Supplementary materials

Initial re-curation of mutation data in IMEx

The data described in this paper has increasingly been made available to the research community since 2007 in PSI-MI XML2.5 files, but the capture of mutant data is incomplete in this format. Although the coordinate data is captured and a Controlled Vocabulary (CV) term describing the effect, the actual amino acid change is not captured. This issue has been addressed and corrected in the recently released PSI-MI XML3.0. In order to populate the replacement amino acid information, initial versions of our automated quality control pipeline were repeatedly applied over the entire data set, enhancing over 75% of the annotations. In addition to this, a significant number of entries have been manually re-curated, when there were too many changes in the reference sequence to allow automatic fixes. The full re-curation effort allowed to recover over 90% of existing annotations. The 2,310 annotations for which it was not possible to determine the exact amino acid change are excluded from the data set but kept in IMEx records as 'undefined mutation' and are scheduled for eventual re-curation.

Automated quality control pipeline for mutation entries in IMEx

UniProtKB entries change over time and accession numbers are obsoleted, merged and de-merged. The underlying protein sequences are often updated and positional features need re-mapping to the new sequence. In order to keep the data correctly annotated and in sync with current proteome builds as provided by UniProtKB, we have developed a 'mutations update' pipeline that is run before every IntAct release. This pipeline is run immediately after the 'protein update' pipeline, which keeps proteins in IntAct in sync with the UniProtKB entries they reference. Both pipelines are able to deal with most sequence changes, with difficult cases being referred to a human curator for manual checking. A diagram of how the 'mutations update' pipeline works can be seen in supplementary figure 1.

Every participant feature of type 'mutation (MI:0118)' or its children is checked using this pipeline before an IntAct release. After a number of preliminary sanity checks,

mutation features are then checked for range consistency, concordance between the HGVS-compliant short label and the 'resulting sequence' field and correct use of amino acid code. If any problems are found or there are changes due to an update in the reference UniProtKB entry, an annotation is added at the feature level and a new short label is proposed, if possible. All corrected entries undergo manual check and correction, if needed. Annotations that cannot be fixed are labelled as 'unspecified mutation' and discarded from the dataset using the special annotation 'no-mutation-export'. We retain a record of previous annotations in case they can be fixed in the future. During the design phase of this pipeline and the first bulk updates of historical mutation annotations, approximately 16,000 annotations were updated, with 1,400 requiring manual intervention. Since the introduction of the pipeline into routine IntAct production process in September 2016 and up to June 2018, 1,090 mutation annotations were automatically updated, with a further 634 requiring manual intervention.

Supplementary figure legends

Supplementary figure 1. 'Mutation update' pipeline diagram

* Interactor annotations: An interactor can hold several different annotations, which help us to determine its characteristics, such as if it can be kept in synch with a referenced entry in UniProtKB. If an interactor is marked with the annotation 'no-uniprot-update', it means it is not possible to keep it in sync with UniProtKB and we do not consider it for the short label generation process. ** Feature annotations: A feature can hold several different annotations, which provides context for the quality control procedure. If a feature is marked with the annotation 'no-mutation-export', we do not consider it for the short label generation. In case it holds the annotation 'no-mutation-update', we still check the feature for its consistency, but do not calculate a new short label for it.

Supplementary figure 2. Annotation depth by species

a: Relative percentage of mutation annotations per species (upper panel), along with distribution of proteins by number of annotations and species (lower panel); b:

Relative percentage of variants per species (upper panel), along with distribution of proteins by number of variants and species (lower panel).

Supplementary figure 3. Amino acid replacements frequencies

a: Detailed matrix plot for amino acid replacement frequencies over the whole data set; b: Detailed matrix plot for normalized replacement frequencies by mutation effect. Substitutions with non-standard amino acids and deletions are not shown for simplicity.

