IRDiRC

Draft Preparatory Document for Workshop on Data Mining/ Repurposing Initiatives in the Field of Rare Diseases

Prepared by

IRDiRC Scientific Secretariat

On behalf of the

Data Mining/ Repurposing for Rare Diseases Task Force

December 14, 2015

Table of Contents

Notes

This document is intended to provide readers with the necessary background information on the work to date in the field of data mining and repurposing initiatives in order to prepare discussions of the upcoming workshop.

The present document includes preparatory documentation concerning data mining and repurposing issues and a description of initiatives to repurpose existing drugs for rare diseases outside of their original intent. It also includes an overview of the need for new medication in the field of rare diseases and of the questions to be debated to advance the field.

The IRDiRC Task Force

The International Rare Diseases Research Consortium (IRDiRC) was set up to maximize scarce resources and coordinate research efforts in the rare diseases field, with the clear goal to boost the research and development process to help deliver effective therapies as soon as possible. IRDiRC aims to stimulate and coordinate basic and clinical research, by promoting links between existing resources, fostering the molecular and clinical characterization of rare diseases and encouraging translational, preclinical and clinical research.

The Therapies Scientific Committee of IRDiRC has issued recommendations on essential actions selected for their high leverage effect to unlock the potential of rare disease therapy development. Among them, the Therapies Scientific Committee recommends:

- **Encouraging, supporting and establishing early and continuous dialogue on clinical development strategy and wide evidence generation** (e.g. natural history, registry, clinical trial design, clinical endpoints, surrogate endpoints, patient relevant outcomes, regulatory strategy, medical practice, public health strategy, **data mining, drug repurposing**) with all relevant stakeholders such as patients' representatives, medical experts, researchers, scientific societies, regulators, health technology assessors, payers and sponsors when appropriate. This could be done through dedicated workshops, safe harbors where knowledge could be shared in a non-competitive manner.
- **Data mining,** extracting hidden patterns from large amounts of data to identify potential targets based on existing knowledge. These data mining approaches take advantage of methodologies and tools available from bioinformatics, chemoinformatics and network and systems biology, aided by various databases. By connecting data on drugs, proteins and diseases, data mining methods may determine new drug targets or enable the repurposing of pharmaceutical compounds.
- **Repurposing drugs,** finding new uses for existing drugs, so that more therapies are available for patients with rare diseases. Data mining for new targets and repurposing opportunities are essential steps to reach IRDiRC's goal of having 200 new therapies by 2020.

In order to make a decisive step to reach these objectives, the IRDiRC Executive Committee decided to set up a Task Force on Data Mining and Repurposing in the field of rare diseases, established in May 2015, composed of the following members acting as Steering Committee:

- Dr Dorian Bevec (Therametrics, Switzerland)
- Prof Benoît Deprez (APTEEUS, France)
- Dr Peter Bram 't Hoen (LUMC, The Netherlands)
- Ms Caroline Kant (EspeRare, Switzerland)
- Dr Frédéric Marin (GMP-Orphan, France)
- Dr Madhu Natarajan (Shire, USA)
- Dr Jordi Quintana (Plateforma Drug Discovery, Spain)
- Dr Noel Southall (NIH/NCATS, USA)

The following IRDiRC Scientific Committee members will also participate in this Task Force:

- Dr Jeffrey Krischer (University of South Florida, USA)

In addition to the Steering Committee, the following individuals have been proposed to participate in the discussion process:

- Dr Diego di Bernardo (Universita' "Federico II" of Naples, Italy)
- Mr Philippe Bissay (HAC Pharma, France)
- Dr Evan Bolton (PubChem, USA)
- Dr Joaquin Dopazo (CIPF Valencia, Spain)
- Dr Joel Dudley (Mount Sinai hospital, USA)
- Dr Karen Eilbeck (University of Utah, USA)
- Dr Tudor Groza (Garvan Institute of Medical Research, Australia)
- Dr Jayne Hehir-Kwa (Radboud University Medical Center, The Netherlands)
- Dr Virginie Hivert (EURORDIS, France)
- Mr Yann Le Cam (EURORDIS, Belgium)
- Dr Subha Madhavan (Georgetown University Medical Center, USA)
- Prof Christopher McMaster (Dalhousie University, Canada)
- Dr Ramaiah Muthyala (University of Minnesota, USA)
- Dr May Orfali (Rare Disease Global Medical lead, MDG, SCBU, Pfizer Inc., USA)
- Dr Helen Parkinson (EMBL-EBI, UK)
- Dr Karin Rademakers (UMC Utrecht, The Netherlands)
- Dr Marco Roos (Leiden University Medical Center, The Netherlands)
- Dr Philippe Sanseau (GSK, UK)
- Dr Nick Sireau (patient representative and chairman AKU Society, UK)
- Dr Elia Stupka (Boehringer Ingelheim, Germany)
- Dr Stelios Tsigkos (Orphan Medicines Office, European Medicines Agency, UK)

The members of the Task Force are requested to:

1. Identify the topics to be explored in order to identify ways to boost data mining for the identification of new therapeutic targets or to reposition drugs.
2. Review the pre-workshop report; preparation of document by the IRDiRC Scientific Secretariat.
3. Participate in the expert discussions at an invited workshop, to be held in Q2, 2016
4. Review the post-workshop report, including actions to be implemented, and of the subsequent publication for a peer-reviewed journal; preparation by the IRDiRC Scientific Secretariat.

