

White Paper

The Integrated Druggable Genome Database

Introduction

A druggable gene can be defined as a human gene that contributes to a disease phenotype and can be modified by a small molecule drug. The concept of the “druggable genome” has been used to denote a list of genes that can serve as suitable targets for developing therapeutic drugs. The term has been in vogue the last few years after the concept was published in a key paper by Hopkins and Groom in 2002¹. The Hopkins & Groom list of genes and gene families has been regularly revised and updated by others (Russ and Lampel, 2005)². Today, however, information on druggable genes is still scattered throughout many public databases, and there is no integrated repository where all this information can be accessed by the cancer research community.

The *Integrated Druggable Genome Database Project* has been undertaken by Sophic at the encouragement of NCI Principal Investigators and is intended to support a wide range of cancer research initiatives, providing information from several public sources. The gene list has been developed using Sophic’s knowledge integration methods, and software tools have been provided by Biomax Informatics AG, Munich Germany.

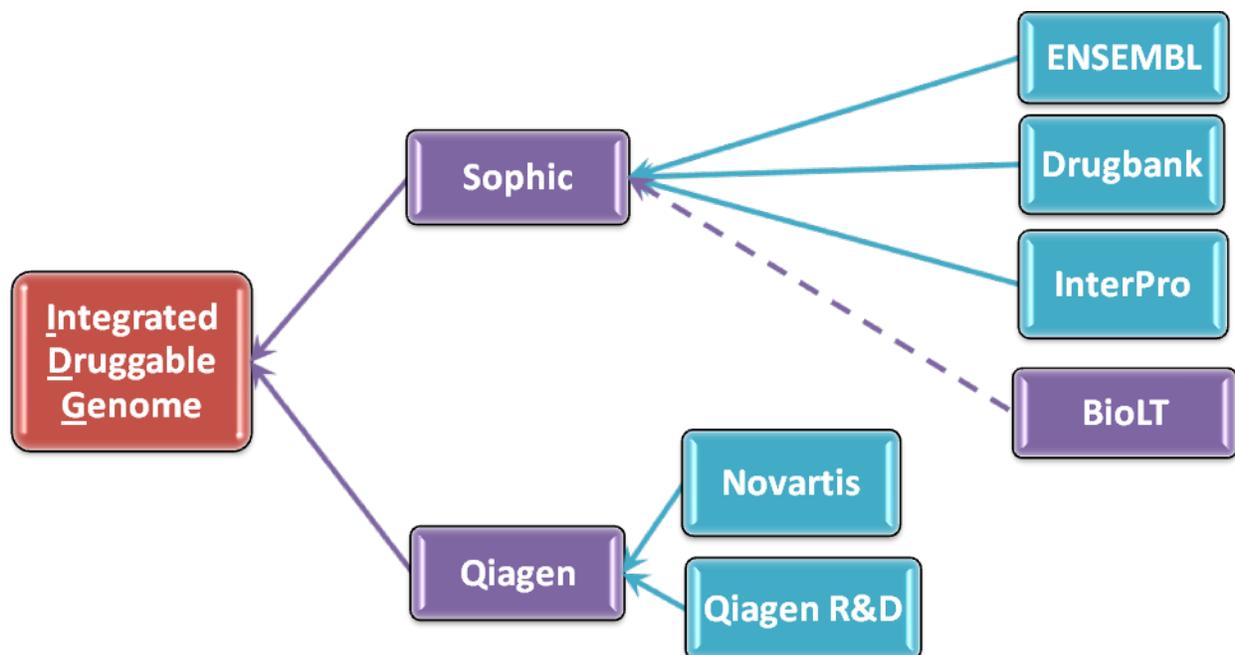
¹ Nat Rev Drug Discov. 2002 Sep;1(9):727-30

² Drug Discov Today. 2005 Dec;10(23-24):1607-10

Approach and Method

The Sophic approach was to generate gene lists retrieved from many different sources and integrate them into a single relational database. The structure of the database includes separate tables created to represent the different sources, which were then mapped to the HUGO³ gene names. Figure 1 is a diagram showing the data sources used to create the database.

Figure 1 Diagram of Data Sources used to create the Integrated Druggable Genome



- a. **HUGO Gene List:** A complete list of official HUGO approved names was downloaded from the HUGO website. All genes from different sources were mapped to the Hugo genes symbols. This is the “anchor” for the genes from sources listed below.

