Human Genetics

Systematic identification of genetic systems associated with phenotypes in patients with rare genomic copy number variations

--Manuscript Draft--

Manuscript Number:	HUGE-D-20-00020
Full Title:	Systematic identification of genetic systems associated with phenotypes in patients with rare genomic copy number variations
Article Type:	Original Article
Funding Information:	Fundación Progreso y Salud (PI-0075-2017) — Not applicable Instituto de Salud Carlos III (SAF2016-78041-C2-1-R) — Dr Juan Antonio Garcia Ranea Instituto de Salud Carlos III (SAF2016-78041-C2-2-R) — Dr Florencio Pazos Junta de Andalucía (CTS-486) — Dr Elena Rojano Fundación Ramón Areces — Dr Juan Antonio Garcia Ranea
Abstract:	Background: Copy number variation (CNV) related disorders tend to show complex phenotypic profiles that do not match known diseases. This makes it difficult to ascertain their underlying molecular basis. A potential solution is to compare the affected genomic regions for multiple patients that share a pathological phenotype, looking for commonalities. Here we present a novel approach to associate phenotypes with functional systems, in terms of GO categories and KEGG and Reactome pathways, based on patient data. Methods: The approach uses genomic and phenomic data from the same patients, finding shared genomic regions between patients with similar phenotypes. These regions are mapped to genes to find associated functional systems. Results: We applied the approach to analyse patients in the DECIPHER database with \textit{de novo} CNVs, finding functional systems associated with most phenotypes, often due to mutations affecting related genes in the same genomic region. Manual inspection of the ten top-scoring phenotypes found multiple FunSys connections supported by previous studies for seven of them. The workflow also produces reports focussed on the genes and FunSys connected to the different phenotypes, alongside patient specific reports, which give details of the associated genes and FunSys for each individual in the cohort. These can be run in “confidential” mode, preserving patient confidentiality. Conclusions: The workflow presented here can be used to associate phenotypes with functional systems by using data at the level of a whole cohort of patients, identifying important connections that could not be found when considering them individually. The full workflow is available for download, enabling it to be run on any patient cohort for which phenotypic and CNV data is available.
Corresponding Author:	Fernando Moreno Jabato University of Malaga Málaga, SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Malaga
Corresponding Author's Secondary Institution:	
First Author:	Fernando Moreno Jabato
First Author Secondary Information:	
Order of Authors:	Fernando Moreno Jabato Pedro Seoane James Richard Perkins

	Elena Rojano
	Adrian Garcia Moreno
	Monica Chayogen
	Florencio Pazos
	Juan Antonio Garcia Ranea
Order of Authors Secondary Information:	
Author Comments:	
Suggested Reviewers:	Antonio Rausell Director of the Clinical Bioinformatics Lab, INSERM antonio.rausell@inserm.fr Expert in rare-disease genomics.
	Julia Foreman DECIPHER Project Manager, Wellcome Sanger Institute jf11@sanger.ac.uk Project Manager of DECIPHER, a key data resource used in this project.
	Nikolas Maniatis Professor of Human Genetics, University College London n.maniatis@ucl.ac.uk Expert in mapping genes to complex disease using massive datasets.
	Belen Perez Group Leader, Universidad Autonoma de Madrid bperez@cbm.csic.es Expert in rare-disease and the analysis of their underlying molecular bases.
	Valentina Cipriani Senior Bioinformatics Research Fellow, Queen Mary University of London v.cipriani@qmul.ac.uk Expert in bioinformatics applied to rare disease; member of HPO disease phenotyping team.

RESEARCH

Systematic identification of genetic systems associated with phenotypes in patients with rare genomic copy number variations

F.M. Jabato[1†], Pedro Seoane[2,1†], James R. Perkins[2,4], Elena Rojano[1], Adrián García Moreno[3], M. Chagoyen[3], Florencio Pazos[3] and Juan A.G. Ranea[1,2,4*]

[*]Correspondence: ranea@uma.es
[1]Department of Molecular Biology and Biochemistry, University of Malaga, Bulevar Louis Pasteur, 31, 29010 Malaga, Spain
[4]Institute of Biomedical Research in Malaga (IBIMA), C. Dr. Miguel Díaz Recio, 28, 29010 Malaga, Spain
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: Copy number variation (CNV) related disorders tend to show complex phenotypic profiles that do not match known diseases. This makes it difficult to ascertain their underlying molecular basis. A potential solution is to compare the affected genomic regions for multiple patients that share a pathological phenotype, looking for commonalities. Here we present a novel approach to associate phenotypes with functional systems, in terms of GO categories and KEGG and Reactome pathways, based on patient data.

Methods: The approach uses genomic and phenomic data from the same patients, finding shared genomic regions between patients with similar phenotypes. These regions are mapped to genes to find associated functional systems.

Results: We applied the approach to analyse patients in the DECIPHER database with *de novo* CNVs, finding functional systems associated with most phenotypes, often due to mutations affecting related genes in the same genomic region. Manual inspection of the ten top-scoring phenotypes found multiple FunSys connections supported by previous studies for seven of them. The workflow also produces reports focussed on the genes and FunSys connected to the different phenotypes, alongside patient specific reports, which give details of the associated genes and FunSys for each individual in the cohort. These can be run in "confidential" mode, preserving patient confidentiality.

Conclusions: The workflow presented here can be used to associate phenotypes with functional systems by using data at the level of a whole cohort of patients, identifying important connections that could not be found when considering them individually. The full workflow is available for download, enabling it to be run on any patient cohort for which phenotypic and CNV data is available.

Keywords: Functional Systems; Human Phenotype Ontology; Gene Ontology; Copy Number Variants; Rare Disease

Background

Copy number variation (CNV) refers to structural genomic mutations that can range from a few thousand to millions of base pairs in length. They include deletions, duplications and translocations of chromosomal regions, which can be inherited or occur spontaneously *(de novo)*. They are frequently observed in healthy individuals, and in terms of total nucleotides, contribute more to inter-individual genomic diversity than SNPs [1].

However, they can also cause disease. This can occur through a variety of mechanisms, including gene dosage effects, insertion or deletion of regulatory regions, and changes in physical proximity of genes and regulatory elements [2, 3]. Pathological symptoms can vary depending on the genetic background of the patient, meaning CNVs in the same region can lead to various phenotypic outcomes. For example, deletions can unmask different recessive alleles depending on the patient's lineage [3].

It can be difficult to discriminate between benign and pathogenic CNVs and to understand how the latter lead to disease [4]. To this end, the DECIPHER project was launched in 2004, creating a global repository of phenotypic information for thousands of patients from around the world with rare and often undiagnosed diseases. Importantly, it also includes high-throughput genomic data, including CNVs [5].

DECIPHER patients tend to show combinations of phenotypes that are not found within known genetic diseases. This precludes the study of the underlying genetic causes by comparison with the mechanisms of already characterized diseases. Their phenotypes are annotated using the controlled vocabulary of human abnormalities from the Human Phenotype Ontology (HPO) database [6]. This HPO is organised hierarchically, ranging from very broad descriptions such as Organ abnormality (HP:0000118), to more specific ones such as 2-3 toe syndactyly (HP:0004691). The use of this vocabulary allows the systematic integration and analysis of phenotypic information for many patients.

In previous work, we found CNVs in DECIPHER patients to average around 700,000 bps in length, compared to 5,000 bps in healthy individuals [7]. Moreover, when considering *de novo* mutations the average size reached 3,000,000 bps. Clearly, larger CNVs are more likely to affect multiple genes and functional genomic regions, increasing the severity and complexity of the phenotypes observed in these patients. The tendency of functionally related genes to sometimes cluster together within the genome can also affect CNV impact, with a recent study finding that CNVs covering multiple functionally related genes led to developmental disorders [8].

These factors, together with the low number of patients and the complex phenotypic profiles associated with them [9], makes the identification of the causal genetic mechanisms and their relationship with the phenotypes observed an unsolved challenge for patients with rare pathogenic CNVs. However, it is one that must be tackled, if we are to improve diagnosis and treatment.

Many previous studies have tried to link genes with HPO phenotypic terms using various techniques [10, 11, 12]. Our group has previously associated clinical phenotypes with genomic regions based on patient data, employing systemic approaches based on network analysis using the DECIPHER database [13, 14, 7].

Furthermore, there has been much interest in connecting patient phenotypes, often in the form of HPO terms, to the functional systems deemed responsible. Many of these are based on the premise that human diseases are not independent of each other but are associated through a complex network of shared genes [15, 16, 17]. Diseases with similar clinical manifestations tend to be caused by similar molecular mechanisms [18]. Moreover, clinical signs form modules in the human interactome [19], indicating that their associated genes tend to be functionally related.

One previous method, PhenoGO, uses natural language processing techniques to obtain gene-GO-phenotype connections from the literature, and this technique has been applied to study disease, although a controlled disease vocabulary was not used [20, 21]. Another method, HPO2GO [22], used the HPO-gene mappings provided by the HPO website based on known gene-disease relations from OMIM/Orphanet, comparing them with GO term to gene mappings and associating HPO and GO terms using a co-occurrence similarity measure. However, to the best of our knowledge, no studies or software exist to link phenotypes with functional systems based on cohort-specific genotype-phenotype data.

In this work, we present a methodology that links phenotypes with functional systems using patient data. The approach is based on the idea that genes associated to a given disease or phenotype tend to concentrate in particular molecular systems. It has been implemented as a workflow, named PhenFun. The approach first associates patient phenotypes with affected genomic regions using a tripartite network model, as described previously [13, 14]. These genomic regions are then mapped to genes and overrepresentation analysis is used to find enriched functional systems (GO terms and KEGG/Reactome pathways). We used this workflow to analyse patients from the DECIPHER database [5], representing a cohort of *de novo* patients with undiagnosed rare and complex diseases and a heterogeneous range of symptoms. Results are compared to those of random models, which are also used as part of a filtering step. We are able to suggest biologically relevant FunSys for most HPO phenotypes in DECIPHER and predict underlying genes and novel genetic-phenotype relationships. We present several of these in the manuscript; the full list is output by the PhenFun workflow as a report, along with patient-centric reports. Here, we present example case studies for two DECIPHER patients; the workflow also has a confidential mode option, which ensures all output reports do not include specific patient data. The source code is also freely-available, enabling the workflow to be applied to any patient cohort for which HPO annotation and CNV data is available.

Methods

Description of patient records

This study was performed using a patient cohort from the DECIPHER database [5] (15th January 2017), which contains records for 20,520 patients with undiagnosed rare genetic disorders. Of these, 13,915 contain both phenotype annotation from the Human Phenotype Ontology (HPO) and genotype information, largely copy number variations (CNVs), in the form of genomic coordinates. Of these, we selected patients whose genomic mutations are tagged as *de novo*, as they can be considered likely to have a direct relationship with the observed phenotypes, resulting in a final set of 3,393 patients, linked to a total of 3,888 CNVs. Of these, 2,529 patients were linked to 2,689 CNV deletions, 1083 were linked to 1,199 CNV duplications and 219 patients showed both deletions and duplications.

Workflow Overview

The workflow used in this study, PhenFun (Figure 1), was developed using the workflow manager AutoFlow [23] and can be downloaded from the repository

https://github.com/fmjabato/PhenFun, which includes details on how to install the framework in any Unix-like system. Conceptually, the workflow can be divided into two main parts. The first (Figure 1, orange box) relates the pathological phenotypes with genomic regions based on their overlap, and finds genes within these genomic regions. These overlapping genomic regions are called short overlapping regions (SOR). They are shared by at least two different patients and any mutation region that is present in only a single patient is discarded as it is not suitable for statistical analysis. The second part (Figure 1, blue box), takes the genes annotated in the phenotype-associated genomic regions and computes the putative Phenotype-associated Functional Systems (FunSys) looking for enrichment of these genes in biological systems defined in Gene Ontology (GO) [24], Reactome [25] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [26]. These putative Phenotype-FunSys relations are then filtered using three random models (Figure 1, far right side).

Identifying Phenotype-Gene relations

The first step is to decompose the mutation data to only include regions in the genome shared by at least two patients, based on the methodology proposed in [13, 14]. These regions are known as short overlapping regions (SORs). Thus any region that is mutated in only a single patient is discarded, as it would be of no use for the subsequent steps of the analysis, which involve connecting phenotypes with mutations based on patient overlap.