Acronyms

BPCA	Best Pharmaceuticals for Children Act
CDD	Collaborative Drug Discovery
CUI	Concept Unique Identifiers
CWHM	Center for World Health and Medicine
EMA	European Medicines Agency
EU	European Union
FDA	USA Food and Drug Administration
GEAS	Gene Set Enrichment Analyses
GO	Gene Ontology
GWAS	Genome-wide association studies
IE	Information extraction
IP	Intellectual property
IR	Information retrieval
IRDiRC	International Rare Diseases Research Consortium
MATADOR	Manually Annotated Targets and Drugs Online Resource
MRC	Medical Research Council, UK
NCATS	National Centre of Advancing Translational Sciences
NCGC	NIH Chemical Genomics Center
NGS	Next-Generation Sequencing
NIH	National Institutes of Health, USA
NPC	NGCG Pharmaceutical Collection
OMIM	Online Mendelian Inheritance in Man
OPEN ACT	Orphan Product Extensions Now Accelerating Cures and Treatments Act
PCA	Principle Components Analysis
PDTD	Potential Drug Target Database
RDRD	Rare Diseases Repurposing Database
TarFisDock	Target Fishing Docking
TTD	Therapeutic Target Database
TRND	Therapeutics for Rare and Neglected Diseases
UMLS	Unified Medical Language System

General Background on Data Mining/ Repurposing

Introduction

Rare diseases concern a relatively small number of persons; effectively a disease is called rare when it affects one in 1500 to 2500 persons. It is estimated that there are between 6000 and 7000 rare diseases, a number that continues to increase thanks to our knowledge of understanding of the biology of disease. In most cases, therapeutic options for rare diseases are few at best, thus highlighting the vast need and opportunity to provide new drugs. Rare diseases therapeutics embodies an example of the power of individualized therapies; however the development of orphan drugs, drugs indented to threat rare diseases, is complicated on many facets[1]. Development is challenging when it concerns clinical trial organization, time-consuming, and generally leads to a lower commercial return compared to treatments targeting more common diseases, due to its low prevalence[2]. Just as common drugs, the failing of drugs in clinical trials provides additional complications and costs to the development of drugs[3].

To improve market conditions for the development of orphan drugs, intended to treat rare diseases, the USA approved the Orphan Drug Act in 1983 to provide financial incentives for companies to develop therapies for rare diseases, followed suit by the regulations of the European regulation of Orphan Medical Products in 1999[4,5]. Measures linked to these acts entail implemented tax credits for clinical testing costs, provision of scientific advice by drug regulatory bodies, authorized expedited regulatory review for orphan drugs and a period of market exclusivity. Although these measures have stimulated orphan drug development for rare diseases, they are still only available for a small fraction of all 6000 to 7000 rare diseases patients; around 100 orphan drugs in Europe have achieved market authorization, versus approximately 500 drugs in the USA[5,6].

Data Mining

Data mining, the process of extracting hidden patterns from large amounts of data, is one of the most promising ways to identify potential targets based on existing knowledge. Data mining is sometimes loosely equated to analytics, but is actually only being a subset of it. It is a convergence of several fields of academic research, such as applied mathematics, computer science, artificial intelligence, statistics and machine learning. In the current era of big data, more and more data is already available from numerous sources. The continuous development of data mining and machine learning methods and databases have promise to see patterns and targets that could result in new therapeutic options.

Technological advances in the field of medicine have led to the availability of various data sources, such as literature data, in vitro laboratory data, animal data, structure data and clinical data on drugs. These available data sources are multidimensional and not readily accessible just by simply looking at the output. By connecting these data on drugs, proteins and diseases, data mining methods may enable the discovery of new or the repurposing of previously known pharmaceutical compounds and put them in a context for further exploration, as such showing opening up their potential.

These data mining approaches take advantage of methodologies and tools available from chemogenomics, medicinal chemistry, bioinformatics, chemoinformatics and network and systems biology, aided by various databases. Several open-source databases, a number of which are described in the paragraph about enabling tools for drug repurposing, provide target and drug profiles, associated with related diseases, biological functions and associated signaling pathways, that can be used to find new or existing candidates using the different data mining approaches[7].

Repurposing

Drug repurposing is the process of finding new indications for existing drugs, including abandoned or potential candidate drugs[8]. Abandoned drugs consist of drugs that have not proven efficacy for a specific indication in phase II or III clinical trials but have not shown major safety concerns. It also includes drugs that have been put aside due to commercial reasons[7,9]. Candidate drugs include drugs in clinical development, whose mechanisms of action could be relevant to multiple diseases, or drugs for that are close to losing their incentive for exclusive market authorization[7,9].

