DGIdb - Mining the druggable genome

Malachi Griffith[*,†,1,2], Obi L. Griffith[*,†,1,3], Adam C. Coffman[1], James V. Weible[1], Josh F. McMichael[1], Nicholas C. Spies[1], James Koval[1], Indraniel Das[1], Matthew B. Callaway[1], James M. Eldred[1], Christopher A. Miller[1], Janakiraman Subramanian[3], Ramaswamy Govindan[3], Runjun D. Kumar[3], Ron Bose[3,4], Li Ding[1,2,3], Jason R. Walker[1], David E. Larson[1,2], David J. Dooling[1], Scott M. Smith[1], Timothy J. Ley[1,3,4], Elaine R. Mardis[1,2,4], Richard K. Wilson[1,2,4]

* These authors contributed equally to this work.
† To whom correspondence should be addressed.

Affiliations:
1. The Genome Institute, Washington University School of Medicine, St. Louis, MO
2. Department of Genetics, Washington University School of Medicine, St. Louis, MO
3. Department of Medicine, Division of Oncology, Washington University School of Medicine, St. Louis, MO
4. Siteman Cancer Center, Barnes-Jewish Hospital, Washington University School of Medicine, St. Louis, MO

List of Supplementary Materials

Supplementary Tables

Supplementary Table 1. Description of sources integrated in DGIdb
Supplementary Table 2. Data breakdown for sources integrated in DGIdb
Supplementary Table 3. Druggable gene categories and gene counts by source
Supplementary Table 4. List of annotations used to categorize drugs as belonging to the anti-neoplastic class
Supplementary Table 5. Breast cancer druggable candidates (available online)
Supplementary Table 6. Breast cancer known druggability by patient (available online)

Supplementary Figures

Supplementary Figure 1. Conceptual overview of DGIdb
Supplementary Figure 2. Database schema of DGIdb
Supplementary Figure 3. The DGIdb interface
Supplementary Figure 4. Percent of the potentially druggable genome that has actually been targeted according to DGIdb
Supplementary Figure 5. Overlap between genes with a known drug interaction and genes belonging to a potentially druggable category
Supplementary Figure 6. Recurrently mutated genes in breast cancer with known drug-gene interactions according to DGIdb
Supplementary Figure 7. Druggable and potentially druggable status of recurrently mutated genes in breast cancer

References

Supplementary Tables

Supplementary Table 1. Description of sources integrated in DGIdb

Data type	Source[ref]	Trust Level	Description of import process for integration into DGIdb
Gene definitions	NCBI Entrez Gene[1]	Non-curated	Entrez gene records formed the basis for all gene concepts in DGIdb to which all other gene instances were mapped. These records were imported from 'gene_info' and 'gene2accession' files obtained from the NCBI ftp site. Gene-gene interactions were also obtained from Entrez from the 'interactions' file.
	Ensembl[2]	Non-curated	Ensembl gene ids were imported and linked to Entrez gene records to improve mapping of sources based on Ensembl. These records were imported using the transcript GTF file available through Ensembl's ftp site.
Drug definitions	PubChem[3]	Non-curated	PubChem drug alternate names were imported from PubChem using the 'CID-Synonym-filtered' file obtained from the NCBI ftp site. Due to the size of PubChem, only drugs corresponding to a drug-gene interaction were imported.
Druggable gene categories	dGene[4]	Expert Curated	dGene druggable gene categories and members of these categories were imported from materials provided directly by the authors.
	Russ & Lampel[5]	Expert Curated	A single 'druggable genome' list was obtained directly from the authors of this source. To our knowledge it is no longer available online outside of DGIdb.
	Hopkins & Groom[6]	Expert Curated	Druggable gene categories were obtained from supplementary materials of the publication in the form of InterPro families and protein IDs. InterPro IDs were manually reviewed and updated. Updated records for each ID were obtained using the Ensembl BioMART Perl API.
	GO[7]	Non-curated	Manually selected categories (terms) and their corresponding protein products were imported from the Gene Ontology (GO). XML files were downloaded for each term by automated query of the AMIGO web service. The XML files were parsed and imported into DGIdb with a custom importer module.
Drug-gene interactions	My Cancer Genome[8]	Expert Curated	Raw targeted therapy data was obtained by parsing web content at mycancergenome.org with a custom Ruby module. The resulting data was manually curated to obtain interactions between drugs and genes. This involved resolving non-standard drug names and targets identified as complexes or pathways.
	TALC[9]	Expert Curated	Drug to target interactions were obtained by manual curation of PDF tables associated with the 'Molecular targeted agents and biological therapies in lung cancer' (TALC) publication. This involved resolving non-standard drug names and targets identified as complexes or pathways.
	TEND[10]	Expert Curated	Drugs, genes and interactions were manually curated from PDF tables of the 'trends in the exploitation of novel drug targets' (TEND) publication.
	PharmGKB[11]	Non-curated	The complete current dataset was obtained with permission from PharmGKB in flat file format. Gene and drug data files were downloaded directly from www.pharmgkb.org while relationship (i.e., interactions) data files were obtained by request. Only relationships linking drug entities to gene entities and classified as "associated" were imported.
	TTD[12]	Non-curated	A complete database dump was obtained from the TTD website in flat file format. Drug-gene interactions were defined by a custom DGIdb parsing module that performed data cleanup, sanity checks, mapping of targets to standard IDs, etc.
	DrugBank[13]	Non-curated	A complete database dump was obtained from DrugBank in XML format. Drug-gene interactions were defined by a custom DGIdb parsing module that performed data cleanup, sanity checks, mapping of targets to standard IDs, etc.

