Section B. Refinement of the hierarchical *P*-values in enrichments of GO or of SNOMED terms (Table 1 and Table 7 in Text S2, Supporting Figure 8 in Text S1).

In this study, we conducted enrichment statistics over two ontologies: Gene Ontology and SNOMED. These ontologies can be represented as directed acyclic direct graphs composed of nodes (classes of genes in the ontology such as a GO term or a SNOMED term) and edges (relationships to classes in the ontology) [1]. We developed a refinement algorithm to identify and filter out false positive p-values derived from enrichment studies in ontologies (hierarchical classifications) due to the inheritance of genes in ancestry classes of a significantly enriched class, a rarely mentioned problem of enrichment statistics that has also been reported by others [2,3]. In this manuscript, this algorithm identified from 0% to 59.3% of false positive enrichment results, with an average of 30.2% per enrichment (data not shown).

Equations S1, S2 and **S3** describe our refinement algorithm over the statistically significant results in an enrichment study. The ontologies are viewed as directed acyclic graphs where nodes are the entities (e.g. Gene Ontology terms or SNOMED concepts) and edges of the graphs are hierarchical relationships between these entities. In our enrichment studies, genes are classified to these entities/nodes. Nodes must meet the following two inclusion criteria: (i) the adjusted *P*-value of their gene enrichment (**Equation 4**, **Material and Methods**) is significant (adjusted $P \leq 0.05$), and (ii) the number of genes classified to this entity/node ≥ 3.

Definitions: V is the set of nodes, and E is the set of edges. Each node, $v_i \in V$, is assigned two types of descriptors of its relation with its neighboring nodes: descriptors of v_i are (i) one or more edges notes as $e_{i,j} \in E$ (when v_i is parent of v_j) or $e_{j,i} \in E$ (when v_j is parent of v_i), and (ii) an adjusted *P*-value from the enrichment study symbolized as p_i (**Equation 4**, **Material and Methods**). Each node, v_i, is also defined in terms of " sets" of hierarchical relationships (capital letters) as follow: (i) A_i for all parent nodes (1st degree ancestors), (ii) C_i for all children, and (iii) D_i for all descendents. These hierarchical sets are respectively described as $A_i = \{v_j \in V \mid \exists e_{j,i} \in E\}$, $C_i = \{v_j \in V \mid \exists e_{i,j} \in E\}$ and $D_i = \{v_j \in V \mid \exists e_{i,k_1}, e_{k_1,k_2}, \ldots, e_{k_{n-1},k_n}, e_{k_n,j} \in E\}$. The nodes that have the most statistically significant adjusted *P*-values (lower values) as compared to their hierarchic neighbors were identified as "Regional Minimum", noted V_{RM}, as defined in **Equation S1**. Among Regional Minimum nodes, we further excluded parents that have the same adjusted *P*-values as their children to conserve the most informative nodes: Refined Regional Minimum (V_{RRM}, **Equation S2**). The subsumed significant associations (Significant Descendants of Refined Regional Minimum or SDRRM) are defined in **Equation S3**. Finally, **Equation S4** defines the subset of included nodes (retained nodes) after refinement: those found in either **Equation S2** or **Equation S3**.

$$V_{RM} = \{ v_i \in V \mid \forall v_j \in A_i \cup C_i, p_j \geq p_i \}$$ (**Equation S1**)

$$V_{RRM} = \{ v_i \in V \mid \exists v_j \in V_{RM}, v_j \in A_i, v_j \in V_{RM} \}$$ (**Equation S2**)

$$V_{SDRRM} = \{ v_i \in V \mid \exists v_j \in V_{RRM}, v_i \in D_j, v_i \notin V_{RRM}, v_i \notin V_{RM} \}$$ (**Equation S3**)

$$V_{included} = \{ v_i \in V \mid v_i \in V_{RRM} \cup VSD_{SDRRM} \}$$ (**Equation S4**)

References of Protocol S1 Section B

1. Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9: 509-515.
2. Barry WT, Nobel AB, Wright FA (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 21: 1943-1949.
3. Prufer K, Muetzel B, Do HH, Weiss G, Khaitovich P, et al. (2007) FUNC: a package for detecting significant associations between gene sets and ontological annotations. BMC Bioinformatics 8: 41.