Finding a new use for an existing drug could be favorable for many reasons. Drug repurposing starts with compounds for which bioavailability, safety profiles and toxicity are known, and that have proven formulation and manufacturing routes and well-characterized pharmacology. As such this route has as potential to significantly reduce the risks associated with drug discovery, as such potentially allowing drugs to enter clinical phases faster, followed by a potential faster approval by the FDA or EMA, and at a lower cost[10]. In comparison to the number of approvals for De Novo drugs, repurposed drugs are more frequently approved (11% versus 30%) giving companies a significant benefit to invest in finding avenues for drug repurposing[8,11]. Currently drug repurposing plays an important role in drug development and discovery, as it is estimated that around 30% of new drugs approved by the FDA in recent years are repurposed drugs[12]. Therefore it is unlikely that repurposing will find a cure for thousands of rare diseases currently without a treatment from this known repurposed compound tool.

As repurposed drugs could potentially dramatically decrease the cost of drug development, this could more easily lead to treatments for diseases where De Novo drug discovery could not lead to a commercial success on investment, as such opening doors towards rare disease treatments. Repurposing can also focus on highly successful and commercial drugs that have lost their exclusive market incentive. Creating new incentives to repurpose generic drugs can deliver affordable treatments to rare disease, acute disease and neglected disease patient populations. This last advantage is however a double-edged sword, as despite the lowered starting costs, the potential lack of financial return of a repurposed drug without secured market exclusivity might not lead to the same economic returns as new drug development, thus still lacking investors for this repurposing[13].

Other disadvantages are the legal and intellectual property (IP) barrier connected to repurposing. Many of the potential repurposing uses have already been published in scientific literature, or adopted in clinical practice. Although these drugs have not been proven to work through clinical testing, they can no longer be patented, which reduces the chances of marketing the repurposed drug successfully[14].

Avenues of Data Mining for Drug Development

The majority of data is unstructured and can either be textual or numerical. In order to structurise and classify these different kind of data, several data mining techniques are and will be developed. Hereby, the focus is placed not only on the extraction of information; but also on the discovery of knowledge, in which these techniques and algorithms are expected to unearth entirely new facts and relations that were previously not known by human experts. With the rising quantity of biomedical data and information, we might be on the verge of an exciting era of omics drug discovery.

Data mining techniques can be applied to different stages of the drug development process (see Figure 1). Inevitably, data mining approaches will become the first phase of future drug discovery pipelines, by playing different roles in the lead generation and lead optimization process, to help to select the best targets. Data mining has already made way in the application to identify target for therapeutic invention, and with continues development of new techniques is expected to even do more so. Owing to the different limitations of the various techniques, combinatory or integration approaches of different data mining techniques should be able to overcome the individual drawbacks. Data mining could thereby assist researchers and drug developers to make earlier, faster and crucial decisions in the drug development process[15].

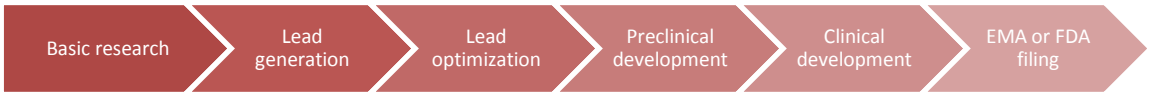

Figure 1: Drug discovery process from target ID and validation through to filing of a compound. Adapted from Hughes *et al*. (2011)[16].

Avenues of Drug Repurposing

The two scientific principles on which drug repurposing is primarily based are:

- The 'promiscuous' nature of the drug; a single drug often interacts with several pathways or targets. Secondary effects of a drugs that are undesirable for one indication could prove to be sought-after in another one[17–19].
- Targets relevant to a particular disease or pathway can also be of vital importance in other biological pathways or phenotypes[20].

Following suit on the several scientific principles behind drug repurposing, there is a range of avenues of drug repositioning. These avenues are differentially divided on either the method of repurposing, e.g. serendipitous, experimental or computational, or based on the starting point of drug repurposing, e.g. drug oriented, disease oriented or treatment oriented.

Most possibilities for drug repurposing so far have arrived through serendipity[21]. For example antiemetic thalidomide, which gained new indications for multiple myeloma and leprosy, has arrived through serendipitous observation[8]. Others have arising from observations, discussions and other collaborations,

such as imatinib, which was first approved for chronic myeloid leukemia, targeting the BRC-Abl fusion protein and was subsequently approved for the treatment of gastrointestinal stromal tumor, due to its potent inhibition of c-KIT. In addition, in clinical trials side effects have been observed, previously not discovered in animal models, which could lead to possibilities for repositioning. A famous example includes sildenafil, which although developed for hypertension, became a blockbuster drug for erectile dysfunction[22].

Other possibilities result from experimental approaches that systematically elucidate new drug-target interactions, identifying drugs with a desired biological activity, for instance by performing high-throughput compound screenings. These fall into several categories. The first is to find direct binding partners of present drugs, such as high-throughput screenings of an approved drug library against protein tyrosine phosphatizes or high-throughput direct-binding assays to test drugs against a panel of kinases, based on the fact that many kinase drugs are multi-targeting[23,24]. A second are cell-based approaches that induce a desired inhibition in cellular phenotype, for example autophagy, apoptosis or proliferation[21,25–27]. Thirdly, gene expression analysis can be exploited to identify drugs that portray similar gene expression profiles in cell lines to other approved drugs[28]. Gene expression analysis can also be used to pick up drugs that will lead cell lines to show an opposite gene expression profile to that of the disease[29].