Supplementary Table 2. Data breakdown for sources integrated in DGIdb

		Data instances				
Data type	**Source[ref]**	**Genes**	**Drugs**	**Drug-Gene Interactions**	**Gene-Gene Interactions**	**Categories**
Gene definitions	NCBI Entrez Gene[1]	43,060	N/A	N/A	283,408	N/A
	Ensembl[2]	59,573	N/A	N/A	N/A	N/A
Drug definitions	PubChem[3]	N/A	10,192	N/A	N/A	N/A
Druggable gene categories	dGene[4]	2,257	N/A	N/A	N/A	10
	Russ & Lampel[5]	3,361	N/A	N/A	N/A	1
	Hopkins & Groom[6]	2,676	N/A	N/A	N/A	22
	GO[7]	5,982	N/A	N/A	N/A	26
Drug-gene interactions	MyCancerGenome[8]	169	226	835	N/A	N/A
	TALC[9]	153	263	573	N/A	N/A
	TEND[10]	437	989	2,243	N/A	N/A
	PharmGKB[11]	600	589	1,952	N/A	N/A
	TTD[12]	691	2,720	3,049	N/A	N/A
	DrugBank[13]	2,101	4,245	9,709	N/A	N/A
	Totals (unique)	**7,668**	**6,307**	**14,144**	**283,408**	**39**

Notes: The Russ & Lampel source does not provide separate categories of genes but rather provides a list of genes belonging to the "Druggable Genome". The Hopkins & Groom source was also all considered part of the generic "Druggable Genome" category in addition to having 21 specific druggable gene categories. The unique totals listed for genes and drugs represent just those involved in an interaction and/or belonging to a druggable gene category.

Supplementary Table 3. Druggable gene categories and gene counts by source

Category Name	dGene	Russ & Lampel	Hopkins & Groom	GO	Total	Go ID
ABC Transporter	N/A	N/A	24	102	102	GO0042626
B30_2 SPRY domain	N/A	N/A	89	N/A	89	N/A
Cell Surface	N/A	N/A	N/A	472	472	GO0009986
Cytochrome P450	N/A	N/A	57	N/A	57	N/A
DNA Directed DNA Polymerase	N/A	N/A	4	N/A	4	N/A
DNA Directed RNA Polymerase	N/A	N/A	2	N/A	2	N/A
DNA Repair	N/A	N/A	N/A	390	390	GO0006281
Drug Metabolism	N/A	N/A	N/A	34	34	GO0017144
Drug Resistance	N/A	N/A	N/A	351	351	GO0042493
Druggable Genome	2,257	3,027	2,668	N/A	3,850	N/A
Exchanger	N/A	N/A	14	N/A	14	N/A
External Side of Plasma Membrane	N/A	N/A	N/A	193	193	GO0009897
Fibrinogen	N/A	N/A	32	N/A	32	N/A
G-Protein Coupled Receptor	857	N/A	718	866	899	GO0004930
Growth Factor	N/A	N/A	N/A	164	164	GO0008083
Histone Modification	N/A	N/A	N/A	260	260	GO0016570
Hormone Activity	N/A	N/A	N/A	114	114	GO0005179
Ion Channel	N/A	N/A	181	392	401	GO0005216
Kinase	N/A	N/A	385	821	821	GO0016301
Lipase	N/A	N/A	10	N/A	10	N/A
Lipid Kinase	N/A	N/A	N/A	6	6	GO0001727
Myotubularin Related Protein Phosphatase	16	N/A	N/A	N/A	16	N/A
Neutral Zinc Metallopeptidase	N/A	N/A	5	181	181	GO0008237
Nuclear Hormone Receptor	48	N/A	48	48	54	GO0004879
Phosphatidylinositol 3 Kinase	14	N/A	18	N/A	24	N/A
Phospholipase	N/A	N/A	26	96	96	GO0004620
Protease	572	N/A	231	588	634	GO0008233
Protease Inhibitor	153	N/A	36	172	181	GO0030414
Protein Phosphatase	82	N/A	82	171	174	GO0004721
PTEN Family	7	N/A	N/A	N/A	7	N/A
RNA Directed DNA Polymerase	N/A	N/A	N/A	5	5	GO0003964
Serine Threonine Kinase	417	N/A	N/A	433	464	GO0004674
Short Chain Dehydrogenase Reductase	N/A	N/A	53	N/A	53	N/A
Thioredoxin	N/A	N/A	30	N/A	30	N/A
Transcription Factor Binding	N/A	N/A	N/A	405	405	GO0008134
Transcription Factor Complex	N/A	N/A	N/A	284	284	GO0005667
Transporter	N/A	N/A	53	1,194	1,195	GO0005215
Tumor Suppressor	N/A	N/A	N/A	727	727	GO0051726
Tyrosine Kinase	91	N/A	N/A	147	148	GO0004713
Total	**2,257**	**3,361**	**2,676**	**5,982**	**6,761**	

Supplementary Table 4. List of annotations used to categorize drugs as belonging to the anti-neoplastic class