Supplementary figure 4. Computational annotations and the IMEx mutations data set (related to figures 4f and 4g)

a: Number of variants located in binding interfaces (curated and predicted), by effect; g: Normalized frequencies of variants reporting effects over interactions and their localization in binding interfaces.

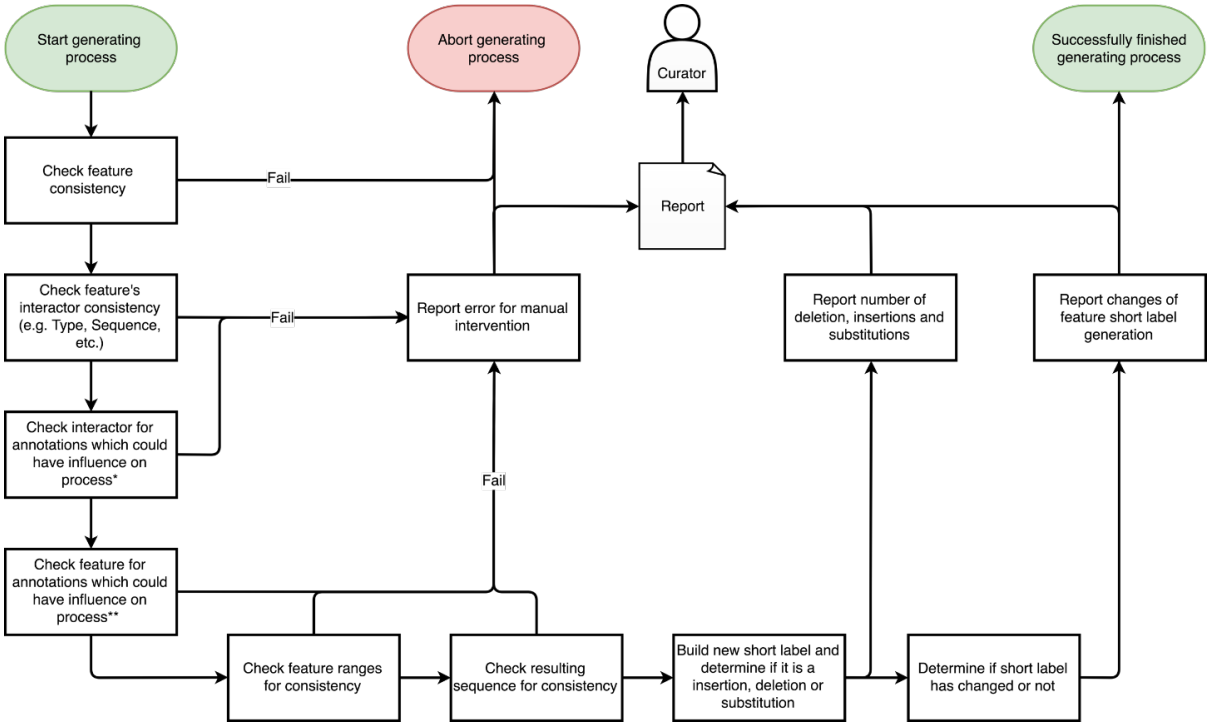
Supplementary figure 5. PathDIP annotation analysis of mutation-influenced interactions (related to figure 5)

p-value (log scale) for top 10 pathways in each set, grouped by topic if possible. Analysis in this figure was performed taking into account human proteins only.