Once the SORs have been obtained, they are used to build a tripartite network with three entities: SORs, patients and phenotypes (Figure 1, orange box, left).

Thus, each patient is connected to his or her genomic mutations (represented by the SORs) and phenotypes (represented by the HPO terms), resulting in a tripartite network. The patient-phenotype relations can then be increased using parental HPO term expansion, based on the ontological relationships in the HPO OBO file (2019-02-12 version). In short, if a specific term has ancestor terms in the OBO, the phenotypic profile of each patient is expanded to include all of these ancestor terms. Network models and results obtained using this expanded dataset will be referred to as parental expansion.

Once the tripartite networks have been built (with and without parental expansion), hypergeometric index (HyI) association values between SORs and phenotypes are then calculated, as described in [27], via a projection-layer, using the NetAnalyzer software [13] (Figure 1, orange box, top right). Associations with a HyI below 2, corresponding to an alpha (α) value of 0.01, are then discarded.

Finally, the significant Phenotype-SOR term relations are converted to Phenotype-Gene relations, by taking genes overlapping the SORs, using human genome annotation obtained from the NCBI for genome build GRCh37.p13. In this way, each SOR is replaced by the set of genes found within the region and a new Phenotype-Gene association is created for each gene located in the SOR (Figure 1, orange box, bottom right).

Random models

Three randomised Phenotype-Gene datasets (random models, Figure 1, solid green box) are also constructed for comparison and for a subsequent filtering step de-

scribed below. These were produced using the following methods: 1) Phenotype-SOR randomisation (PhenSOR), in which the significant Phenotype-SOR relations are randomised without modifying the number of SORs connected to each phenotype, and connecting phenotypes to genes according to their new SORs; 2) SOR coordinate randomisation (CorSOR), in which the genomic coordinates of the original SORs are randomised, keeping the size of the genomic regions the same, and connecting the Phenotypes to the genes in the new regions; and 3) Phenotype-Gene randomisation (PhenGene), in which the Phenotype-Gene relations obtained in the first part of the workflow are shuffled, keeping the number of genes connected to each phenotype the same. For each random dataset, 100 different replicates were generated.

Inferring Phenotype-Functional system relations from Phenotype-Gene relations

In the second part of the workflow we use the genes mapped to each phenotype to infer putative Phenotype-FunSys relations (Figure 1, blue box) through over-representation (enrichment) analysis. To achieve this, the genes annotated in SORs significantly associated with the same phenotype are mapped to genes belonging to different processes/pathways in KEGG using ClusterProfiler [28], Reactome using ReactomePA [29] and GO using TopGO [30]. These packages are used with default parameters and Phenotype-FunSys relations with a p-value less than 10^{-3} are considered significant.

These results were then filtered (Figure 1, solid red box), using the Phenotype-FunSys relations found using randomized datasets described above. Any Phenotype-FunSys relation that is found in at least one of the 100 replicates for any of the three random datasets is discarded.

Workflow output

PhenFun generates three types of HTML report to help the user understand and analyse their data: 1) A general report describing the entire dataset and various aspects of the analysis, the number of Phenotype-Gene and Phenotype-FunSys relations found and details of the comparisons between the real patient data and the random datasets. 2) Three phenotype reports (one for each annotation source, GO, KEGG and Reactome) that describe, for each phenotype, the associated FunSys and which diseases are linked in OMIM. 3) Three patient reports (one for each annotation source), that integrate the Phenotype-FunSys connections for each patient, and present the results of this integration alongside other pertinent information, allowing the user to inspect the patients individually. Taken together, these reports allow the user of the workflow to inspect the results and hypothesize the molecular causes of the observed phenotypes.

For the Phenotype-FunSys connections in the patient reports, all HPO parent-child connections are analysed, in order to remove redundant parental terms using the HPO. In addition, for the Phenotype-GO term connections, redundant GO parental terms are also removed based on the ontological structure of GO.

We include examples of these reports as supplementary information, based on the analysis of the *de novo* DECIPHER patients. However, due to patient confidentiality, we have removed all patient-specific data from the phenotype reports. In

addition, in the patient report we only include data for two case studies described below, for whom we have obtained permission from DECIPHER to present here. The workflow has an option to include or omit patient-specific data in the output reports. As such, the interested reader could rerun the workflow on the DECIPHER data under signed agreement and produce the full reports.

The reports also give details on the information content of each Phenotype-FunSys relation, which can be interpreted as its specificity. It is described in the following section.

Information content calculation

In formal ontologies like HPO or GO, the different terms are hierarchically related in a directed acyclic graph (DAG). We calculate the information content (IC) of the phenotypes (HPO terms) and FunSys (GO terms) detected by our system using the following formula:

$$IC(term_i) = -\log(P(term_i)) \tag{1}$$

For HPO terms, the probability of occurrence ($P(term_i)$) is calculated by dividing the number of patients with a given phenotype in the initial dataset by the total number of patients. For GO terms, it is calculated by dividing number of genes annotated with a given GO term by the total number of genes annotated with any GO term in the human genome. We calculated the average IC for each patient by taking the mean IC of all their associated phenotypes.

Integrated FunSys enrichment score calculation

Phenotypes could be associated with more than one FunSys. Therefore, for the purpose of ranking phenotypes, we calculated an integrated enrichment score by combining the enrichment p-values for all FunSys associated with a given phenotype as follows:

$$(\sum_{i=1}^{N} -2log[p_i])/N \tag{2}$$

Where p_i represents the overrepresentation (enrichment) p-value for the ith FunSys and N represents the total number of FunSys for a given phenotype. This integration is based on Fisher's combined probability test.

Results

Comparison of different phenotype-patient-mutation networks

DECIPHER contains deletion and duplication CNVs. It might be expected that these different mutation types lead to distinct pathological outcomes, even when occurring in the same genomic region, due to differences in gene dosage effects or changes to chromosome structure. Another important consideration is that DECIPHER patients have been characterized by different clinical groups, who may follow

different annotation protocols or use the HPO in subtly different ways. As such, it may be that one group uses more general HPOs to describe a patient, whilst another group uses more specific ones, which are in fact child terms of the more general terms used by the former group. If the hierarchical structure of the HPO was ignored, an algorithm would consider phenotypes defined at distinct levels as independent terms rather than related. This could occur with the term *intellectual disability* (HP:0001249) and its child terms, *severe intellectual disability* (HP:0010864) and *mild intellectual disability* (HP:0001256), for example.

In order to evaluate the impact of CNV type and HPO parental-child relations, the PhenFun workflow for identifying Phenotype-FunSys relationships was executed with different settings: only considering deletions, only considering duplications, and considering all *de novo* mutations. Of the 3,393 patients in the initial cohort, 2,529 had deletions and 1,083 had duplications. Furthermore, 219 patients have one duplication and one deletion.

For each of these three divisions of the original cohort, the workflow was executed twice: i) without taking into account parent-child HPO term relations (raw version) and ii) adding ancestors in the HPO hierarchy to each phenotype in the patient profile (parental expansion version).

Table 1 shows the total genomic and phenotypic coverage achieved for each of the workflow-executions, before performing FunSys overrepresentation analysis. Using the full cohort containing all *de novo* mutations and executed using the parental expansion dataset yields the highest coverage, in terms of number of nucleotides in the genome and number of genes significantly associated with an HPO term. The overlap between the different results is shown in Figure 2. It was possible to find associations for 94% of the total number of genes found (26,435 out of 28,034) in the all *de novo* mutations plus parental expansion execution, with a set of 2,243 additional genes found by this execution only.

In terms of phenotypic coverage, again considering the full cohort including all *de novo* CNVs led to the highest coverage, identifying a total of 900 HPO terms associated with at least one gene. In fact, it identifies all HPO terms found in the other versions, plus an additional 30 (Figure 2).

These results suggest that the use of all *de novo* CNVs, i.e. deletions and duplications together, allows us to find almost all the phenotype-gene connections that would be found by treating them separately, plus specific connections for this set. Furthermore, by expanding the annotation to include parental HPO terms, we were able to find additional connections. Therefore, we chose to focus on the results generated using all *de novo* mutations and parental expansion for subsequent analysis.

Using these parameters, 3,347 of the original 3,393 patients could be associated with at least one of a total of 5,520 SORs and at least one of a total of 900 phenotypes, generating 18,533 significant Phenotype-SOR relations. These 5,520 SOR mapped to 26,435 different genes allowing a total of 279,779 Phenotype-Gene connections to be established.

Comparison between the results obtained for real dataset and random models

The 279,779 Phenotype-Gene associations obtained when running the workflow using parental expansion and all *de novo* mutations were used to infer phenotype-

functional system relationships in the second stage of the workflow. The Phenotype-Gene and Phenotype-SOR associations obtained from the original DECIPHER data will be referred to as the real data, to distinguish them from those obtained with the randomized data. Then, we took the real data and randomised them using the three random model generation procedures described above, in order to obtain different random datasets of Phenotype-Gene connections. One hundred replicates were generated for each of the three random models; each of these in turn were used as input to run the same protocol as applied to the real data to identify phenotype-functional system relationships for three databases, GO, KEGG and Reactome.

The random datasets PhenGene (randomised Phenotype-Gene connections) and PhenSOR (randomised Phenotype-SOR connections) contain the same number of genes as the real dataset (Table 2), which is to be expected given that they are based on the same Phenotype-Genotype connections, but randomised. However, the CorSOR dataset, which uses random genomic coordinates for each SOR (keeping the size of each SOR the same) shows a lower total number of genes. This suggests that the SORs in the real patient data tend to overlap with coding regions, in agreement with previous studies [31, 32, 33].

The comparison of results obtained from the real and random models (Table 2) shows that the PhenSOR and CorSOR random models produce similar number of Phenotype-FunSys associations compared with the real data. However, the PhenGene random model yields a much lower number of Phenotype-FunSys pairs compared to the real and the other random models.

The PhenGene model represents the hypothesis that FunSys annotations are completely independent of the gene location in the genome, which we know not to be the case [34]. This model randomizes Phenotype-Gene relationships, breaking the links between adjacent genes. However, when FunSys enrichment is calculated using adjacent genes within a SOR, as occurs with the PhenSOR and CorSOR random models, the number of significant Phenotype-FunSys identified rises to almost reach that of the real data.

Due to the low number of Phenotype-FunSys identified using KEGG and Reactome, we show the results obtained using these two databases in the supplementary files only (Supplementary Files 3 and 4); in the rest of this work we will focus on the GO term analysis.

Interestingly, when stricter p-value thresholds are used for identifying enriched GO terms, the gap between real and random models increases. Moreover, the results obtained using the real data show a relatively larger number of Phenotype-GO term associations. However, the absolute number decreases for all datasets, real and random (see Figure 3, A). The same is not true for the total number of phenotypes associated with at least one GO term (see Figure 3, B): here the ratio remains similar for the real data and random models regardless of the threshold used, with the exception of the PhenGene model.

These results suggest that genes involved in the same FunSys tend to be located more closely in the genome, in keeping with the findings of Thevenin *et al.*, [34], increasing the probability that they will occur within the same SOR. In order to estimate the extent of this effect, we calculated the percentages of SORs containing two or more genes that contribute to GO enrichment in the different datasets. In

Figure 4 we see that for the real data, 32,7% of SORs associated with a Phenotype contain two or more genes linked to the same GO term. This percentage is similar for the PhenSOR model, dropping slightly to 23.2% for the CorSOR random model. Interestingly, this tendency mirrors the number of Phenotype-GO terms associations found in the different random models in Figure 3 above, which shows that the CorSOR model finds a slightly lower number of Phenotype-GO pairs at all enrichment p-value thresholds than the real and PhenSOR models.

Identifying significant and specific Phenotype-FunSys relationships

As genes within the same FunSys tend to be located in the same SORs, the Phenotype-FunSys results from the PhenSOR, CorSOR and PhenGene random models were used to filter the Phenotype-GO term results obtained from the real data. After this filtering, 5,826 of the initial 23,967 Phenotype-GO pairs remained (Table 3).