Nowadays, given the large number of available drugs and the even larger number of diseases, it is feasible to find connections using experimental or computational approaches. More and more computational approaches have been published in recent years, using data mining to specify new roles for existing target proteins or target pathways in different diseases, as such leading to repositioning possibilities. This is especially interesting for rare diseases, as this strategy is supported by the observation that causative genes in many rare diseases share pathways with common disease targets, thus creating opportunities for repurposing[17].

Concepts of Interest for Data Mining/ Repurposing

The following terms outline concepts that could be used for data mining for the identification of new or repurposed targets for drug development of orphan drugs.

- Distinct rare diseases might share similar or identical biological mechanisms

Groups of apparently distinct rare diseases might share similar or identical biological mechanisms[30]. Drug development challenges could therefore be overcome by grouping diseases based on their underlying cause – or etiology –, rather than concentrating on one treatment for one rare disorder at a time; for example the recently-approved ataluren to treat Duchenne muscular dystrophy. The authors indicate that ataluren has also demonstrated efficacy in treating cystic fibrosis, suggesting that several clinically distinct disorders result from common underlying causes and might respond positively to a same drug. The authors believe that adopting this approach for drug development could have multiple benefits: greater industrial interest in providing (repurposed) therapeutic options in rare diseases, larger patient pools to conduct clinical trials, improved understanding of the relationship between disease and drug response, and potential therapeutic benefits for a greater number of patients.

- Normalization

The preprocessing and preparation of data, before the actual data mining step, is essential for the overall knowledge discovery process. One of the first steps in this process is normalization, a commonly used methodology to analyze high-throughput data. This step is vital when dealing with parameters of various scales and units. It also adjusts individual profiles to balance them properly, so that consequential biological comparisons can be made. Commonly used algorithms for this process are linear regression analysis, non-linear regression analysis and lowest normalization[15].

- Unsupervised Clustering

Clustering is useful for exploring data, but can also be used as a preprocessing or preparation of the data before mining. Clustering analysis is set up to identify clusters that are embedded in the data, but when no obvious natural groupings can be seen. A cluster is a collection of data objects that are in one way or another similar to one another. Members of a cluster are therefore more like each other than they are like members of a different cluster. Unsupervised clustering is different from supervised classification, in that the outcome of the process is not guided by known results; therefore there is no target element, therefore, there is no predefined constraint on samples[31]. Examples of algorithms are pair wide clustering, principle component analysis and self-organizing maps[15].

- Supervised Classification

Classification of data us the process of dividing the items that make up the data collection into classes or categories. In the context of data mining, this classification of data is done based on a model that that been created by historical data. The goal of this approach is to precisely predict the categories of each new data find, so that it can be recorded together with the previous data. Supervised classification is started using build data, otherwise known as training data or a training set. The test set or test data is then used to validate the classifiers[31]. Examples of this methodology are linear discriminant analysis, K-nearest neighborhood prediction and trained neural network[15].

- Principal Component Analysis

Principal component analysis (PCA) is a linear algebra technique used to emphasize variation and bring out strong patterns in a dataset. It is a simple, non-parametric method, used to make it possible for data being easily explored and visualized. PCA can provide a way to reduce the complexity of the dataset to a lower dimension, to reveal the simplified dynamics underlying the data. This concept can be used to visualize and cluster the chemical space analysis of predicted drugs and as such is part of several methodologies on a matrix of drugs versus targets[32,33].

- Homopharma

A new concept in the field of drug repurposing is “Homopharma”. This concept is based on the fact that a set of proteins which have the conserved binding environment can be matched with a set of compounds are often able to inhibit these proteins[34]. According to the authors this method can identify potential targets of compounds and reveal key binding environment and thus be instrumental in for discovering new usages for existing drugs. The experimental work of the authors showed that the four flavonoid derivatives, which can be used as anticancer compounds, selected by the authors, was able to inhibit multiple protein-kinases with similar physiochemical properties efficiently. The authors believe that “the Homopharma concept can have the potential for understanding molecular binding mechanisms and providing new clues for drug development”.

- Genome-wide association studies

Over the last couple of years, various genome-wide association studies (GWAS) have been performed, aiming to give insight into the biology of disease, but also intend to lead to concrete opportunities for drug development and repurposing in multiple therapeutic areas[35]. However, the direct influence of GWAS on the launch of clinical trials cannot be determined with certainty, given that the rationale of initiating a drug discovery project is not always clear-cut. With the continuous stream of new GWAS and other Next-Generation Sequencing (NGS) studies, new gene-disease associations may be revealed, with the potential to give rise to additional changes for repurposing[35].

- Interactome Networks

Most phenotype/ genotype relationships are based on complex biological systems. To visualize and give insights in their complexity, different interactome networks are mapped and integrated with each other[36]. Many different interactome networks exist, all given rise to specific types of information. Examples of interactome networks are metabolic networks, protein-protein networks, gene regulatory

networks, cellular interactome networks, transcriptional profiling networks, phenotypic networks or genetic interaction networks. From these interactome network models, global properties can emerge, as well as knowledge about how these properties can relate to human disease or therapeutic options[36].