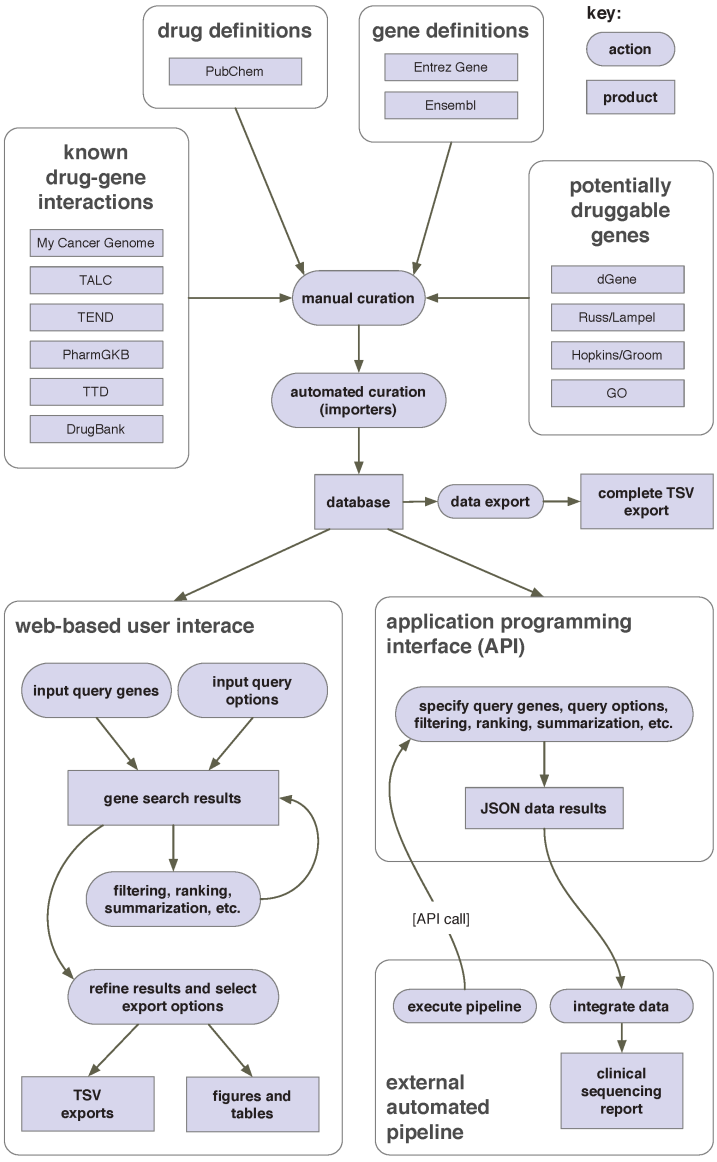
Drug type/class/category	Source
antineoplastic agents	TEND, DrugBank
antineoplastic agents, protein kinase inhibitors	TEND
antineoplastic adjuncts	TEND, DrugBank
antineoplastic agents, hormonal	TEND
antineoplastic agents, homeopathic agents	TEND
antineoplastic agent	TEND
antineoplastic agents	TEND
antineoplastic agent	TEND
antineoplastic agents, phytogenic	TEND
anticancer agents	DrugBank
antineoplastic	DrugBank
anticarcinogenic agents	DrugBank
antineoplastics	DrugBank

Note: All drugs reported from My Cancer Genome and TALC sources were considered anti-neoplastic since these sources report exclusively on anti-cancer agents. For TTD and PharmGKB the anti-cancer properties of drugs from these sources could not be determined from information provided.