For HPO and GO terms it is possible to calculate their Information Content (IC) as described in section Information Content Calculation in Methods. The higher the IC of a given HPO or GO term the more specific it is; conversely, lower IC values are found for more general and prevalent HPO or GO terms. IC values were calculated for each HPO and GO term in all Phenotype-GO term associations obtained before and after the random filtering step (see Figure 5).

The random filtering process tends to remove less informative, i.e. more general phenotypes and GO terms, and hence their connections, leaving only the more informative and specific Phenotype-GO term pairs. The density distribution bias towards more informative Phenotype-GO term pairs following filtering indicates that noise due to the genes involved in the same FunSys and located in the same SOR regions could generate less-specific (low IC) Phenotype-GO term associations.

In some cases, the set of Phenotype-FunSys pairs contains redundancy due to multiple ancestor terms of a given HPO also being connected to the same FunSys. In order to avoid redundancy we selected the pair involving the most specific terms with the following procedure: For a given Phenotype-FunSys term pair in the random filtered set, we identified all other pairs connected to the FunSys that involved ancestors/descendants of the phenotype. Of these, we selected the pair containing the most specific HPO term in terms of the ontology structure. These values are shown in Table 3. The final result was a dataset of 3,247 significant and highly specific Phenotype-GO term pairs, comprised of 609 distinct Phenotypes and 1,931 GO terms.

Genes underlying the Phenotype-FunSys connections in the patient cohort

Each Phenotype-GO relation is derived from patients with mutations affecting a pool of genes involved in a specific functional system. We therefore inspected the patients and compared their affected genes with those involved in each Phenotype-GO association, allowing us to asses which genes, out of all mutated genes, are likely to contribute to the specific phenotypes of the patient profile and thus infer which molecular mechanisms have been altered, leading to the manifestation of the clinical phenotype.

We explored the relationship between the total number of affected genes per patient and the percentage of these genes that were involved in a Phenotype-GO relation.

In general, for patients that have mutations affecting a large number of genes, a smaller percentage of these genes actually contribute to Phenotype-GO associations, whilst patients with mutations affecting few genes tend to show a much larger range of percentages (Figure 6A).

This might suggest that for patients with mutations affecting a large number of genes, only a small percentage may actually be affecting the underlying molecular mechanisms that are responsible for the phenotypic profile.

However, it is important to take into account the lack of information provided by a large number of patients. As shown in Figure 6B, over half of the patients have no GO terms that are found in the final Phenotypes-GO associations. Moreover, these patients tend to have phenotypic profiles consisting of HPO terms with low IC. Conversely, the patients with high IC terms tend to have far more associated FunSys. This fact underlies the importance of deeper patient phenotyping for the low IC cases. Furthermore, this explains the poor performance of PhenFun when it attempts to describe which functional systems are affected in a specific patient.

Inspecting phenotype-GO term associations

PhenFun produces multiple reports in order to allow the user to examine the information generated by the workflow. Here, we comment on the phenotype report generated for the DECIPHER data, describing various Phenotypes and their related GO terms. This information is contained in Supplementary File 2, which contains all relevant data generated by PhenFun for the analysis of each phenotype, including information on the parental terms in the ontology, OMIM diseases that have been mapped to the phenotype [35] and associated GO terms. Supplementary Files 3 and 4 contain analogous information for KEGG and Reactome, respectively. The workflow is also able to add details of the patients with a given phenotype that also show mutations in genes annotated with the associated GO terms. However, this information is not included in the report for the DECIPHER data-set due to patient confidentiality.

Given the results of the IC analysis shown in Figure 6B, it would appear that the phenotypes with higher IC are held by patients that contribute to a greater number of Phenotype-GO connections. As such, the phenotypes in the above-mentioned reports are ordered according to their IC. If multiple phenotypes have the same IC, they are subsequently ordered according to their integrated enrichment score. Details of the integration are described in methods.

The top ten results found using this ordering are shown in Table 4. Manual inspection of these phenotypes reveals multiple Phenotype-GO relations of interest.

The top result, *Arthritis* (HP:0001369) is related to multiple GO terms that have been shown to play a role in its treatment. For example, *coumarin metabolic process* (GO:0009804) was shown to be associated; remarkably, derivatives of coumarin have wide anti-inflammatory properties and have been used to treat arthritic conditions [36]. Similarly, several flavonoid metabolism related GO terms are also associated with this phenotype and share several genes with coumarin metabolism;

these compounds have been shown to have anti-inflammatory properties and have been reported to alleviate arthritic symptoms [37]. The term *bilirubin conjugation* (GO:0006789) was also associated; interestingly, low serum levels of bilirubin are associated with rheumatoid arthritis [38]. Finally, this phenotype is also associated with *retinoic acid binding* (GO:0001972); recent research has shown retinoic acid-binding protein 2 to be a potential target for rheumatoid arthritis synovial hyperplasia [39].

The second result, *Generalized tonic seizures* (HP:0010818), shows associations with multiple GO terms related to ion-channel activity and neuron action potential. These associations might be expected, given that the neurobiology underlying seizures can be related to a state of hyper-excitability [40]. Looking in more detail at the genes involved in these associations (Supplementary File 2), we see that all of the genes affected are voltage gated sodium channel subunits, many of which have been shown to be associated with epilepsy and seizures [41].

The third phenotype in the table, *Urinary incontinence* (HP:0000020), is associated with multiple GO terms related to antibodies and their receptors. Although links have been investigated between IgE levels and diseases related to urinary incontinence, such as interstitial cystitis and bladder pain syndrome, results remain somewhat inconclusive [38].

This is in contrast with the fourth phenotype in the table, *Premature birth* (HP:0001622), which shows associations with similar immune system related categories, including immunoglobulin production and their receptors, but through different genes. Intrauterine inflammation is associated with premature birth, and as such this result is unsurprising, although it is interesting that the pathways involved are related to immunoglobulins and their receptors, rather than T-cell and HLA related processes [42].

The sixth phenotype, *Intellectual disability, profound* (HP:0002187) is associated with the same six GO terms as Generalized tonic seizures, with the addition of *sensory perception of pain* (GO:0019233). Further inspection shows that the same genes are involved, with the exception of *SCN7A*, demonstrating that these phenotypes may be caused by pleiotropic affects of these genes, potentially leading to seizures and/or intellectual disability by affecting sodium channel activity and neuronal action potential. Certainly, sodium channels have been associated with intellectual disability, such as those encoded by *SCN2A* and *SCN3A* [43, 44]. Somewhat more confusing is the association between the seventh phenotype in the table, *Supernumerary nipple* (HP:0002558), and these same GO terms. Further inspections shows that this phenotype always co-occurs with *Intellectual disability, profound* in DECIPHER, as such it also appears enriched for similar terms.

The eighth most highly ranked phenotype, *Central hypotonia* (HP:0011398) shows multiple associations with GABA signalling, principally GABA-A related pathways. Multiple disorders related to GABA metabolism involve hypotonia [43]. Moreover, there is a rare disease, *Gamma-aminobutyric acid transaminase deficiency* (ORPHA:2066), which includes hypotonia as one of its constituent phenotypes.

Phenotype nine in the table, *Bicuspid aortic valve* (HP:0001647) is only associated with two GO terms, *serotonin receptor signaling pathway* (GO:0007210) and *serotonin-gated cation-selective channel activity* (GO:0022850). Excessive serotonin

is known to lead to tachycardia [45], and several studies have associated serotonin related genes and processes with heart disease, including heart valve problems [46]; as such the association with the bicuspid aortic valve might give clues as to exactly how serotonin exerts its influence on the cardiovascular system. Digging down further into the three genes associated with this phenotype, they code for sub-units of the 5-Hydroxytryptamine (serotonin) receptor, which itself has been shown to exert multiple effects on the cardiovascular system [47].

The two other phenotypes in Table 4, *Depressed nasal tip* (HP:0000437) and *Stridor* (HP:0010307) showed associations with GO types for which no clear connection could be found. It is possible that these phenotypes tend to co-occur with other phenotypes that are related to these GO terms, or represent novel associations that are yet to be demonstrated experimentally.

Patient case studies

PhenFun also produces patient reports. These show, for each patient, which of their phenotypes were associated with a FunSys through at least one gene that is mutated in that patient's CNVs. It then presents a series of figures showing the relationship between these phenotypes, FunSys and genes. This is limited to patients that have a minimum coverage value of 50%, in accordance with the results shown in Figure 6B, which suggests that patients with lower coverage do not add much information to the generation of Phenotype-GO term associations. By coverage, we refer to the percentage of phenotypes of a patient that have associated GO terms, where the patient has one or more mutated genes belonging to at least one of those GO terms.

We obtained permission from DECIPHER to analyse two patients with maximal coverage according to their clinically observed phenotypes. Full details are given in Supplementary File 5. Although the workflow also generates patient reports for KEGG and Reactome, we have only included the GO report, as this provided the most relevant information for the two case studies.

Case study: Patient 1

For the first case study we selected a patient with a complex phenotypic profile comprising three phenotypes: *Cerebellar hypoplasia* (HP:0001321), *Intellectual disability, severe* (HP:0010864) and *Congenital microcephaly* (HP:0011451).

This patient has 21 mutated genes, caused by deletion, however only two are involved in Phenotype-GO relations: monoamine oxidase A and B (*MAOA* and *MAOB*, Figure 7A). In addition, Figure 7C shows that both genes are involved in all the Phenotype-GO terms connections for this patient; these GO terms are related to neurotransmitter-related metabolic activity: *Primary amine oxidase activity* (GO:0008131), *neurotransmitter catabolic process* (GO:0042135) and *dopamine catabolic process* (GO:0042420).

As such, we hypothesize that the symptoms presented by this patient are due to problems related to these processes, due to mutations in these genes. In fact, a search of the literature reveals that the genes are involved in Brunner syndrome, which involves severe mental disability and a number of other neurological diseases including Parkinsonian dopamine deficiency disorders [48]. In Figure 7E we see that the genes used to obtain an association between the GO terms and phenotypes during the workflow execution include all the genes affected in this patient.

Case study: Patient 2

For the second case study we have selected a patient with a phenotype profile comprised of two phenotypes: *Lissencephaly* (HP:0001339) and *Cerebellar hemisphere hypoplasia* (HP:0100307).

This patient has 38 affected genes, caused by deletion, of which two are involved in significant Phenotype-GO relations, as shown in Supplementary File 5. These genes are Delta-like protein 1 precursor (*DLL1*) and TATA-Box Binding Protein (*TBP*). Figure 7B shows than *TBP* is only related to one GO term, *Obsolete general RNA polymerase II transcription factor activity* (GO:0003702). However *DLL1* is associated with 6 GO terms: *Notch signaling pathway involved in arterial endothelial cell fate commitment* (GO:0060853), *Endothelial tip cell fate specification* (GO:0097102), *Columnar/cuboidal epithelial cell development* (GO:0002066), *Cerebellar Purkinje cell layer structural organization* (GO:0021693), *Cerebellar molecular layer formation* (GO:0021688), *Auditory receptor cell differentiation* (GO:0042491). In Figure 7D we observe that both genes are related to both phenotypes.

When the phenotype-GO connections are inspected, we see that both phenotypes are connected to the GO terms *general RNA polymerase transcription factor activity*, *Notch signalling involved in arterial endothelial cell fate* and *cerebellar molecular layer formation* as shown in Figure 7F. All the genes that support these Phenotype-GO relations are affected by the patient mutation. There are also specific Phenotype-GO relations for the individual phenotypes: *Lissencephaly* has weaker relations with *columnar cuboidal epithelial cell development* and *auditory receptor cell differentiation*, with the patient's mutation only affecting half of all genes involved in these Phenotype-GO associations within the entire network. For the phenotype *Cerebellar hemisphere hypoplasia*, there are two strong specific connections with the GO terms *endothelial tip cell fate specification* and *cerebellar Purkinje cell layer*.

The associations with GO terms related to tissue development and cell differentiation, in particular for cerebellum-related processes, support the patient's phenotypes. This is reinforced by previous studies that suggest that Purkinje cells and Notch alterations are connected to Autism and Intellectual disability phenotypes [49, 50], which could be missed in the phenotype patient characterization. The results suggest that the main way by which DLL1 leads to the patient's condition is by affecting the the molecular systems related to the GO terms. In fact, recent studies have related this gene to neurodevelopmental alterations, although the underlying mechanisms are unknown[51].