A related article published in Science presents a network-based framework to identify the location of disease modules within the interactome – a network integrating all interactions within a cell – to understand and predict disease modules relationships[37]. According to Menche *et al*., a complete and accurate map of the interactome, which could have tremendous impact on our ability to understand human disease at a molecular level, is at least a decade away[37]. The authors show that the current data from an "incomplete interactome" may be able to map out some disease module relationships using network science. The authors demonstrate that the "network-based location of each disease module determines its pathobiological relationship to other diseases, where associated disease models segregate in the same neighborhood of the human interactome," whilst unrelated modules form in different neighborhoods. The authors believe that the proposed network-based distance allows us to envisage the relationships between diseases even if they do not share genes. The authors believe that the study is significant as "the introduced network-based framework can be extended to address numerous questions at the forefront of network medicine, from interpreting genome-wide association study data to drug target identification and repurposing."

Models for Data Mining/ Repurposing

With the drug-related data growth and open data initiatives, many data have become for mining. In order to support this mining process, several methodologies have been developed to more efficiently exploit the available literature. These methods can gather evidence supporting the find of new uses or indications of existing drugs. The methodologies are based on different terminologies found in literature that are connected to each other via different models. A number of models suggested for data mining and repurposing are proposed below:

- Graph Theory & Computational logistics

Data mining methodologies have been targeted at making the most of the existing data and knowledge to identify new therapeutic targets and to repurpose drugs. The double-layer literature-based research methodology developed by Gramatica et al has as objective to efficiently exploit natural-language expressed biomedical knowledge to identify possibilities for repurposing[38]. This methodology leverages on developments in Computational Linguistics and Graph Theory, to build a network of knowledge representations, which takes concepts seen in diverse publications and connects them into a graph symbolized with paths and nodes, whenever they co-occur in a given sentence. The information resulting from this graphical representation is then explored and ranked for statistically relevant indirect connections, thereby analyzing and revealing potential hidden relations between any drug and disease. Based on the length of the path between different nodes in this graphical network, a rationale can be suggested to indicate a biological mechanism or mode of action, e.g. it points out a specific biological interaction through which a pharmacological compound produces its pharmacological effect amongst molecular targets. This methodology allows for both the validation of existing rationales and the discovery of new ones, taking the serendipitous component out of the equation. This methodology has provided different proofs of concept, having led to new potential treatment options.

- Text Mining

Available scientific literature represents a rich source of knowledge on links between biomedical concepts such as genes, diseases and cellular processes. Text-mining, one of the methods to establish relationships and retrieve knowledge between biomedical concepts is co-occurrence, which could lead to the discovery of hidden relationships[39]. Based on this coupling of terms, relationships are thought out that could give new insights for drug repurposing. Text mining exists of two major steps, being information retrieval (IR) and information extraction (IE). IR finds literature or abstracts around a topic of interest with the help of either a general search engine, such as Pubmed, Uniprot, HighWire Press, E-Biosci or InterPro, or a specific text mining tool. IE then follows to identify or define the facts or concepts from these resources. This is based on the principle of ABC, in which A and C are not directly related, but connected via B.

- Molecular Docking

Molecular docking is a computational method that calculates how two molecules interact with each other in a three-dimensional space[40]. It is a virtual screening approach in the field of drug discovery, to see how a chemical, whose structure is known form crystallography, is docked against a specific

protein binding site, in order to determine whether the chemical can inhibit the target[41]. The best scoring chemicals or compounds can then be obtained and submitted for experimental testing. Several software for docking approaches are available, however high false positive rates remains a limitation of this methodology, due to limitations such as incomplete binding pocket prediction, inadequate ligand confirmation sampling, inaccurate scoring functions, lack of protein flexibility and lack of solvent molecules during the simulation[42–44]. Developments in this methodology try to take these limitations into account, by developing computational pipelines that can run large-scale cross-docking of compounds to targets, with stringent filtering criteria[44].

- Signature-based methodologies

With the advancement of NGS, GWAS, and the already available microarray data, a large volume of relevant genomics data has become available, including data from gene signatures and Gene Set Enrichment Analyses (GSEA) that can be used to discover unknown mechanisms. Signature-based methodologies make use of these gene signatures derived from disease omics data with or without treatments, to discover unknown disease mechanisms of action of molecules and drugs[45]. The methodology uses a global clustering of drugs and diseases by predicted therapeutic scores, which reveals new and already known therapeutic relationships and also provides the pathophysiological context to support its interpretation[46].

- Pathway- or network-based methods

Pathway or network-based drug-repositioning methods arrange various disease omics data, signaling or metabolic pathways data and protein interaction networks to visualize disease-specific pathways that provide new possibilities for repositioned drugs[47,48]. This methodology can assist in narrowing down large numbers of general signaling networks based on a large sum of proteins to a specific network with limited proteins and targets, as such making it possible to pick up potential candidates.

- Targeted mechanism-based methods

Targeted mechanism-based drug-repositioning methods combine omics data of disease treatments with signaling pathway data and protein information networks, to discover unknown mechanisms of action of drugs[49–51]. This methodology not only attempts to identify new mechanisms of action; but also to identify treatments of drugs to specific diseases. As such, it sets up a computational model to predict drug effects and related targeted pathways.

