpentier, G., Bottos, A. *et al.* (2014) Neuroligin 1 induces blood vessel maturation by cooperating with the α6 integrin. *J. Biol. Chem.*, **289**, 19466-19476.

59. Huerta-Sanchez, E., DeGiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., Cardona, A., Montgomery, H.E., Cavalleri, G.L., Robbins, P.A. *et al.* (2013) Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. *Mol. Biol. Evol.*, **30**, 1877-1888.

60. Garcia-Alonso, L., Alonso, R., Vidal, E., Amadoz, A., de Maria, A., Minguez, P., Medina, I. and Dopazo, J. (2012) Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic Acids Res.*, **40**, e158.

61. The Gene Ontology, C. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331-D338.

Table 1: Estimates of the effects of the five parameters on the precision (PPV) and sensitivity (TPR) obtained under a logistic regression framework. For each parameter, the coefficient, p-value and the percentage of explained total deviance (%TD) are indicated.

	PPV			TPR		
	Estimate	p-value	%TD	Estimate	p-value	%TD
N [1]	-0.025	$< 2.10^{-16}$	17.5	$4.3.10^{-3}$	$< 2.10^{-16}$	< 1
k [2]	0.14	$< 2.10^{-16}$	21.3	-0.03	$1.4.10^{-15}$	< 1
μ_{HSS} [3]	0.75	$< 2.10^{-16}$	45.8	1.29	$< 2.10^{-16}$	66
d [4]	-0.029	0.29	< 1	5.10^{-2}	0.31	< 1
i [5]	$-1.8.10^{-6}$	2.10^{-5}	< 1	$1.5.10^{-6}$	0.05	< 1

[1]Network size; [2]HSS size; [3]HSS mean score; [4]Network density; [5]Number of iterations

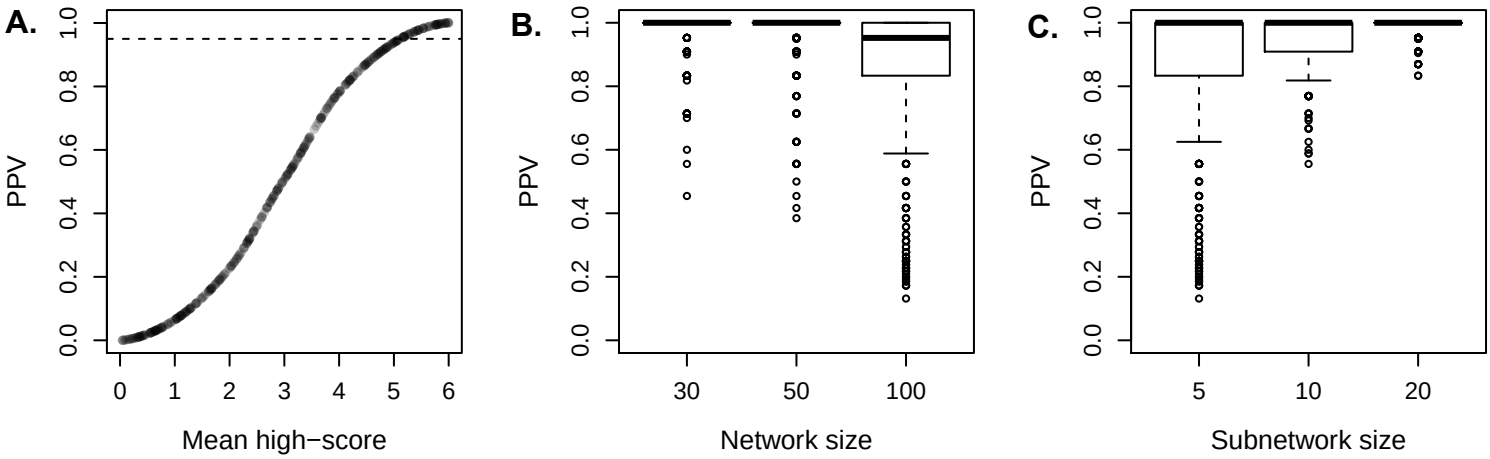

Figure 1: Impact of different parameters on the precision of the estimation (PPV). The predictions of the GLM for the influence of μ_{HSS} on the precision is represented (A), as well as the PPV as a function of network size (B) and subnetwork size (C) when μ_{HSS} is fixed to 5.