Discussion

We have presented PhenFun, a new method for linking human disease-related phenotypes with functional systems (FunSys), such as Biological Process and Molecular Function GO terms, and pathway data from KEGG and Reactome, to aid in the elucidation of molecular mechanisms underlying CNV-related pathologies.

Given a cohort of patients with detailed CNV-associated pathologies and phenotypic and genomic information, the workflow identifies significant phenotype-genotype connections by exploiting all available association data in the phenotype-patient-genotype network. Then, by transferring the mutations to the gene level

and performing overrepresentation analysis, it is ultimately able to associate the phenotypes with FunSys, which are then filtered using a battery of random models in order to remove associations that are likely to occur by chance.

Here, we have analysed thousands of highly genetically and phenotypically heterogeneous *de novo* patients, registered in the DECIPHER database. These patients were used to build various phenotype-patient-short overlapping region (SOR) tripartite networks. The SORs were built by decomposing the CNVs of the DECIPHER patients to find regions that were shared by at least two individuals.

As such, it was necessary to decide whether to separate CNVs into deletions and duplications, or to integrate them, i.e. should deletions and duplications in the same region be considered different mutations. By considering them separately, one might expect that we have increased power to detect certain Phenotypes-SOR associations, i.e. those that only occur due to a deletion in a certain region, but not a duplication, and vice-versa.

This loss of power could occur in the following manner: a region that is connected to a given phenotype by deletions in five patients, but connected to a different phenotype via duplications in five other patients would have a lower association value when deletions and duplications were considered together. Conversely, for phenotypes that can occur due to either type of mutation in the same region, the Phenotype-SOR association signal would be strengthened by considering duplications and deletions together. Many examples of both cases have been found [52, 53].

Another key decision in the analysis was whether to use HPO parental expansion. The HPO is designed in such a way that a patient with a specific term is also said to suffer from its parental terms. Therefore, when constructing the tripartite network, a given patient should be connected to his diagnosed phenotypes, and optionally to all of the parental terms of these phenotypes in the HPO.

Parental expansion has the potential effect of increasing the number of SOR connections for parental HPO terms, which could help detect additional Phenotype-SOR associations at higher levels of the ontology. Table 1 and Figure 2 show the effects of using deletions, duplications or integrated; with and without parental expansion. Strikingly, the best results were obtained when combining deletions and duplications with the parentally expanded network, yielding the greatest phenotypic, patient and Phenotype-SOR connection coverage. Based on this performance comparison, we used the integrated and expanded network for all further analysis.

The 279,779 significant Phenotype-Gene associations identified by this approach, using the integrated and parentally expanded network, were used for the enrichment analysis step. We focussed the analysis on GO, as the results found using KEGG and Reactome annotations were more limited (Table 2).

The functional overrepresentation analysis protocols implemented employ the Fisher's exact test, based on the hypergeometric distribution [52]. This test assumes independent links between genes, SORs and FunSys, however this is not necessarily the case. We calculated enrichment p-values between Phenotypes and FunSys for the real data and compared the results to those obtained using three different random models.

The PhenGene random model allows a free distribution of Phenotype-Gene pairs by breaking the links between genes and their SOR locations. Curiously, this model

found very few significant Phenotype-FunSys associations at all assayed p-value thresholds, much fewer than the results for the real dataset (Figure 3).

However, when we calculated enrichment using the the two other random models, which maintain the links between genes and their genomic location, as is the cases of the PhenSOR and the CorSOR random models, the number of Phenotype-GO term associations drastically increases for all p-value thresholds, becoming much closer to the real dataset. The reason for this increase is likely due to the tendency for genes in the same FunSys to occur more closely in the genome. Thus the assumptions of the Fisher's exact test are not met, as it assumes independence between genes.

The high percentage of functionally linked genes observed in SOR genomic locations (Figure 4) supports the idea that genes involved in the same FunSys tend to locate more closely in the human genome. This has been described previously by Thevenin et al. [34]. This functional grouping of genes in chromosomes has important misleading statistical consequences, as it can lead to increased false positive Phenotype-FunSys associations being detected using the Fisher's exact test for the real data.

Thus, in order to reduce the number of false positives, the three random models were used to filter out the Phenotype-FunSys pairs, removing any association appearing within the whole dataset made up of 300 random replicas (100 per model). The filtering step leads to a drastic reduction in Phenotype-FunSys pairs, removing 76% of them (Table 3), most of which show low information content (Figure 5). These low informative pairs are composed of highly prevalent phenotypes in the patient cohort, moreover they are connected to FunSys involving many genes in the human genome, as described in the IC calculation in Methods. On the other hand, this filtering process kept the majority of the highly informative pairs, comprising more specific phenotypes and FunSys.

The genes that support Phenotype-FunSys connections can be thought of as the most likely to underlie the pathogenic phenotype. We plotted the percentage distribution of such genes versus the total number of genes located in the SORs corresponding to the CNVs of each patient (Figure 6A). We see that not all the genes mutated in a given patient are found in significant Phenotype-FunSys connections. In fact, the percentage of genes associated with a phenotype drops sharply when the total number of mutated genes increases (Figure 6A).

This suggests that if all genes located in a CNV region were to underlie pathogenic effects it would be impossible for patients with large CNVs to survive. This uneven pathogenic effect could also be explained by haplo-insufficiency, variation in the recessive gene pools inherited in the lineage of each patient, or a combination of both.

In any case, for CNVs the identification of the potentially pathogenic genes among all affected genes in the mutated regions is an important clinical challenge. It is key to ensuring the correct genetic diagnosis of these patients. The approach implemented in this work is able to distinguish those genes relevant to Phenotype-FunSys associations from all genes located in a given CNV mutation. As would be expected, we have also observed that the capability of our approach to identify significant associations depends on the quality of the phenotypic annotation of the patients within the cohort. When patients are annotated with very general HPO terms with low

information content, the workflow performance is poor, improving when patients are annotated with more specific and complete phenotypic records (Figure 6B). Therefore, the workflow performance would greatly benefit from larger cohort sizes, made of deeply phenotyped patients.

We have also shown the potential use of the system to aid clinicians in genetic diagnosis and helping researchers to understand the genetic mechanisms underlying disease symptoms. The top ten phenotypes, according to information content and association strength, show multiple connections to FunSys and genes that have been reported in previous studies (Table 4).

The system allows us to prioritise genes linked to Phenotypes, based on their involvement in a Phenotype-Funsys enrichment, compared to all of the genes located in the SOR pathogenic regions associated with that phenotype, thereby reducing the search space when trying to better understand the underlying mechanisms (see two last columns in Table 4).

The system can also help us to better understand the causes underlying the phenotypic profile of a given patient, by showing the relationships between FunSys, genes and affected phenotypes (Figure 7).

Conclusions

To conclude, we have shown that PhenFun is capable of linking phenotypes with genes and functional systems, in particular for processes and functional terms in GO. As we have explained, this is especially important and useful in the case of undiagnosed and rare CNV genomic disorders, where uncertainty of the pathogenic causing genes can be very high. To the best of our knowledge, PhenFun represents the only approach available to perform such a task using patient data.

Availability of data and materials

The data that support the findings of this study are available as additional material. However, whilst information for the patients results have been removed for confidential reasons. The datasets used and/or analysed during the current study are available from the DECIPHER database under signed agreement. All code underlying the workflow is freely available from https://github.com/fmjabato/PhenFun, written in R and bash script, employing a workflow manager, AutoFlow, to be run on UNIX-like systems. All dependencies are explained in the README file of the Github repository.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by The Spanish Ministry of Economy and Competitiveness with European Regional Development Fund [SAF2016-78041-C2-1-R to J.A.G and SAF2016-78041-C2-2-R to F.P.]; the Andalusian Government with European Regional Development Fund [CTS-486]; biomedicine research project [PI-0075-2017] (Fundacion Progreso y Salud); and the Ramón Areces foundation for rare disease investigation (National call for research on life and material sciences, XIX edition). The CIBERER is an initiative from the Carlos III Health Institute (Instituto de Salud Carlos III).

Acknowledgements

The authors thank the Supercomputing and Bioinnovation Center (SCBI) of the University of Malaga for their provision of computational resources and technical support (www.scbi.uma.es/site).
This study makes use of data generated by the DECIPHER community. A full list of centres who contributed to the generation of the data is available from http://decipher. sanger.ac.uk and via email from decipher@sanger.ac.uk. Funding for the project was provided by the Wellcome Trust. Those who carried out the original analysis and collection of the data bear no responsibility for the further analysis or interpretation of it by the Recipient or its Registered Users.

Author details

[1]Department of Molecular Biology and Biochemistry, University of Malaga, Bulevar Louis Pasteur, 31, 29010 Malaga, Spain. [2]CIBER of Rare Diseases (ISCIII), Av. Monforte de Lemos, 3-5, Pabellon 11, Planta 0, 28029 Madrid, Spain. [3]Computational Systems Biology Group, Systems Biology Program, Spanish National Centre for Biotechnology (CNB-CSIC), 28049 Madrid, Spain. [4]Institute of Biomedical Research in Malaga (IBIMA), C. Dr. Miguel Díaz Recio, 28, 29010 Malaga, Spain.

References

1. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., González, J.R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W., Hurles, M.E.: Global variation in copy number in the human genome. Nature **444**(7118), 444–454 (2006). doi:10.1038/nature05329
2. Gamazon, E.R., Stranger, B.E.: The impact of human copy number variation on gene expression. Briefings in functional genomics **14**(5), 352–357 (2015). doi:10.1093/bfgp/elv017
3. Shaikh, T.H.: Copy Number Variation Disorders. Current genetic medicine reports **5**(4), 183–190 (2017). doi:10.1007/s40142-017-0129-2
4. Zarrei, M., MacDonald, J.R., Merico, D., Scherer, S.W.: A copy number variation map of the human genome (2015). doi:10.1038/nrg3871
5. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Vooren, S.V., Moreau, Y., Pettett, R.M., Carter, N.P.: DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. The American Journal of Human Genetics **84**(4), 524–533 (2009). doi:10.1016/j.ajhg.2009.03.010
6. Robinson, P.N., Mundlos, S.: The Human Phenotype Ontology. Clinical Genetics **77**(6), 525–534 (2010). doi:10.1111/j.1399-0004.2010.01436.x
7. Reyes-Palomares, A., Bueno, A., Rodríguez-López, R., Medina, M.Á., Sánchez-Jiménez, F., Corpas, M., Ranea, J.A.G.: Systematic identification of phenotypically enriched loci using a patient network of genomic disorders. BMC Genomics **17**(1), 232 (2016). doi:10.1186/s12864-016-2569-6
8. Andrews, T., Honti, F., Pfundt, R., De Leeuw, N., Hehir-Kwa, J., Silfhout, A.V.V., De Vries, B., Webber, C.: The clustering of functionally related genes contributes to CNV-mediated disease. Genome Research (2015). doi:10.1101/gr.184325.114
9. Shaw-Smith, C., Redon, R., Rickman, L., Rio, M., Willatt, L., Fiegler, H., Firth, H., Sanlaville, D., Winter, R., Colleaux, L., Bobrow, M., Carter, N.P.: Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. Journal of medical genetics **41**(4), 241–248 (2004). doi:10.1136/jmg.2003.017731
10. Notaro, M., Schubach, M., Robinson, P.N., Valentini, G.: Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods. BMC Bioinformatics **18**(1), 449 (2017). doi:10.1186/s12859-017-1854-y
11. Kahanda, I., Funk, C., Verspoor, K., Ben-Hur, A.: PHENOstruct: Prediction of human phenotype ontology terms using heterogeneous data sources. F1000Research **4**, 259 (2015). doi:10.12688/f1000research.6670.1
12. Javed, A., Agrawal, S., Ng, P.C.: Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nature Methods **11**(9), 935–937 (2014). doi:10.1038/nmeth.3046
13. Rojano, E., Seoane, P., Bueno-Amoros, A., Perkins, J.R., Garcia-Ranea, J.A.: Revealing the relationship between human genome regions and pathological phenotypes through network analysis. In: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2017). doi:10.1007/978-3-319-56148-6_17
14. Bueno, A., Rodríguez-López, R., Reyes-Palomares, A., Rojano, E., Corpas, M., Nevado, J., Lapunzina, P., Sánchez-Jiménez, F., Ranea, J.A.G.: Phenotype-loci associations in networks of patients with rare disorders: application to assist in the diagnosis of novel clinical cases. European Journal of Human Genetics **26**(10), 1451–1461 (2018). doi:10.1038/s41431-018-0139-x
15. Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabási, A.-L.: The human disease network. Proceedings of the National Academy of Sciences of the United States of America **104**(21), 8685–8690 (2007). doi:10.1073/pnas.0701361104
16. Lee, D.-S., Park, J., Kay, K.A., Christakis, N.A., Oltvai, Z.N., Barabasi, A.-L.: The implications of human metabolic network topology for disease comorbidity. Proceedings of the National Academy of Sciences (2008). doi:10.1073/pnas.0802208105
17. Pache, R.A., Zanzoni, A., Naval, J., Mas, J.M., Aloy, P.: Towards a molecular characterisation of pathological pathways. FEBS letters **582**(8), 1259–1265 (2008). doi:10.1016/j.febslet.2008.02.014
18. Zhou, X., Menche, J., Barabási, A.L., Sharma, A.: Human symptoms-disease network. Nature Communications (2014). doi:10.1038/ncomms5212
19. Chagoyen, M., Pazos, F.: Characterization of clinical signs in the human interactome. Bioinformatics (2016). doi:10.1093/bioinformatics/btw054
20. Sam, L.T., Mendonça, E.A., Li, J., Blake, J., Friedman, C., Lussier, Y.A.: PhenoGO: an integrated resource for the multiscale mining of clinical and biological data. BMC Bioinformatics **10**(Suppl 2), 8 (2009). doi:10.1186/1471-2105-10-S2-S8
21. Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y., Friedman, C.: PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 64–75 (2006)
22. Doğan, T.: HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences. PeerJ **6**, 5298 (2018). doi:10.7717/peerj.5298
23. Seoane, P., Ocaña, S., Carmona, R., Bautista, R., Madrid, E., M. Torres, A., Gonzalo Claros, M.: AutoFlow, a Versatile Workflow Engine Illustrated by Assembling an Optimised de novo Transcriptome for a Non-Model Species, such as Faba Bean (Vicia faba). Current Bioinformatics **11**(4), 440–450 (2016). doi:10.2174/1574893611666160212235117
24. The Gene Ontology Consortium: The Gene Ontology in 2010: extensions and refinements. Nucleic acids