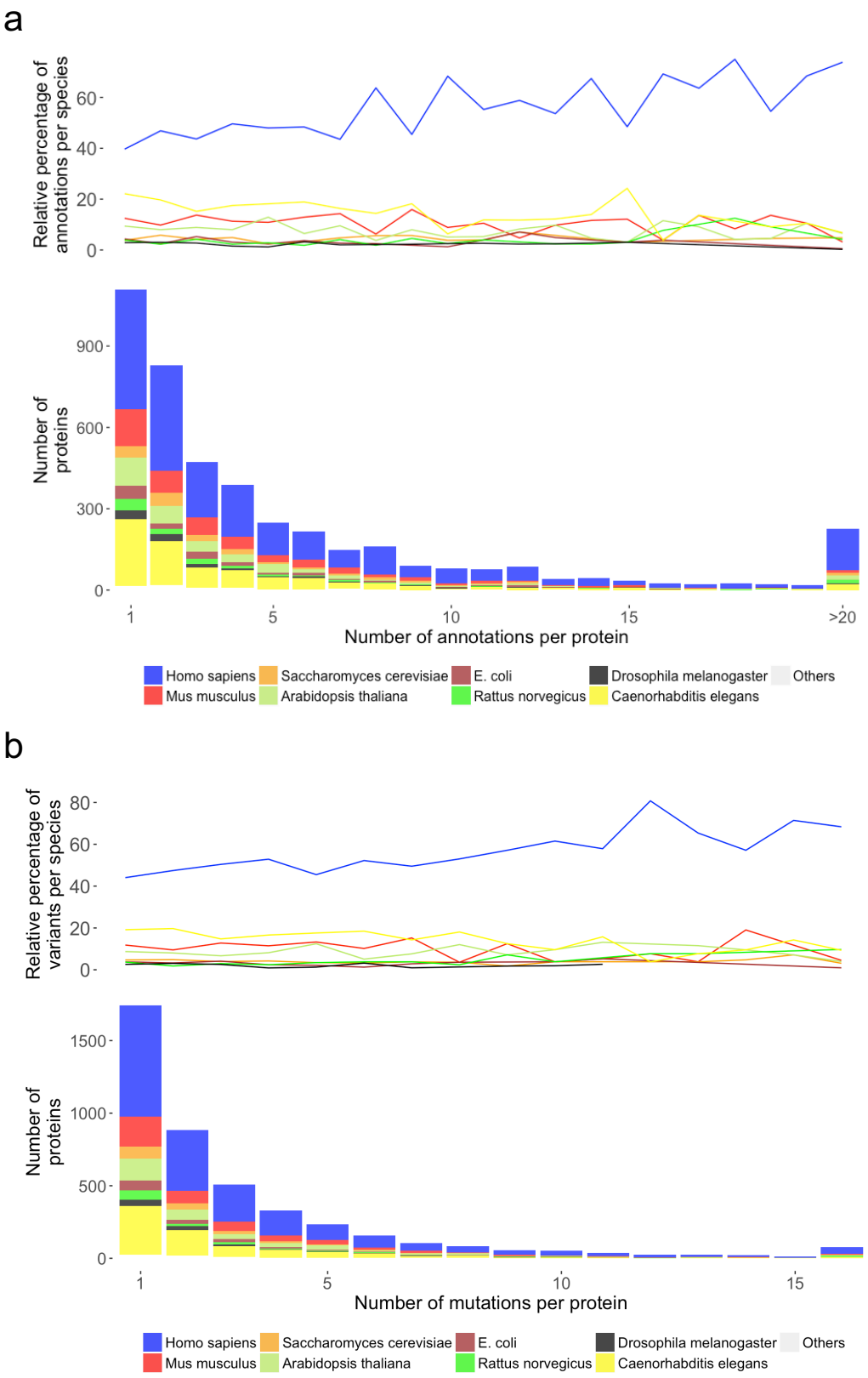
Supplementary figure 6. Original wordles for figure 5c

a. Original wordle for the “no effect” enrichment analysis results; b: original wordle for the “common mutated” enrichment analysis results.