- Knowledge-based methods

Knowledge-based drug-repositioning methods are methodologies that apply a large spectrum of differential approaches, such as bioinformatics and chemoinformatics approaches to include information of drugs, drug–target networks, signaling and metabolic pathways and chemical structures of targets and drugs[52–56]. It can also include information about clinical trials. Knowledge-based models incorporate all this information, attempting to improve a better accuracy as individual methods.

- Concept Modeling-based Drug Repositioning

Concept modeling-based drug repositioning is based on the Unified Medical Language System (UMLS) concept that uses topic modeling to estimate the probability distribution of subjects for each of the

drugs or diseases and assesses disease-drug similarity[57]. The methodology uses Concept Unique Identifiers (CUI) to map biomedical concepts. The CUI are filtered limiting only those belonging to specific semantic groups. It next calculates the differences between the topic distribution in the selected disease and drug profile, to find drug candidates.

- Integrated data mining

Target discovery for drug discovery is a difficult job, due to the complexity of human diseases and the heterogeneity of various biological data. Therefore, no single data mining approach is sufficient to obtain a full understanding of the cellular mechanisms behind biological networks. To retrieve and prioritize data, therefore the best approach is to integrate and analyze data across the different disciplines, taking into account the pros and cons of the different approaches. A combination or integration of text-mining with high-throughput data, such as genomic, proteomic or chemogenomic data, has been increasingly used as a data mining technique. The emergence of systems biology has given scientists a tool to analyze and visualize datasets in the context of the biological pathway or network of their interest[58–60].

Databases and Tools to Enable Data Mining/ Repurposing

Tools for Data Mining/ Repurposing

Many different methodologies are available for the discovery of possible new orphan drugs. Some of these methodologies have resulted in software, in order to stimulate scientific collaboration and to facilitate data mining. Two of these online softwares are TarFishDock, a tool focused on molecular docking, and DrugNet for network based methodologies. Additionally, there is a large scope of text mining tools available.

- Target Fishing Docking (TarFisDock)

TarFisDock is a web-based tool for molecular docking that automatically searches small molecule-protein interactions over a large repertoire of protein structures. It offers PDTD (potential drug target database), a target database containing 698 protein structures covering 15 therapeutic areas and a reverse ligand-protein docking program. In contrast to conventional ligand-protein docking, reverse ligand-protein docking aims to seek potential protein targets by screening an appropriate protein database. The input file of this web server is the small molecule to be tested, in standard mol2 format; TarFisDock then searches for possible binding proteins for the given small molecule by use of a docking approach. The ligand-protein interaction energy terms of the program DOCK are adopted for ranking the proteins. To test the reliability of the TarFisDock server, we searched the PDTD for putative binding proteins for vitamin E and 4H-tamoxifen. The top 2 and 10% candidates of vitamin E binding proteins identified by TarFisDock respectively cover 30 and 50% of reported targets verified or implicated by experiments; and 30 and 50% of experimentally confirmed targets for 4H-tamoxifen appear amongst the top 2 and 5% of the TarFisDock predicted candidates, respectively. Therefore, TarFisDock may be a useful tool for target identification, mechanism study of old drugs and probes discovered from natural products.

- DrugNet

To aid drug repurposing, an article published in Artificial Intelligence in Medicine has described a novel web tool, DrugNet[61]. The authors "*built a network of interconnected drugs, proteins and diseases and applied DrugNet to different types of tests for drug repositioning*"[62]. Their work is based on the principle that biological entities are intricately networked as well as dynamic and heterogeneous. The web tool can be accessed to query for drug-disease or disease-drug prioritizations, which then returns a list of ranked drugs (active substance, not trade names) based on a given disease or provides a ranked list of diseases (possibly new indications that can be pursued) for a drug query. The authors believe that usage of DrugNet could potentially bring respite for patients with no treatment, especially rare disease patients, sooner as the identified drugs have already been shown to be safe and tolerable.

- Textpresso

This tool allows for literature searches of model organism research, giving access to the full text, so that the entire content of the article can be investigated and capabilities using categories of biological concepts and classes; relating or indentifying different objects[63]. Textpresso also allows for text mining of

biomedical literature for database curation, identifying and extracting biological entities and facts from the full text of research articles. Furthermore, it allows for the linking of biological entities in PDF and online journal articles.

- BioRAT

This tool is a framework to describe IE, focusing on the definition of the template patterns used to convert free text into a structured database[64]. These template patterns are used to identify information of interest, with definitions of words and documents, and other typical IE and text mining tasks, such as stemming and part of speech tagging, as well as IE itself. The framework has allowed us to explicitly identify some of the fundamental issues underlying IE and to formulate possible solutions. The framework allows computationally feasible heuristic search methods to be developed for automatic template creation.

- iHOP

iHOP provides a network of concurring genes and proteins that extends through the literature, touching upon phenotypes, pathologies and gene function, as a natural way of accessing millions of PubMed abstracts. By using genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed can be converted into one navigable resource, bringing all advantages of the internet to scientific literature research.

Details about these software tools can be found in Annex I.

Databases for Data Mining/ Repurposing

Various databases that could assist in the field of data mining/ drug repositioning have been developed, both in the academic, industrial and non-profit sector. The nature of these databases is various, from being set-up to give information about the drug-target relation, as such being off assistance for the docking of molecules, or databases more related to chemical structures. Furthermore, the Collaborative Drug Discovery (CDD) initiative is described here, to facilitate with IP related problems. Additionally, although not included in this list, several databases on pathway information and omics data could be of assistance.