research (2010). doi:10.1093/nar/gkp1018
25. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P.: The Reactome Pathway Knowledgebase. Nucleic Acids Research **46**(D1), 649–655 (2018). doi:10.1093/nar/gkx1132
26. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M.: KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research (2016). doi:10.1093/nar/gkv1070
27. Bass, J.I.F., Diallo, A., Nelson, J., Soto, J.M., Myers, C.L., Walhout, A.J.M.: Using networks to measure similarity between genes: association index selection. Nature Methods **10**(12), 1169–1176 (2013). doi:10.1038/nmeth.2728
28. Yu, G., Wang, L.-G., Han, Y., He, Q.-Y.: clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A Journal of Integrative Biology (2012). doi:10.1089/omi.2011.0118
29. Yu, G., He, Q.-Y.: ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Molecular bioSystems **12**(2), 477–479 (2016). doi:10.1039/c5mb00663e
30. Alexa, A., Rahnenfuhrer, J.: topGO: Enrichment Analysis for Gene Ontology. R package version 2.26.0. October (2016)
31. Johansson, A.C.V., Feuk, L.: Characterization of copy number-stable regions in the human genome. Human Mutation (2011). doi:10.1002/humu.21524
32. Nguyen, D.-Q., Webber, C., Ponting, C.P.: Bias of Selection on Human Copy-Number Variants. PLoS Genetics **2**(2), 20 (2006). doi:10.1371/journal.pgen.0020020
33. Nguyen, D.-Q., Webber, C., Hehir-Kwa, J., Pfundt, R., Veltman, J., Ponting, C.P.: Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome research **18**(11), 1711–1723 (2008). doi:10.1101/gr.077289.108
34. Thévenin, A., Ein-Dor, L., Ozery-Flato, M., Shamir, R.: Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. Nucleic Acids Research **42**(15), 9854–9861 (2014). doi:10.1093/nar/gku667
35. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., Hamosh, A.: OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research **43**(D1), 789–798 (2015). doi:10.1093/nar/gku1205
36. Hemshekhar, M., Sunitha, K., Thushara, R.M., Sebastin Santhosh, M., Shanmuga Sundaram, M., Kemparaju, K., Girish, K.S.: Antiarthritic and antiinflammatory propensity of 4-methylesculetin, a coumarin derivative. Biochimie (2013). doi:10.1016/j.biochi.2013.02.014
37. Hughes, S.D., Ketheesan, N., Haleagrahara, N.: The therapeutic potential of plant flavonoids on rheumatoid arthritis. Critical Reviews in Food Science and Nutrition (2017). doi:10.1080/10408398.2016.1246413
38. Juping, D., Yuan, Y., Shiyong, C., Jun, L., Xiuxiu, Z., Haijian, Y., Jianfeng, S., Bo, S.: Serum bilirubin and the risk of rheumatoid arthritis. Journal of Clinical Laboratory Analysis (2017). doi:10.1002/jcla.22118
39. Mosquera, N., Rodriguez-Trillo, A., Mera-Varela, A., Gonzalez, A., Conde, C.: Uncovering Cellular retinoic acid-binding protein 2 as a potential target for rheumatoid arthritis synovial hyperplasia. Scientific Reports (2018). doi:10.1038/s41598-018-26027-x
40. Scharfman, H.E.: The neurobiology of epilepsy (2007). doi:10.1007/s11910-007-0053-z
41. Kaplan, D.I., Isom, L.L., Petrou, S.: Role of Sodium Channels in Epilepsy. Cold Spring Harbor Perspectives in Medicine **6**(6), 022814 (2016). doi:10.1101/cshperspect.a022814
42. Melville, J.M., Moss, T.J.M.: The immune consequences of preterm birth (2013). doi:10.3389/fnins.2013.00079
43. Begemann, A., Acuña, M.A., Zweier, M., Vincent, M., Steindl, K., Bachmann-Gagescu, R., Hackenberg, A., Abela, L., Plecko, B., Kroell-Seger, J., Baumer, A., Yamakawa, K., Inoue, Y., Asadollahi, R., Sticht, H., Zeilhofer, H.U., Rauch, A.: Further corroboration of distinct functional features in SCN2A variants causing intellectual disability or epileptic phenotypes. Molecular medicine (Cambridge, Mass.) (2019). doi:10.1186/s10020-019-0073-6
44. Zaman, T., Helbig, I., Božović, I.B., DeBrosse, S.D., Bergqvist, A.C., Wallis, K., Medne, L., Maver, A., Peterlin, B., Helbig, K.L., Zhang, X., Goldberg, E.M.: Mutations in SCN3A cause early infantile epileptic encephalopathy. Annals of Neurology (2018). doi:10.1002/ana.25188
45. Foong, A.L., Grindrod, K.A., Patel, T., Kellar, J.: Demystifying serotonin syndrome (or serotonin toxicity) (2018)
46. Hutcheson, J.D., Setola, V., Roth, B.L., Merryman, W.D.: Serotonin receptors and heart valve disease-It was meant 2B (2011). doi:10.1016/j.pharmthera.2011.03.008
47. Doggrell, S.A.: The role of 5-HT on the cardiovascular and renal systems and the clinical potential of 5-HT modulation (2003). doi:10.1517/13543784.12.5.805
48. Tong, J., Rathitharan, G., Meyer, J.H., Furukawa, Y., Ang, L.-C., Boileau, I., Guttman, M., Hornykiewicz, O., Kish, S.J.: Brain monoamine oxidase B and A in human parkinsonian dopamine deficiency disorders. Brain **140**(9), 2460–2474 (2017). doi:10.1093/brain/awx172
49. Clifford, H., Dulneva, A., Ponting, C.P., Haerty, W., Becker, E.B.E.: A gene expression signature in developing Purkinje cells predicts autism and intellectual disability co-morbidity status. Scientific Reports **9**(1) (2019). doi:10.1038/s41598-018-37284-1
50. Ding, X.F., Gao, X., Ding, X.C., Fan, M., Chen, J.: Postnatal dysregulation of Notch signal disrupts dendrite development of adult-born neurons in the hippocampus and contributes to memory impairment. Scientific Reports **6** (2016). doi:10.1038/srep25780
51. Fischer-Zirnsak, B., Segebrecht, L., Schubach, M., Charles, P., Alderman, E., Brown, K., Cadieux-Dion, M., Cartwright, T., Chen, Y., Costin, C., Fehr, S., Fitzgerald, K.M., Fleming, E., Foss, K., Ha, T., Hildebrand, G., Horn, D., Liu, S., Marco, E.J., McDonald, M., McWalter, K., Race, S., Rush, E.T., Si, Y., Saunders, C., Slavotinek, A., Stockler-Ipsiroglu, S., Telegrafi, A., Thiffault, I., Torti, E., Tsai, A.C.-h., Wang, X., Zafar, M., Keren, B., Kornak, U., Boerkoel, C.F., Mirzaa, G., Ehmke, N.: Haploinsufficiency of the Notch Ligand DLL1

Causes Variable Neurodevelopmental Disorders. The American Journal of Human Genetics (2019). doi:10.1016/j.ajhg.2019.07.002

52. Mullegama, S.V., Rosenfeld, J.A., Orellana, C., Van Bon, B.W.M., Halbach, S., Repnikova, E.A., Brick, L., Li, C., Dupuis, L., Rosello, M., Aradhya, S., Stavropoulos, D.J., Manickam, K., Mitchell, E., Hodge, J.C., Talkowski, M.E., Gusella, J.F., Keller, K., Zonana, J., Schwartz, S., Pyatt, R.E., Waggoner, D.J., Shaffer, L.G., Lin, A.E., De Vries, B.B.A., Mendoza-Londono, R., Elsea, S.H.: Reciprocal deletion and duplication at 2q23.1 indicates a role for MBD5 in autism spectrum disorder. European Journal of Human Genetics (2014). doi:10.1038/ejhg.2013.67
53. Nevado, J., Mergener, R., Palomares-Bralo, M., Souza, K.R., Vallespín, E., Mena, R., Martínez-Glez, V., Mori, M.Á., Santos, F., García-Miñaur, S., García-Santiago, F., Mansilla, E., Fernández, L., de Torres, M.L., Riegel, M., Lapunzina, Pablo: New microdeletion and microduplication syndromes: A comprehensive review (2014). doi:10.1590/S1415-47572014000200007

Figures

Figure 1: PhenFun workflow. In the first stage (top), HPO phenotype terms (green hexagons) are associated with specific genomic regions (short overlapping regions, SORs), based on overlap between patients. These SORs are then mapped to genes, to obtain Phenotype-gene relations. These are used as input for the second stage (bottom), where each group of genes linked to a given Phenotype are used for enrichment analysis, to obtain putative Phenotype-functional system connections. These connections are then filtered using three random models.

Figure 2: Intersection between the output of the different workflow executions in terms of genes and phenotype, taking into account all *de novo* CNVs (All), all duplications and all deletions. Left side: Overlap in terms of genes associated with at least one phenotype in the raw version (top) and parental expansion version (bottom). Right side: Overlap in terms of phenotypes associated with at least one gene in the raw version (top) and parental expansion version (bottom).

Figure 3: Real vs Random models comparison for GO term enrichment. This figure, along with the equivalent figures for KEGG and Reactome, are generated by the workflow and output in the general report (Supplementary File 1). **A**: Total numbers of Phenotype-GO term associations for the real and random models for different p-value thresholds. **B**: Total number of phenotypes for which at least one GO term could be associated. Black dashed line shows the initial number of phenotypes associated to SORs with HyI equal or greater than 2 in real model. Blue dashed line shows the number of phenotypes that were associated to a GO term after the enrichment analysis in real model.