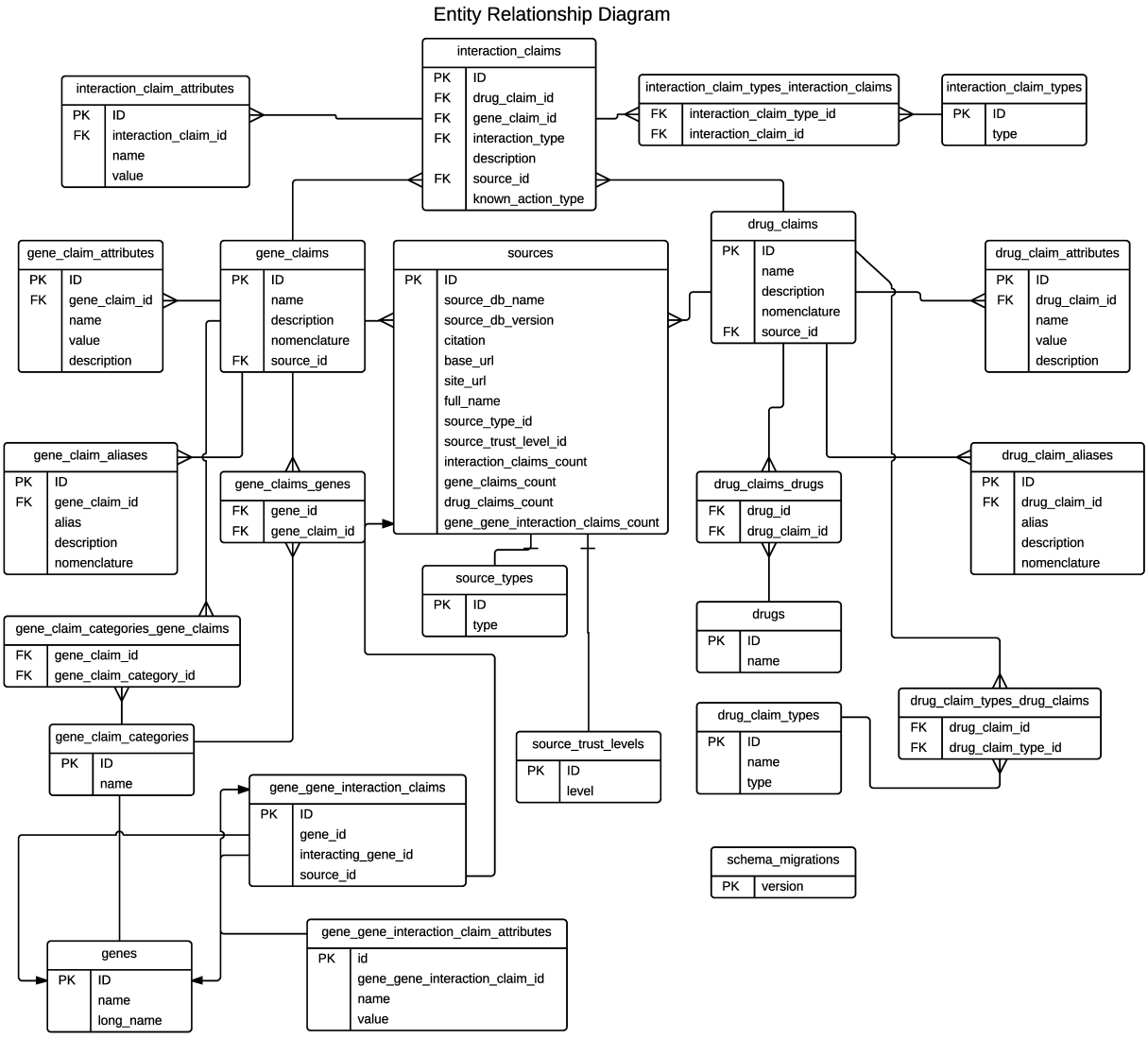
Supplementary Figures

Supplementary Figure 1. Conceptual overview of DGIdb

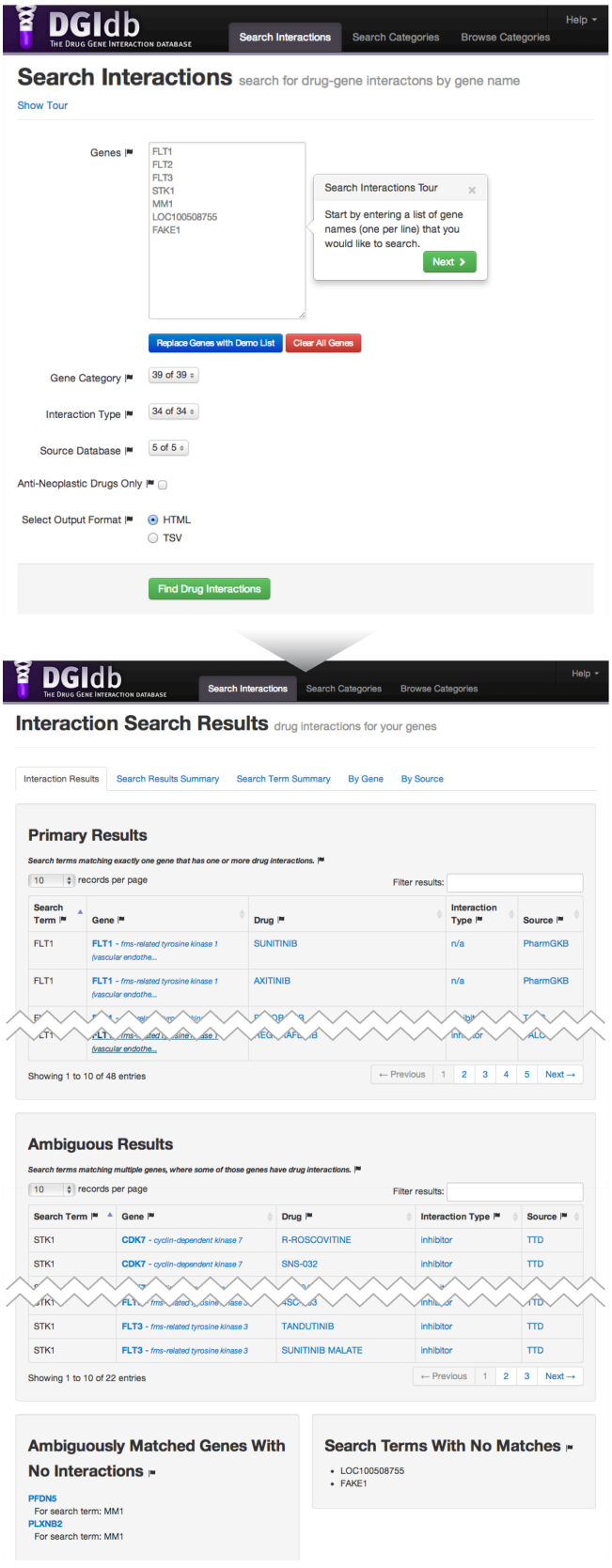
DGIdb aggregates information on known drug-gene interactions and potentially druggable genes from multiple sources using a combination of manual curation and automated parsing with importers customized to each source. All drug, gene, interaction, and category data are stored in a Postgres database that facilitates access through raw data downloads, a simple web interface, or an 'application programming interface' (API). The web interface (www.dgidb.org) allows users to input genes, apply filters, prioritize or otherwise interact with results and export to TSV. The API allows the same functionality to be automated and plugged into analysis pipelines. In our case, DGIdb represents the end point analysis of our clinical cancer sequencing, discovery, annotation and reporting pipeline.