³ HUGO Gene Nomenclature Committee - <http://www.genenames.org/>

- b. **ENSEMBL List⁴:** contains the list of genes from the Hopkins and Groom paper mentioned earlier. The original list, obtained from supplementary data, contained all ENSEMBL IDs that were converted to HUGO symbols.
- c. **Drugbank⁵ List:** contains the list of genes identified as druggable from the public database provided by Prof. Dave Wishart and his group at the University of Alberta, Canada.
- d. **InterPro⁶-BLAST List:** contains the list of genes identified by Sophic scientists as druggable using the following method:
 1. *Extract all protein sequences from Drugbank corresponding to the InterPro families.*
 2. *Extract the Swissprot protein sequences corresponding to each gene name listed in the HUGO database.*
 3. *Carry out a BLAST search for each sequence in the Drugbank against all the downloaded Swissprot protein sequences. This method provides an expanded list of proteins that are then mapped back to their respective HUGO gene symbols.*
- e. **Sophic List:** Includes the integrated, non-redundant Hugo gene list.
- f. **The BioLT⁷™ List:** Includes the list of genes identified as druggable using the BioLT⁷ Text Mining Tool. BioLT is a literature mining tool that uses both a lexical and natural language

⁴ Homepage of the ENSEMBL - <http://www.ensembl.org/index.html>

⁵ Homepage of the Drugbank - <http://www.drugbank.ca/>

⁶ Homepage of the InterPro - <http://www.ebi.ac.uk/interpro/>

⁷ Homepage of BioLT – a Natural Processing Language tool that searches the literature databases - <http://www.biomax.de/products/biolt.php>

processing approach to efficiently mine information from literature sources. This tool was used to mine all gene names in the literature (Pubmed) that co-occurred with the name of a drug (as identified using Drugbank) and the term “inhibit*”. A thousand entries out of the **3,610** BioLT genes were manually checked and 77% were found to be viable “suspect” druggable genes.

g. **The Qiagen List:** Includes the list of genes identified as druggable by Qiagen Inc. and matched against the entries in the HUGO database.

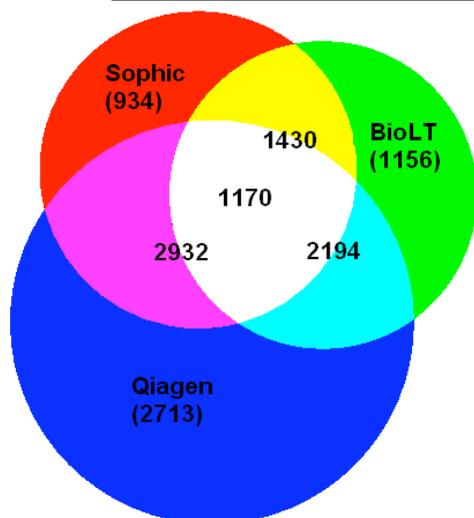
Summary of Results

Integrated Druggable Genome List combined from Different Sources: A total of **4,126** Hugo gene names were identified after combining the results from different sources including ENSEMBL, Drugbank and InterPro-BLAST approach. An additional **3,610** genes were identified using BioLT’s literature mining tool described above. The combined list totals **6,306** genes represented in a non-redundant Hugo gene list.

Comparison of the Qiagen-Gene List with the Sophic List: The Research and Development team at Qiagen Inc provided a separate list of druggable genes. The list was created by combining the list from the Novartis Research Foundation (<http://function.gnf.org/druggable/index.html>) and a separate list developed by Qiagen’s research group. For this project, Sophic selected only the 6,669 Qiagen genes that could be mapped to Hugo gene symbols. (Table 1).

Table 1 Number of Genes identified as Druggable by Sophic and Qiagen

<u>Sophic list</u>	4126
<u>BioLT list</u>	3610
<u>Qiagen list</u>	6669



The Integrated Druggable Genome: Figure 2 shows a Venn diagram of the Integrated Druggable Genome combining the sources from Sophic, Qiagen and the text mining approach. Since the genes identified by the text mining have not been fully curated, this list is shown as a separate entity⁸.

Numbers shown in each colored circle represent the number of genes unique to that category. Figure 2 shows genes common between different lists as well as unique

Figure 2 - Schematic diagram showing the Comparison of Genes identified by Sophic and Qiagen

genes from any of the respective lists (Sophic and Qiagen and BioLT), separately identified.

⁸ The list has been spot checked for about 1000 entries and was found to have an acceptance rate of ~77%. The process involved examining if identified gene names were valid and had a co-occurrence with a valid drug name (as identified from Drugbank) and the term 'inhibit.'

There were **1,170** genes that were found in all the three sources. The Sophic list contained **934** unique genes that were not present in the remaining two lists. Similarly, the Qiagen list and the BioLT list were found to have **3,737** and **1,156** genes, respectively.

Confidence Level Classification of Genes: All the genes in the HUGO column were rank-ordered based on the number of gene lists and the levels of rigor applied to each gene. The numbers of genes found at each Confidence Level are:

Confidence Level	Number of Druggable Hugo Genes
5	240
4	624
3	1113
2	2407
1	4664
Total	9048

Level 5 genes represent the highest confidence since these genes were identified as viable druggable targets across all the sources. Decreasing levels down to level 1 indicate that these genes are worthy of further analysis and may be of value.



Discussion and Future Plans

The Integrated Druggable Genome Database will be made publically available on Sophic's website and is intended to support complex disease research by helping scientists identify or validate potentially valuable druggable gene targets. The database will be uploaded into the BioXM™ Knowledge Management System and mapped to all relevant sources of potentially valuable scientific information. The Druggable Genome Database will be updated by Sophic on a regular basis with the same or improved methods.