Figure 4: SORs containing multiple genes with the same functional annotation. Boxplots show the proportion of Phenotype-GO term relations that involve SORs containing two or more genes annotated with the same enriched FunSys.

Figure 5: IC representation of Phenotype-GO pairs. Scatter plots show the IC values calculated for each HPO term (x-axis) and GO term (y-axis) from the Phenotype-GO term associations in the real data set, before **(A)** and after **(B)** random model filtering. Density plots show the distributions of the IC for HPO and GO separately. This figure is generated by the workflow and output in the general report (Supplementary File 1).

Figure 6: The contribution of each patient to the obtained filtered Phenotype-GO term associations. This figure is generated by the workflow and output in the general report (Supplementary File 1). **A** Number of genes affected in a patient (through SORs) compared to the percentage of them that are actually involved in a Phenotype-GO association from the patient phenotypic profile. **B** Percentage of patient's HPO phenotypes that are involved in Phenotype-GO term relations for which the enriched genes include at least one gene in the patient's mutation. Results are ordered by this percentage and grouped into percentiles. Colour represents the mean IC for the phenotypic profile (phenome) of the patients in each percentile.

Figure 7: Relationships between patient phenotypes, affected genes and relations inferred from the Phenotype-GO terms associations for Patient 1 [A,C,E] and 2 [B,D,F]. GO terms are only shown when at least one affected gene is annotated with them. Figure taken from Supplementary File 5. Such figures are generated automatically by the workflow for each patient. **A-B** Connections between affected genes and GO terms. **C-D** Connections between affected genes and annotated phenotypes. **E-F** Heatmap relating phenotypes and GO terms. Color intensity reflects the proportion of the genes for a GO term association that are affected in the patient.

Tables

Additional Files

Supplementary file 1 — General report
General HTML report with input, middle stages and output studies. Extended explanation is included into file.

Supplementary file 2 — Phenotypes GO report
Phenotype-GO HTML enrichments report with phenotype specific enrichment results. Extended explanation is included into file.

Supplementary file 3 — Phenotypes KEGG report
Phenotype-KEGG HTML enrichments report with phenotype specific enrichment results. Extended explanation is included into file.

Supplementary file 4 — Phenotypes Reactome report
Phenotype-Reactome HTML enrichments report with phenotype specific enrichment results. Extended explanation is included into file.

Supplementary file 5 — Patients GO report
Patient-GO HTML enrichments report with patients specific enrichment results. Extended explanation is included into file.

Table 1: Patients, genomic and phenotypic coverage statistics of the workflow executions for different sets of CNVs, for the raw and parentally expanded datasets. Highest values are indicated in bold. * For SORs mean size, the lowest value is in bold.

	Integrated Del + Dup		**Deletions**		**Duplications**	
	Expansion	Raw	Expansion	Raw	Expansion	Raw
Patients	**3,393**	**3,393**	2,529	2,529	1,083	1,083
CNVs	**3,888**	**3,888**	2,689	2,689	1,199	1,199
Genome coverage by CNV (bp)	**2,547,693,810**	2,225,567,000	1,917,120,878	1,506,862,774	1,490,002,548	958,935,409
Genome coverage by CNV (%)	**82.30%**	71.90%	61.93%	48.68%	48.14%	30.98%
Overlaping patients	**3,347**	2,844	2,440	2,001	1,003	710
SORs	**6,704**	6,231	4,170	3,640	1,788	1,399
Genome coverage by SORs (bp)	**1,849,586,973**	1,551,678,657	1,328,277,944	1,044,997,347	844,144,177	606,288,434
SORs (HyI >=2.0)	**5,520**	4,937	3,259	2,779	1,231	986
SORs mean size (bp)(HyI >= 2.0)	335,070	**314,296**	407,572	376,034	685,739	614,897
Genes	**26,435**	22,323	17,599	14,153	14,715	10,767
HPO terms	**900**	353	823	301	439	133
HPO-SOR links	**382,953**	47,429	175,213	21,513	51,068	5,329
HPO-SOR links (HyI >= 2.0)	**69,939**	18,533	41,288	10,139	11,074	2,179

Table 2: Summary statistics for the Phenotype-FunSys relations obtained for the real data and random models. The numbers shown in this table refer to associations with an adjusted p-value less than 10^{-3}. For the random models, the mean value of the hundred iterations ± standard deviation is shown.

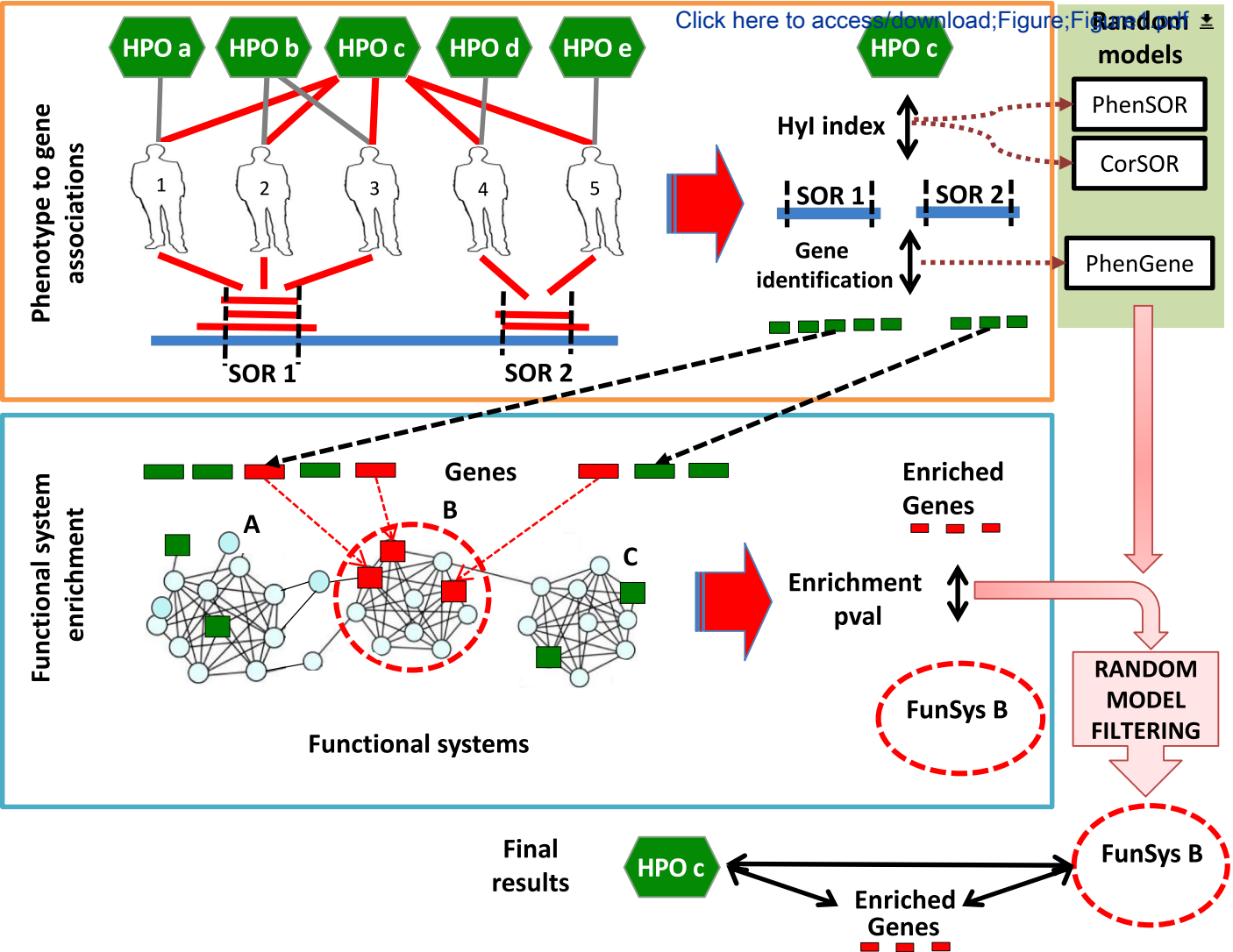
		Real	PhenSOR	CorSOR	PhenGene
	Genes	26,435	26,435	19,564 ± 352	26,435
GO	HPO terms	873	864 ± 11	840 ± 13	718 ± 30
	GO terms	3,989	3,854 ± 145	3,293 ± 198	1,599 ± 126
	HPO-GO pairs	23,967	22,051 ± 1,878	18,812 ± 2,677	3,699 ± 432
KEGG	HPO terms	102	135 ± 34	89 ± 29	3 ± 2
	KEGG terms	46	46 ± 6	31 ± 10	2 ± 2
	HPO-KEGG pairs	429	524 ± 209	247 ± 178	4 ± 9
Reactome	HPO terms	161	198 ± 97	43 ± 44	2 ± 2
	Reactome terms	87	87 ± 27	87 ± 34	4 ± 8
	HPO-Reactome pairs	429	524 ± 210	247 ± 178	4 ± 9

Table 3: Effect of random and parental filters on the final number of significant and unique Phenotype-GO term associations.

Dataset	HPO terms	GO terms	HPO-GO pairs
Raw	873	3,989	23,967
Random filt.	867	2,003	5,826
Random + parental filt.	609	1,931	3,247

Table 4: The top phenotypes associated with GO terms following random model filtering. Results are ordered by phenotype IC, using the integrated enrichment score to reorder in the case of ties. All phenotypes in this table have the maximum information content (7.42), as such this is not shown. Columns represent the following: Integrated-FS - Integrated FunSys enrichment score; FunSys - Number of FunSys associated with the functional system (gene ontology in this case); Genes-Enr - genes associated with the Phenotype that lead to association; Genes - total number of genes associated with the phenotype.

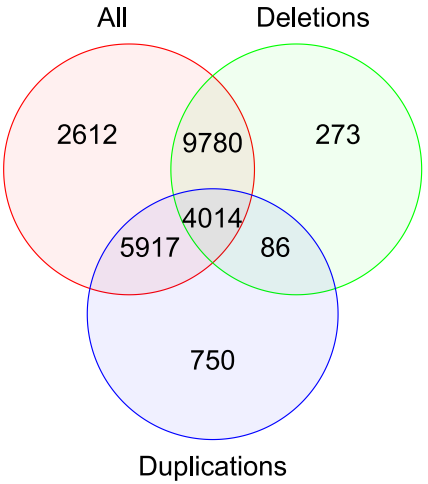
HPO-ID	Name	Integrated-FS	FunSys	Genes-Enr	Genes
HP:0001369	Arthritis	44.32	15	12	26
HP:0010818	Generalized tonic seizures	41.56	6	5	24
HP:0000020	Urinary incontinence	36.73	7	9	20
HP:0001622	Premature birth	36.52	6	17	60
HP:0000437	Depressed nasal tip	33.26	1	3	59
HP:0002187	Intellectual disability, profound	32.05	7	4	18
HP:0002558	Supernumerary nipple	32.05	7	4	18
HP:0011398	Central hypotonia	31.59	10	95	175
HP:0001647	Bicuspid aortic valve	27.24	2	3	50
HP:0010307	Stridor	26.21	11	5	8

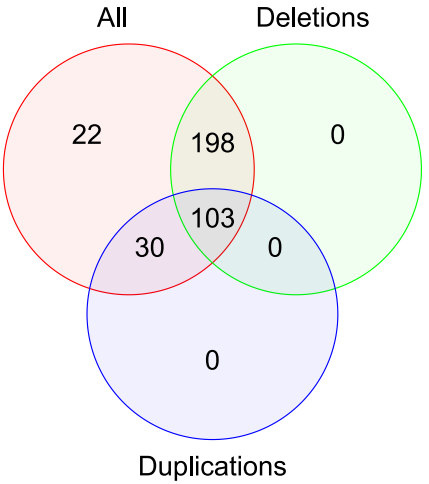


Genes

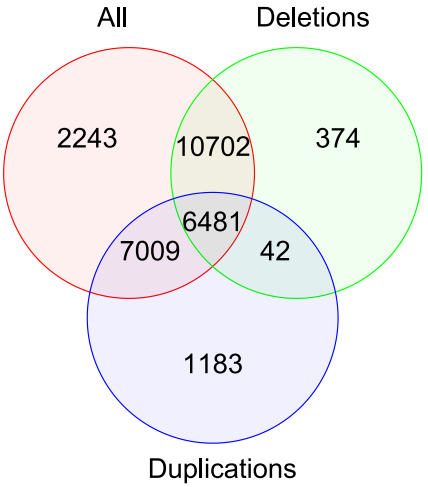
Phenotypes

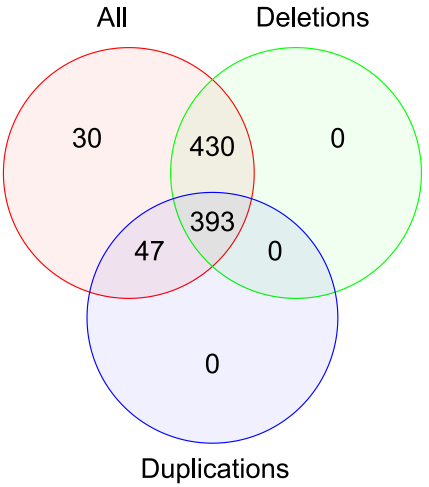
Raw





Parental expansion





P−val <= 0.001

P−val <= 1e−04

P−val <= 1e−05

GO-Phen links ± s.d.