Supplementary Figure 2. Database schema of DGIdb



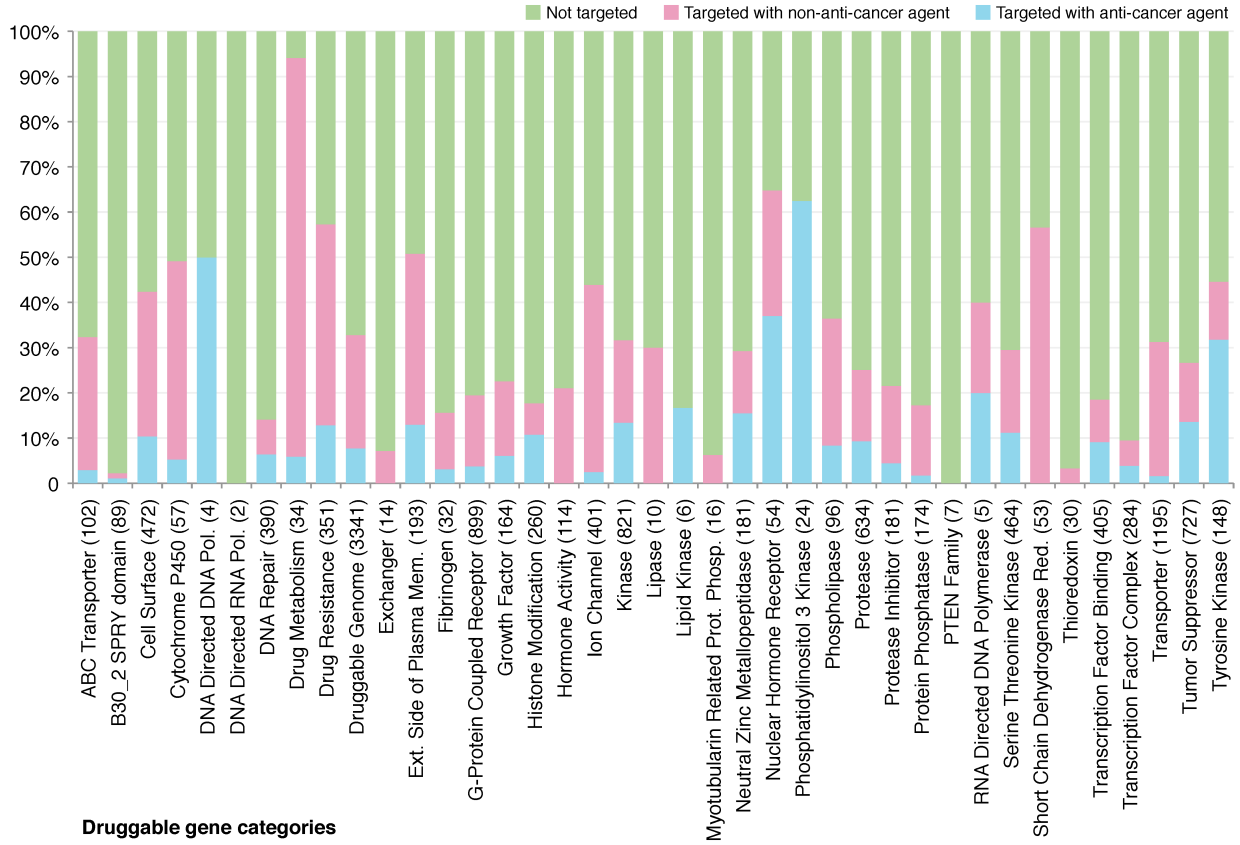
Supplementary Figure 3. The DGIdb interface



The DGIdb web interface allows exploration of the druggable genome with three simple tools (arranged along the site's top bar). Users can input a list of genes and quickly determine which of these genes, 1) have known drug-gene interactions using the 'Search Interactions' tool or 2) belong to potentially druggable gene categories using the 'Search Categories' tool. Alternatively, the lists of potentially druggable genes can be browsed directly by going to the 'Browse Categories' tool. Tutorials, FAQs, sources, downloads, web services, news, and contact details are available under the 'Help' menu. Shown here, the 'Search Interactions' page allows entry of a gene list of one gene to several thousand. A set of default genes can also be entered for illustrative purposes. Results can be limited to specific druggable gene categories, interaction types, or interaction source databases. If interested in only anti-cancer genes, results can be filtered to only 'anti-neoplastic' drugs. Output can be directed to an HTML web view or a tab-delimited (TSV) text file. Once submitted, the results page (or TSV output) displays all known drug-gene interactions for the input gene list. Search terms with ambiguous gene name mapping are shown but indicated as such. Results can be further filtered in real time using the filter results box. Additional display tabs provide a general summary of the search results, and detailed summaries broken down by search term, gene, and source.

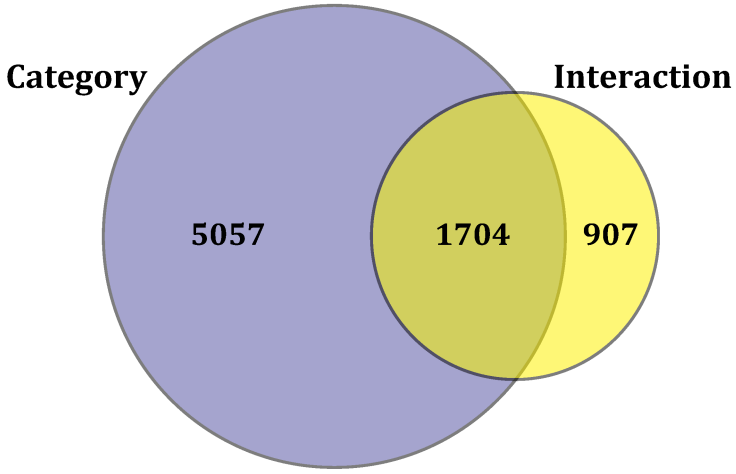
Supplementary Figure 4. Percent of the potentially druggable genome that has actually been targeted according to DGIdb

The 39 potentially druggable gene categories are listed and broken down by percentage of genes not targeted, targeted by a non-anti-cancer agent or targeted by an anti-cancer agent. Some category names have been abbreviated for display purposes. Pol = Polymerase; Phosp = Phosphatase; Red = Reductase; Ext = External; Mem = Membrane;



Supplementary Figure 5. Overlap between genes with a known drug interaction and genes belonging to a potentially druggable category

A Venn diagram summarizes the total number of genes belonging to at least one potentially druggable gene category (6,761), having at least one drug-gene interaction (2,611), and having both a drug-gene interaction and belonging to a druggable gene category (1,704).