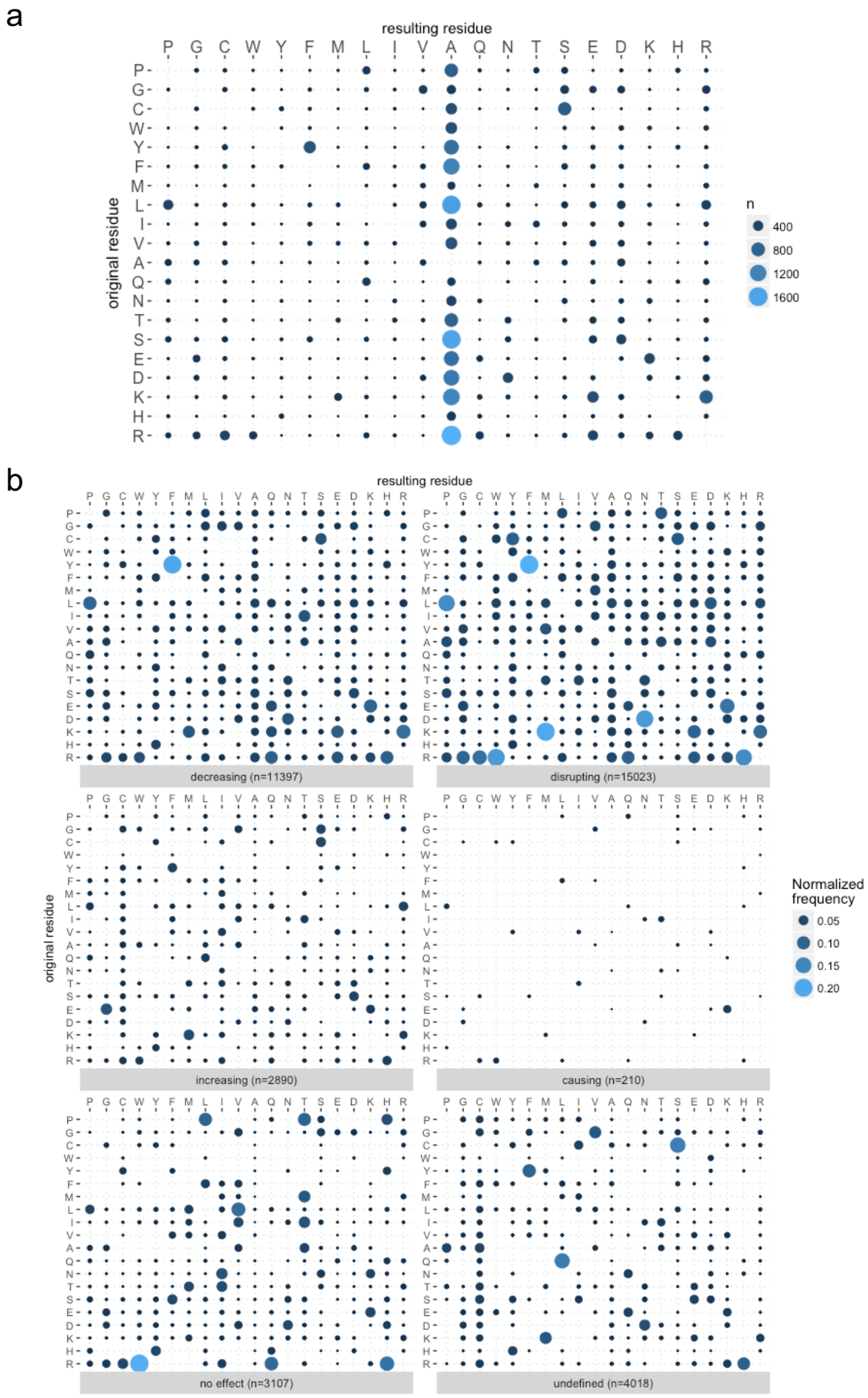
Supplementary figure 1



Supplementary figure 2