- Orphanet database

Orphanet, the portal of rare diseases and orphan drugs, includes all orphan medicinal products in development and on the market in Europe, whether they have or not the orphan drug status.. It provides also an inventory of rare diseases, fully classified using a multi-hierarchy approach, as well as genes involved in the expression of the diseases.

- Supertarget

Supertarget is a database that integrates drug-related information about medical indication areas, adverse drug effects, drug metabolism, pathways and Gene Ontology (GO) terms of the target proteins that aims to help in providing a better understanding of the molecular basis of drug action. It also allows

for a two-dimensional drug screening and sequence comparison of the targets proteins, with pointers to the respective literature source[65].

- Manually Annotated Targets and Drugs Online Resource (MATADOR)

Matador is a database related to Supertarget, in which part of the drugs that were retrievable in that particular database have been annotated with additional binding information and indirect interactions of compounds and chemicals. The available annotations are based on text-mining and manual curation based on PubMed and Online Mendelian Inheritance in Man (OMIM) entries[65].

- DrugBank

DrugBank is a database that combines information about chemical, physical, pharmaceutical and biological data about drugs and drug targets[66]. It provides a large coverage of data for drugs needed to facilitate drug repurposing. Additionally, it also provides data on small molecules and biotech drugs. The database provides several tools for visualization, querying and search options.

- Potential Drug Target Database: PDTD

PDTD is a protein database for *in silico* target identification[67]. It provides an array of protein data extracted from literature and other data sources, covering information of known and potential drug targets, including protein and active sites structures, related diseases, biological functions and associated signaling pathways. Targets are order by both the classification of disease and the biochemical function.

- Therapeutic Target Database (TTD)

TTD is a database that contains information about targets, targeted disease, pathway and the corresponding drugs directed to each one of these targets[68]. As such it provides target validation information, such as drug potency against the target, effect against disease models and effect of target knockdown, knockout or genetic variations. Also data about previous studies are included, cross-linked to the clinical trial information.

- Promiscuous

Promiscuous is a database of protein-protein and drug-protein interactions aimed at providing a uniform data set for drug repositioning and further analysis, compiled via literature and other data sources via text and data mining, including manual curation[69]. It contains three different types of entities, being drugs, proteins and side-effects as well as relations between them. This network-based approach can provide a starting point for drug-repositioning.

- Disease Manifestation Network (DMN)

DMN is a new phenotype network database, created based on the usage of highly accurate disease-manifestation semantic relationships from UMLS instead of mining on textual descriptions[70,71]. The usage of 50,543 highly accurate disease-manifestation semantic relationships UMLS helped capture major aspects of disease phenotypes which can successfully predict disease causes. DMN not only contained existing knowledge but also some novel insights. DMN partially correlates with the genetic network database - Human Disease Network - based on OMIM and GWAS. It is thought to have the

potential to provide new leads to discover unknown causes of specific diseases, thus concluding that a combinatorial approach where mimMiner and DMN disease is used would be a method for gene discovery and drug repositioning.

- CDD

Research collaborations are necessary in the process of repurposing, in order to speed up research, diminish the financial burden and prevent unnecessary repetition of experiments[72,73]. However, there are several problems with IP to be taken into account when sharing data[74,75]. To facilitate these issues, the CDD created software for researchers for storing chemistry and biological data, which can be securely shared and mined while keeping IP status[76].

A detailed list of the databases is provided in Annex II.

Initiatives to Boost Data Mining/ Repurposing

Besides the many databases that are around to search for possibilities, there is a need to make the compounds accessible for testing to researchers. To address this issue, several initiatives have been formed that created chemical libraries of existing drugs. These initiatives aim to allow interested researchers to access these libraries to test their candidates compounds[7]. In the area of rare and/or neglected diseases, these initiatives include the Center for World Health and Medicine (CWHM) at Saint Louis University, the FDA's Rare Diseases Repurposing (RDRD) initiative, the National Institutes of Health (NIH) collection available through the Therapeutics for Rare and Neglected Diseases (TRND) program and the NIH Chemical Genomics Center (NCGC)' Pharmaceutical Collection (NPC).

- Rare Diseases Repurposing Database (RDRD)

Several databases for drug repurposing have been created in recent years, in order to advance drug repurposing. In order to further speed up the development of new FDA-approved drugs for the treatment rare diseases, the FDA has created the Rare Diseases Repurposing Database (RDRD). While the data is a reconfiguration of already FDA-released information, it offers researchers a useful tool for finding opportunities. The RDRD lists 'high-potential' drugs that have received orphan status designation, yet have not received market authorization for the disease, but have already received this authorization for another disease or condition[77]. The RDRD contains three parts, being orphan-designated products with at least one marketing approval for a common disease indication, orphan-designated products with at least one marketing approval for a rare disease indication and orphan-designated products with marketing approvals for marketing for both rare and common disease indications[77]. This database suggests sponsors chances to build up niche therapies for rare diseases[77].