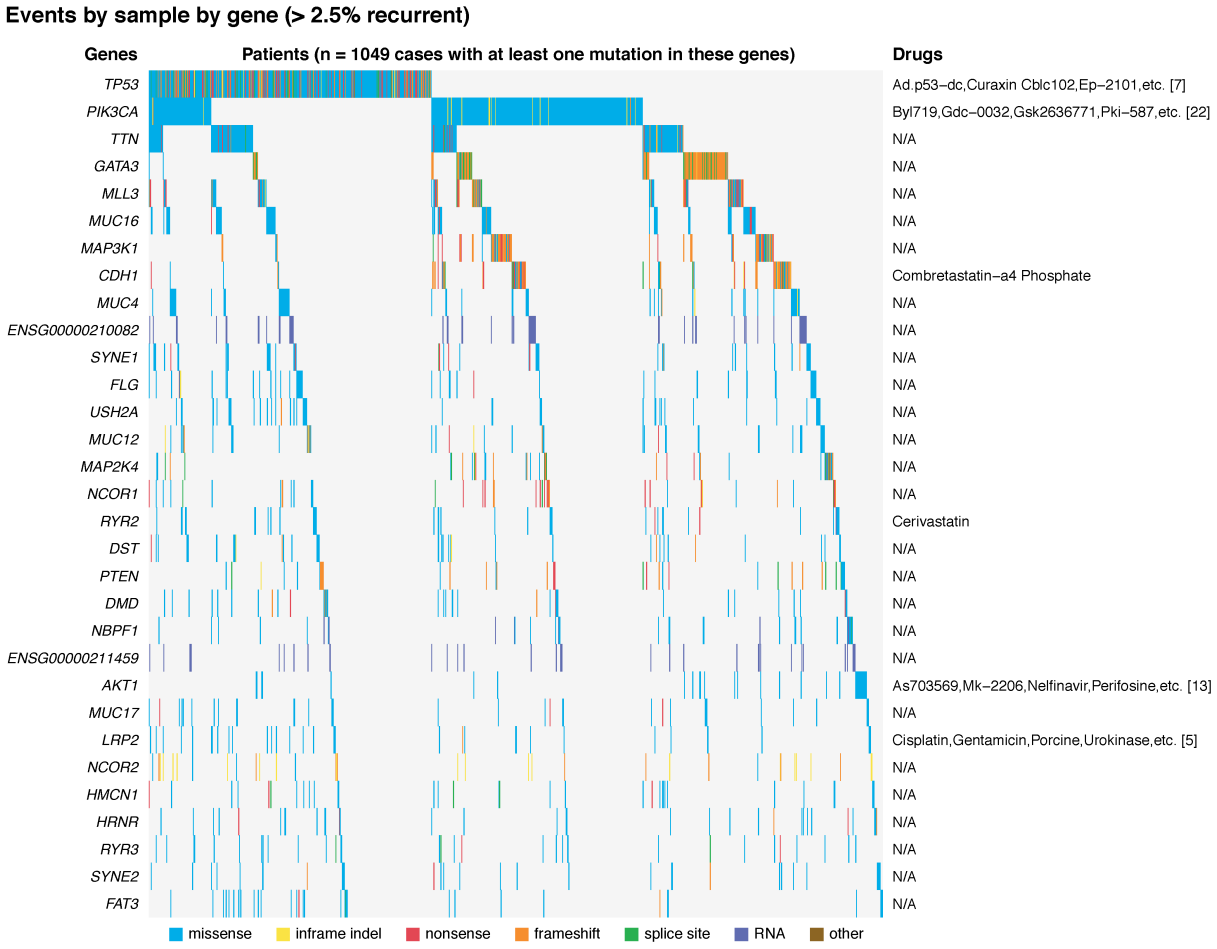


Supplementary Figure 6. Recurrently mutated genes in breast cancer with known drug-gene interactions according to DGIdb

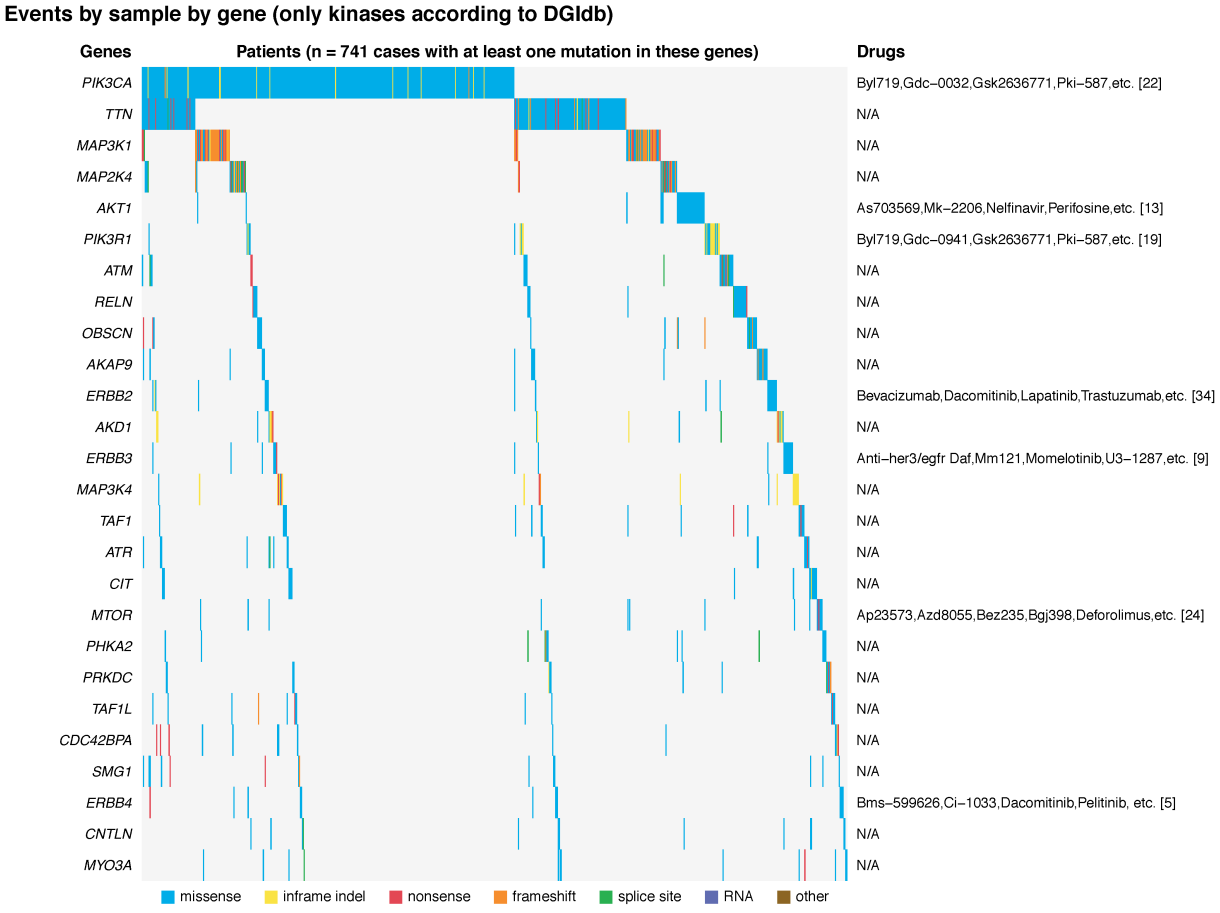
Heat maps were generated to display gene mutations observed in the genomes of individual patients (one patient per column). Each mutation is color-coded according to mutation type (see Methods). Gene names are indicated at the left and the names of drugs targeting those genes are indicated at the right. Where the number of drugs is large, the list has been abbreviated and the total number of drugs is indicated in square brackets after the drug list.