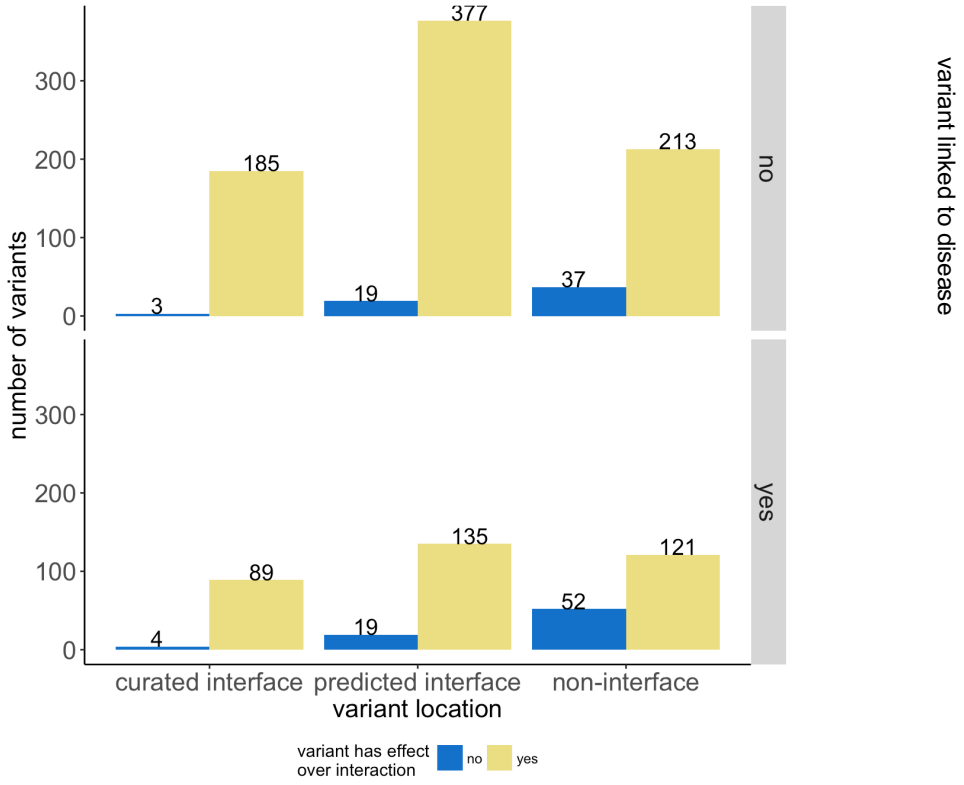


Supplementary figure 3

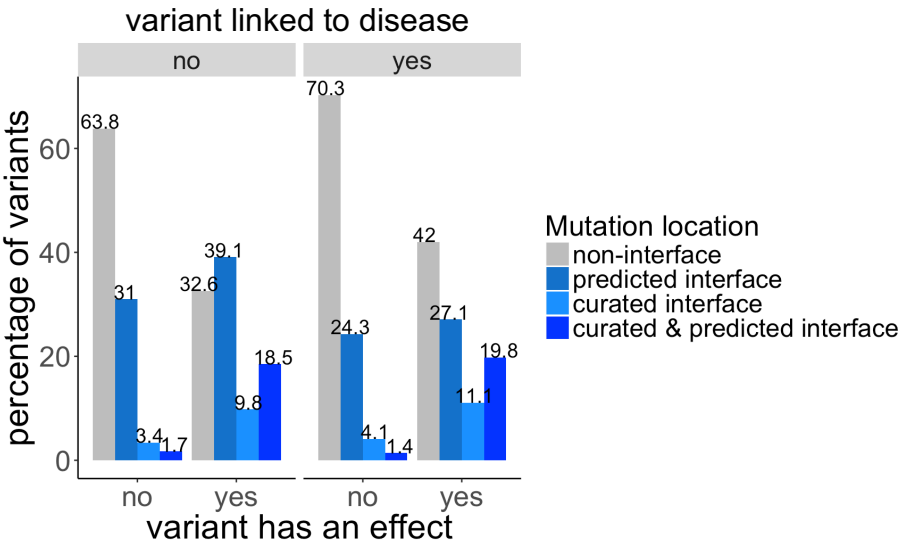


Supplementary figure 4