Phenotypes with any GO terms

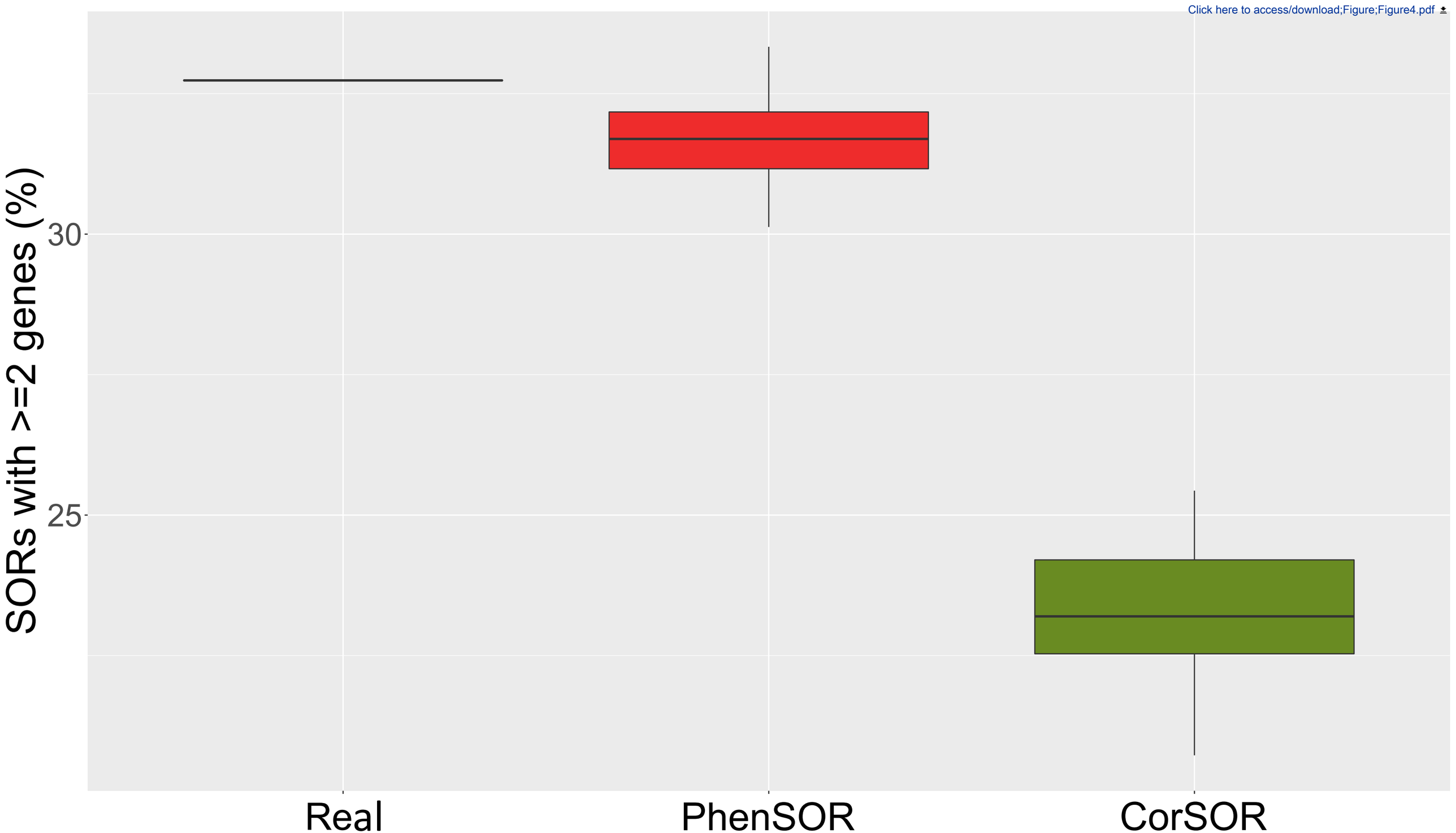
Real source: 900

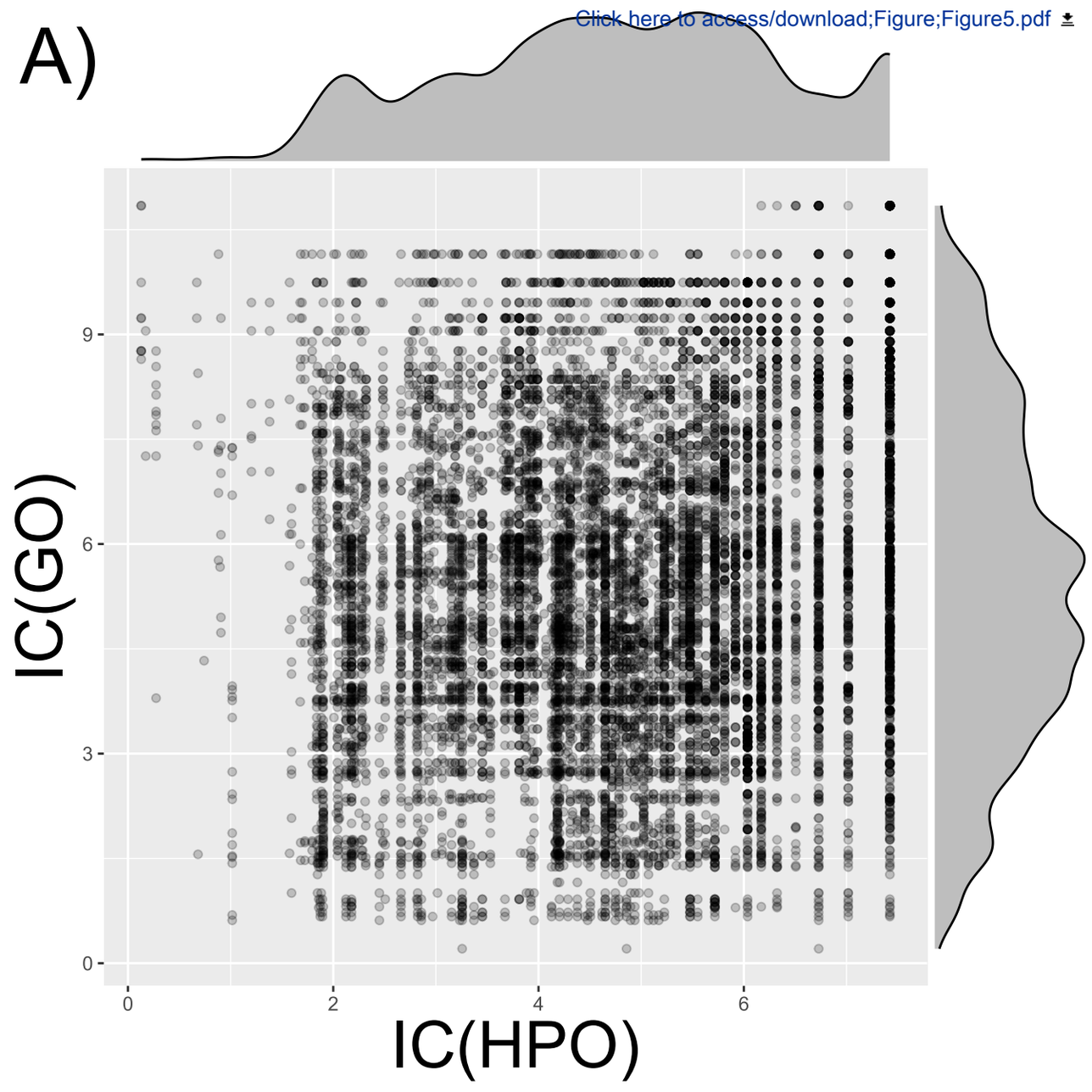
Real: 873

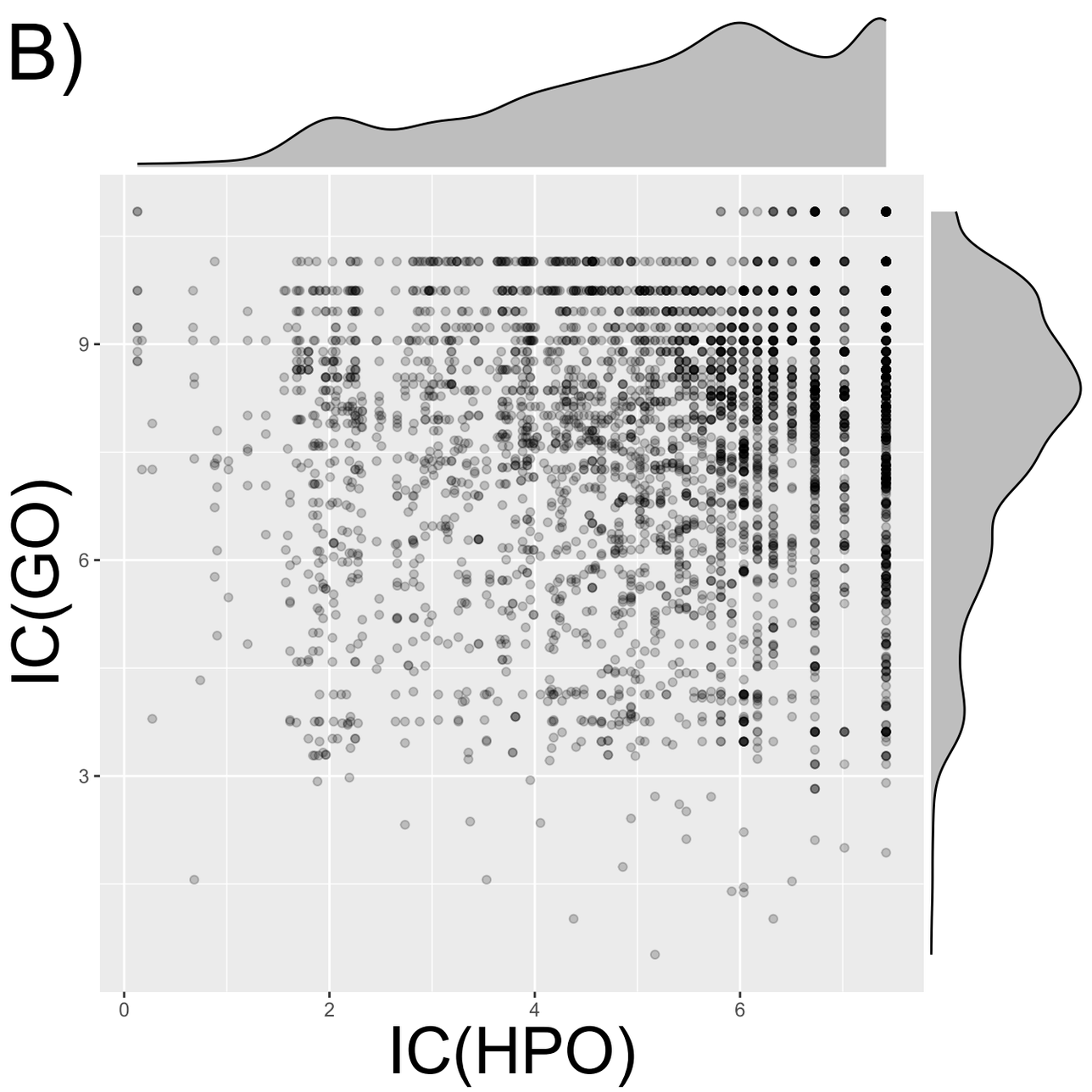
Real: 653

Real: 452

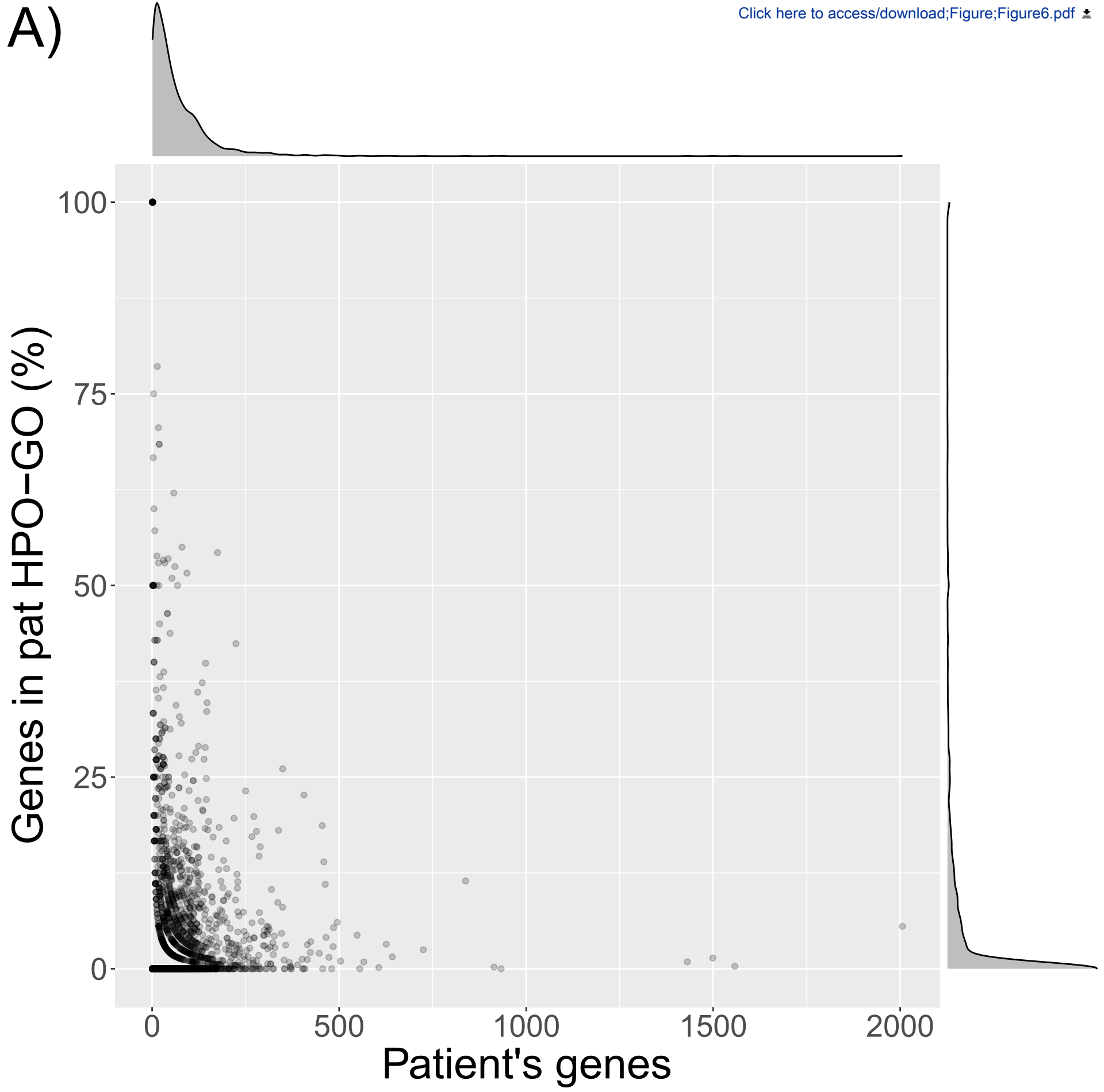
Model: PhenSOR CorSOR PhenGene Real