Supplementary Figure 6A

This heat map is limited to only the 31 genes mutated in at least 2.5% of 1,273 breast cancer cases. 1,049 tumors (82%) have a mutation in at least one of the genes displayed.

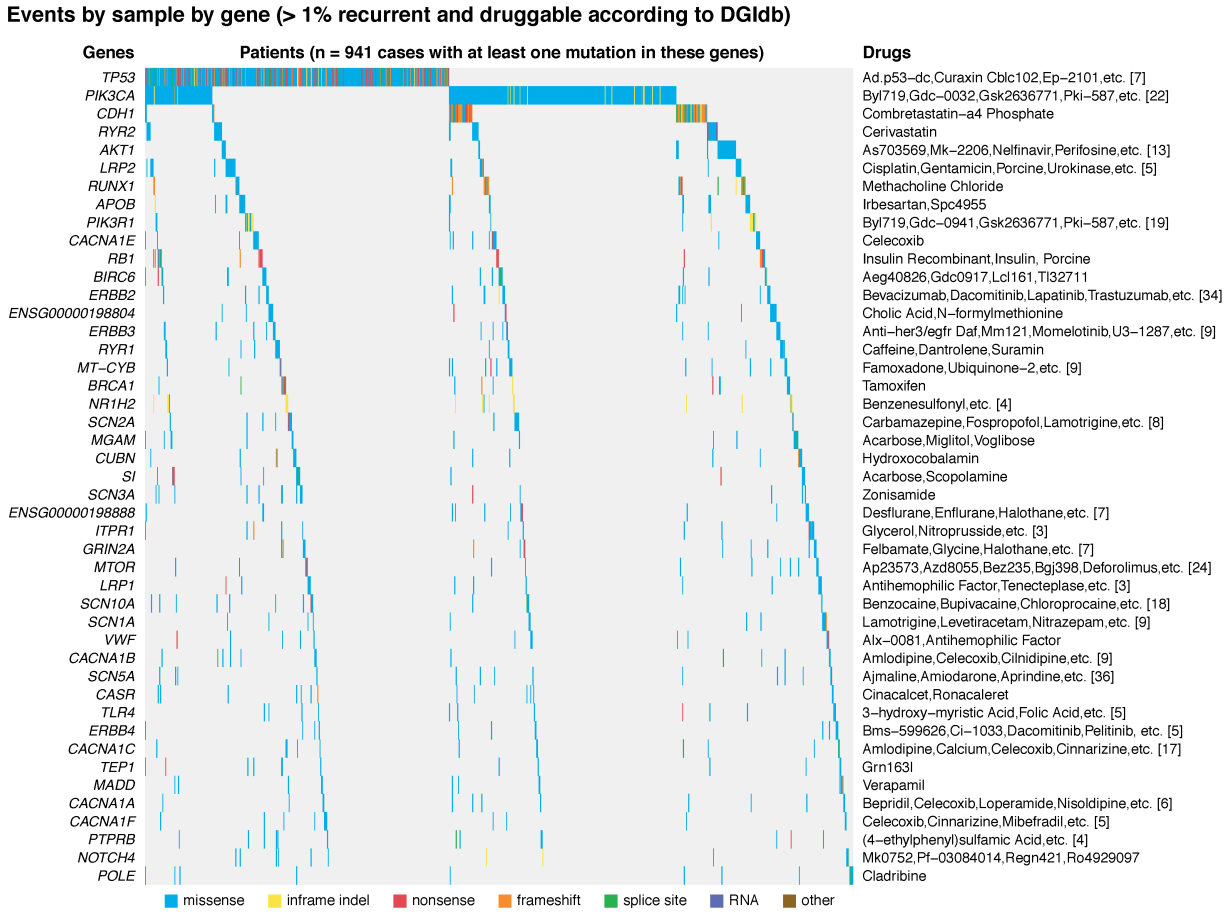


Supplementary Figure 6B
This heat map displays drug-gene interactions for only the 26 genes mutated in at least 1% of 1,273 breast cancer cases where the gene is also considered a kinase according to DGIdb. 741 cases (58%) have a mutation in at least one of the genes displayed. A complete data matrix including drug lists and sample identifiers is available as **Supplementary Table 5** online.