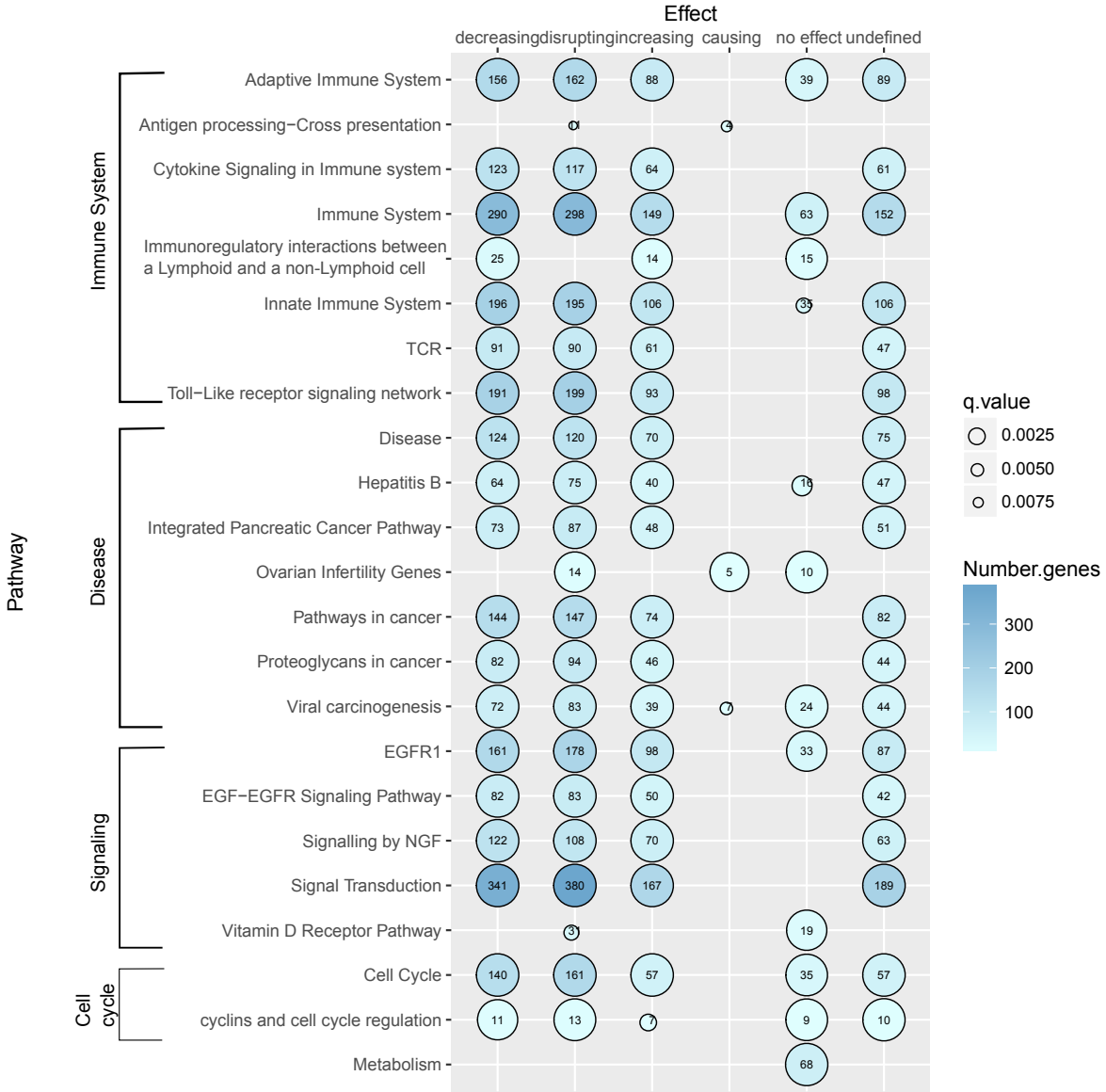
a



b



Supplementary figure 5



Supplementary figure 6

a

b