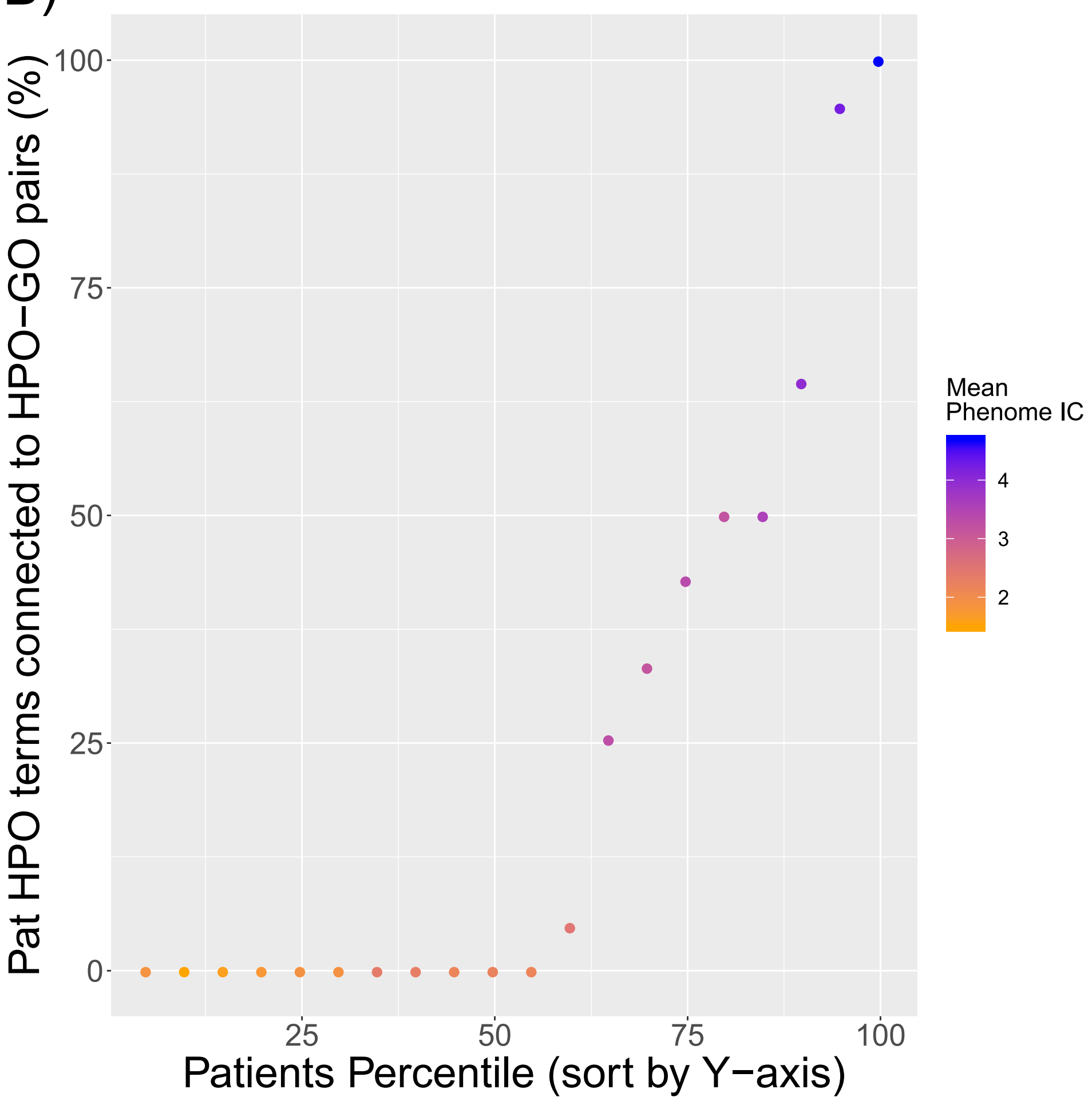


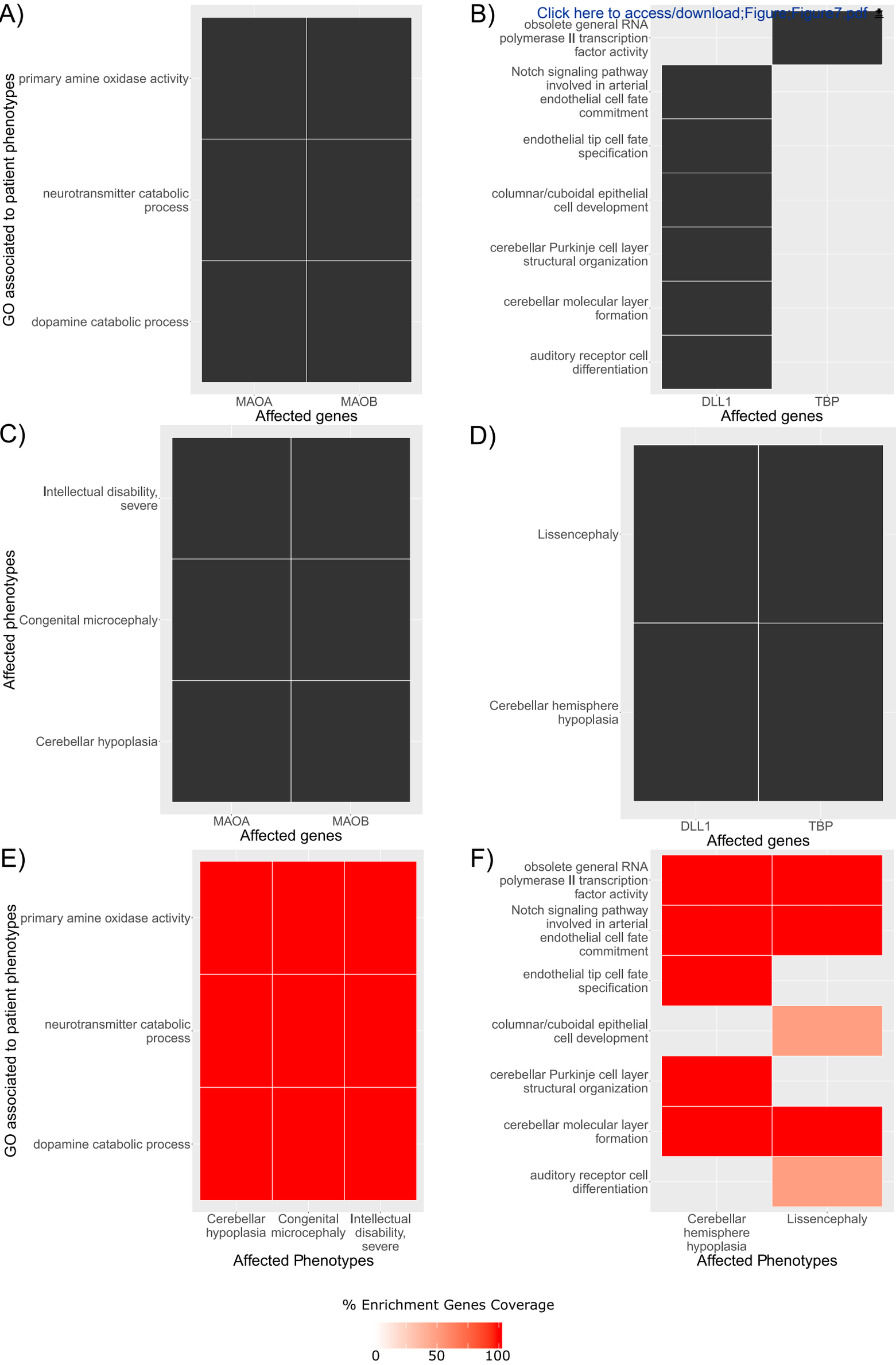


A)



B)





Click here to access/download
Supplementary Material
supplementary_3.html

Click here to access/download
Supplementary Material
supplementary_4.html

Click here to access/download
Supplementary Material
supplementary_2.html