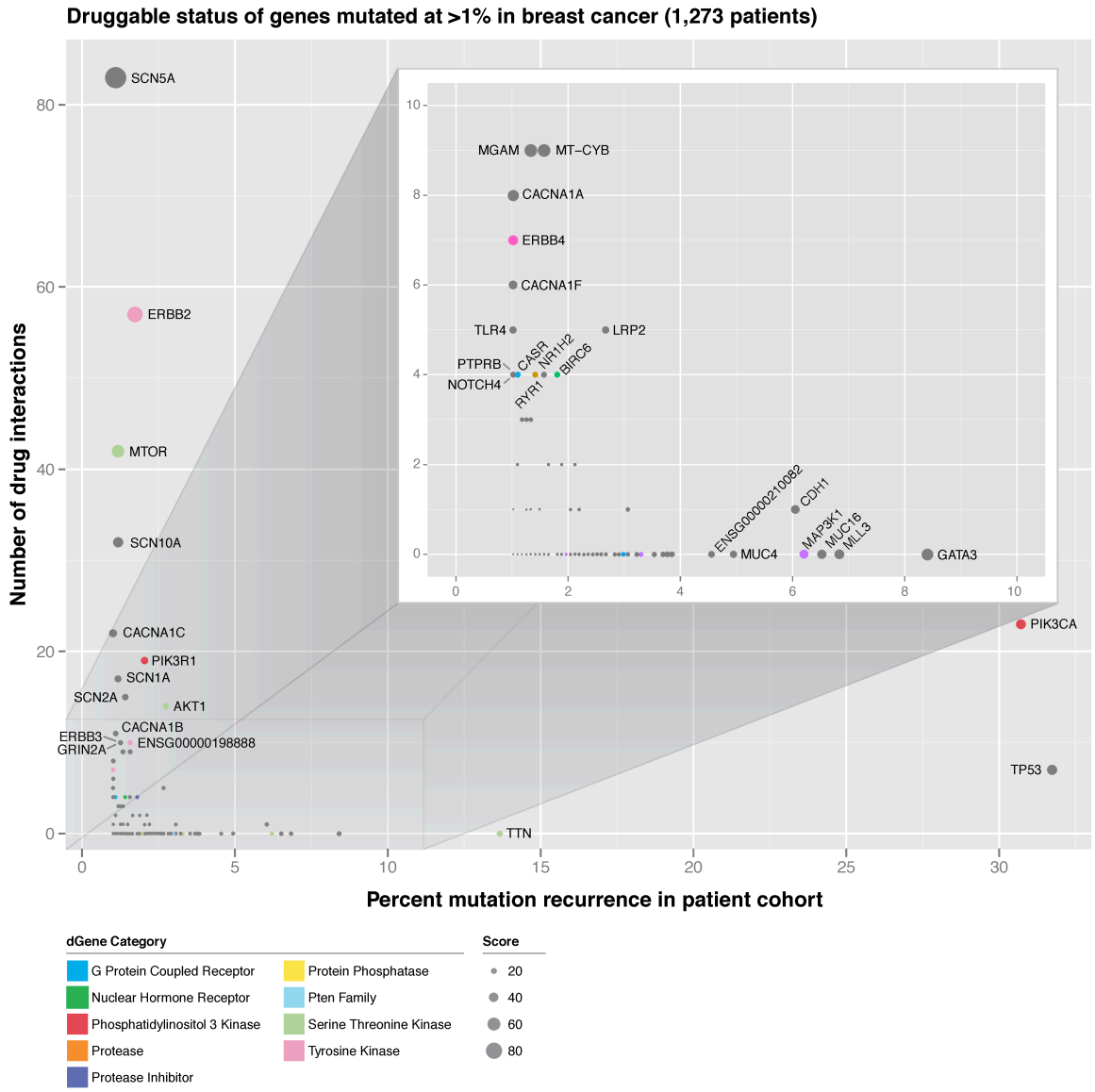
Supplementary Figure 6C

This heat map is limited to only the 45 genes with at least one drug-gene interaction and the 941 cases (73%) with a mutation in at least one of these genes are displayed (the complete dataset is available as **Supplementary Table 5-6** online).



Supplementary Figure 7. Druggable and potentially druggable status of recurrently mutated genes in breast cancer

Starting with a list of 315 genes recurrently mutated in breast cancer (>1% of cases) a scatterplot was generated to display the percent mutation recurrence rate within the cohort of 1,273 cases (ranging from 1% - 32% recurrence) and the number of known drug-gene interactions for each gene (ranging from 0 – 83 interactions). A potentially druggable 'score' based on both recurrence rate and number of drug-gene interactions is indicated by increasing point size (i.e. high scoring candidates have larger points). Finally, color-coding is used to highlight certain potentially druggable gene categories of particular interest for drug development efforts (the complete dataset is available as **Supplementary Table 5-6** online).



References

1. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* **39**, D52-57 (2011).
2. Flicek, P. et al. Ensembl 2011. *Nucleic acids research* **39**, D800-806 (2011).
3. Wang, Y. et al. PubChem's BioAssay Database. *Nucleic acids research* **40**, D400-412 (2012).
4. Kumar, R.D., Chang, L.W., Ellis, M.J. & Bose, R. Prioritizing Potentially Druggable Mutations with dGene: An Annotation Tool for Cancer Genome Sequencing Data. *PloS one* **8**, e67980 (2013).
5. Russ, A.P. & Lampel, S. The druggable genome: an update. *Drug discovery today* **10**, 1607-1610 (2005).
6. Hopkins, A.L. & Groom, C.R. The druggable genome. *Nature reviews. Drug discovery* **1**, 727-730 (2002).
7. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000).
8. Yeh, P. et al. DNA-mutation Inventory to Refine and Enhance Cancer Treatment (DIRECT): A catalogue of clinically relevant cancer mutations to enable genome-directed cancer therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research* (2013).
9. Somaiah, N. & Simon, G.R. Molecular targeted agents and biologic therapies for lung cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **6**, S1758-1785 (2011).
10. Rask-Andersen, M., Almen, M.S. & Schioth, H.B. Trends in the exploitation of novel drug targets. *Nature reviews. Drug discovery* **10**, 579-590 (2011).
11. McDonagh, E.M., Whirl-Carrillo, M., Garten, Y., Altman, R.B. & Klein, T.E. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomarkers in medicine* **5**, 795-806 (2011).
12. Zhu, F. et al. Update of TTD: Therapeutic Target Database. *Nucleic acids research* **38**, D787-791 (2010).
13. Knox, C. et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research* **39**, D1035-1041 (2011).