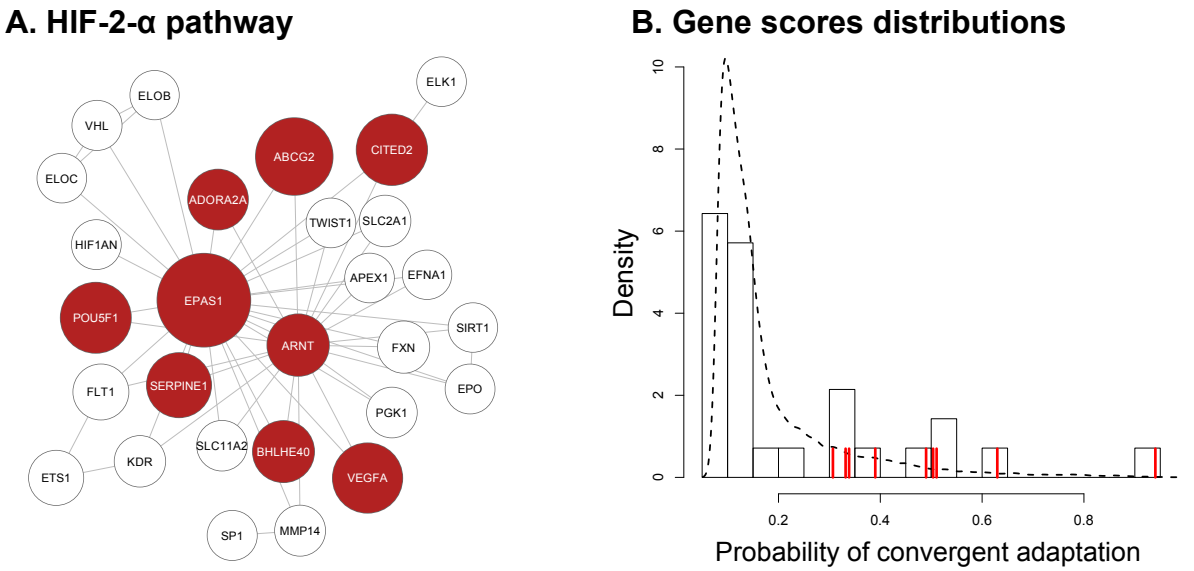

Figure 2: Most significant subnetwork among the three pathway databases. The HIF-2-α transcription pathway is represented as a graph (A), where each node is a gene, and the node size is proportional to the gene score. The highest scoring subnetwork (HSS) of the pathway is shown in red. The gene scores density distribution in this pathway is shown in (B). The dashed line represents the density of gene scores within all the KEGG database, the histogram shows the distribution of genes scores within this pathway, and the vertical red lines indicate the scores of the genes belonging to the HSS.

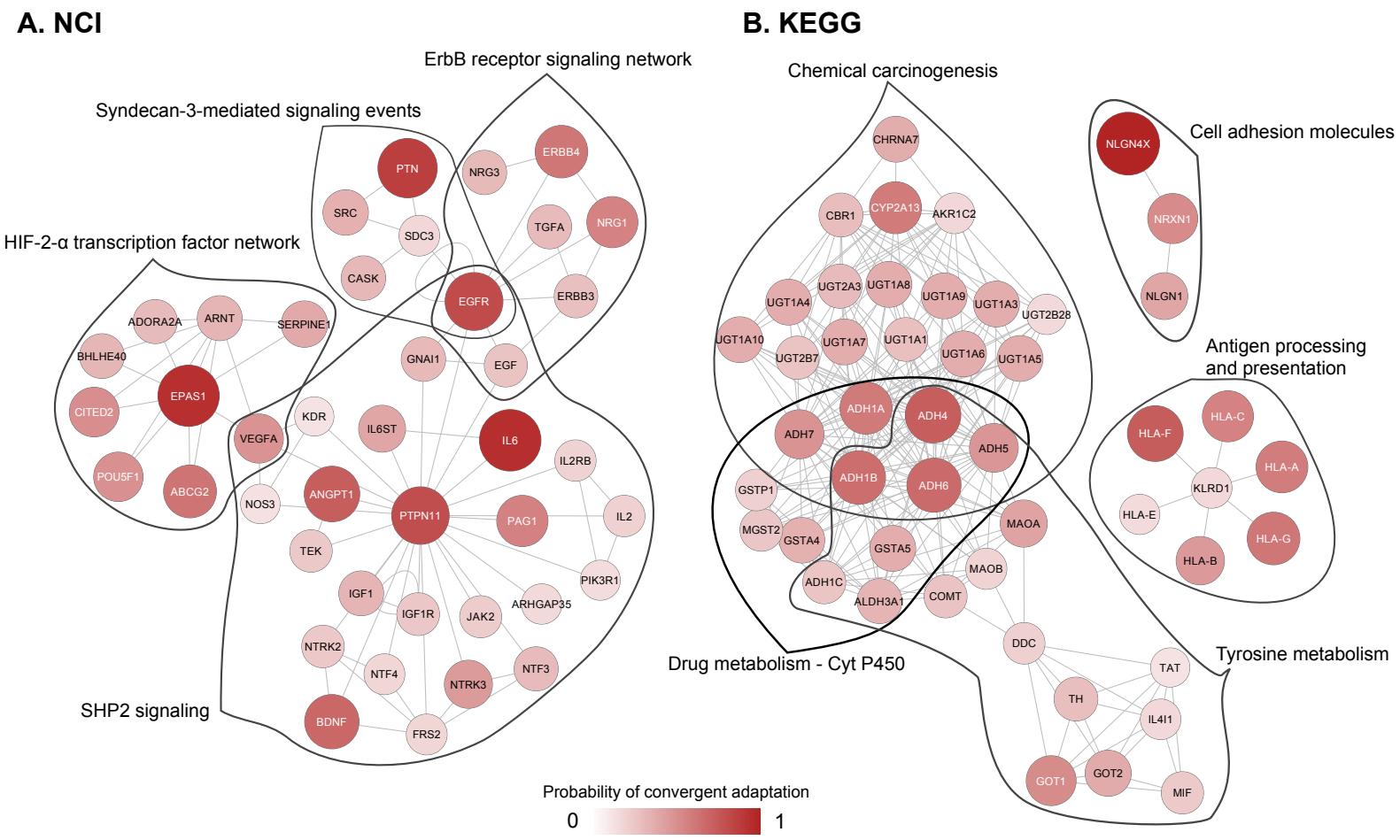

Figure 3: Merged significant subnetworks. For each database, NCI (A) and KEGG (B), we merged the significant subnetworks of genes if they overlapped. The colour intensity and size of the nodes are proportional to the gene score. Red lines delimit the individual significant subnetwork and the names of pathways to which they belong are shown next to it.