- Therapeutics for Rare and Neglected Diseases (TRND)

The TRND aims to speed up the development of new and repurposed treatments for rare and neglected diseases. TRND aims to stimulate research collaboration among academic researchers, non-profit organizations and pharmaceutical companies working on rare and neglected diseases[78]. The program is set up to provide expertise and resources, such as a drug database, working with research partners to move therapeutics through pre-clinical testing, including plans for clinical trials and submission of an IND

application to the Food and Drug Administration. These efforts effectively "de-risk" therapeutic candidates and make them more attractive for adoption by outside business partners.

- The Center for World Health Medicine at St Louis University (CWHM)

CWHM has a large collection of compounds available for collaboration with internal and external partners, with as goal to develop and optimize high-throughput target- or phenotype-based assays. The goal of these high-throughput screens is to identify hits that may be useful molecular probes for drug discovery and repurposing programs. Compounds are available in various formats, plates and individual compounds for validation. Lead optimization resources are also available.

- NGCG Pharmaceutical Collection (NPC)

The NCGC has created the NPC a collection of drugs, that is available both as electronic source and as experimental high-throughput screening resource[79]. This collection represents drugs that are registered for use in humans by regulatory agencies worldwide. Data on the activities of the different drugs generated through screening of the NPC are made publically available through PubChem[79].

A detailed list of the initiatives to boost data mining and repurposing is provided in Annex III.

Initiatives for Data Mining/ Repurposing

Several initiatives in the public, private and governmental sector have been developed to support data mining for new and repurposed drugs, in both the European Union (EU) and the USA. These initiatives, in the public, profit and non-profit sector have helped to set data mining and drug repurposing into the spotlight and have resulted in a number of successful collaborations, and mostly, in the availability of various new and repurposed drugs. A number of initiatives that endeavor to support data mining and drug repurposing on several levels are described below.

Public Sector-Driven Initiatives for Data Mining /Repurposing

The success of data mining for new and repurposed drugs is built on the academic and industrial partners and the contractual agreements and strategies enlisted by initiatives to do so. Several initiatives are launched in the USA and Europe in order to stimulate strategies and partner collaborations for data mining and drug repurposing, and to overcome the different expectations regarding IP and publishing. One of the earliest of such initiatives is the Innovative Medicine Initiative that started in 1996, thereby riding the big data wave from the start. In this initiative, EMBL-EBI forms partnerships for data mining with different small and medium-sized enterprises. The projects range from studying drug safety databases to semantic enrichment of scientific literature, thereby providing data infrastructure and services. The strategic focus of the program is the development of resources and services that will benefit both our members as well as our wider stakeholder communities[80].

Another European initiative in the field of data mining is the Drug Discovery Technology Platform (Plataforma Drug Discovery) in Barcelona Science Park. This initiative aims to provide services and collaborate on projects with companies and research bodies for drug discovery. The platform offers technology for drug design and analysis of databases of compounds of therapeutic interest, and conducts competitive intelligence projects[81]. The platform contains chemical compound libraries, virtual and experimental screening tools and databases of biological and chemical interactions for drug discovery. This platform has a number of in-house services applying data mining technology to research and develop new therapeutic entities and thereby touching both the technological resources as well as providing access to the various chemical and biological databases.

In May 2012, NIH' NCATS launched an initiative focused on drug repurposing[82]. The initiative, entitled "Discovering new therapeutic uses for existing molecules," is aimed to foster collaboration between the pharmaceutical industry, academia and non-profit institutes, to repurpose shelved compounds. Eight pharmaceutical companies joined this initiative, which together made 58 compounds available for a pilot program. These drugs have been thoroughly been through the research and development process, including safety testing in humans, and as such provide thus a starting point for researchers to further the clinical development process for repurposing faster[7].

Each partnership between an academic and industrial partner is joined by the NIH forming a three-part union. This pilot program does not only attempt to repurpose compounds, but also provides administrative and legal assistance for the collaborative process between academia and industry[7,82]. The

NIH provides not only funding, but also confidential disclosure agreements, collaborative research agreements and mechanisms for peer review, and oversight of the program. The industry provides the drug and the relevant data, while the academic partner provides new indications for the drug along with disease biology knowledge. The new indication for the drug will belong to the academic centre repurposing the drug, while the IP will remain with the company that made the compound available. Academic partners are allowed to publish their findings and are also allowed to make their IP available for other non-profit partners for research and educational purposes[82].

A similar initiative was preceded by a joined initiative from the UK's Medical Research Council (MRC) and AstraZeneca[83]. In the program, MRC has solicited proposals by academic researchers to study 22 of AstraZeneca's compounds for their effect on human disease mechanisms and for the development of possible new therapies[84]. More than 100 applications where submitted from 37 UK institutions[83]. A selection was made on the basis of the scientific rationale for using the drug, the availability and supply of the compound, the novelty of the study, clinical trial design, and the risks and benefits for patients. The rights to IP generated using the compounds varies from projects, but generally follows the same scheme as the for the NCATS initiative, being that the IP for the new indication will be owned by academic institution whereas the existing rights stay with AstraZeneca[84].

Governmental Initiatives for Data Mining/ Repurposing

In 2010, the President's Council of Advisors on Science and Technology in the USA wrote a report, entitled "Report to the President Realizing the Full Potential of Health Information Technology to Improve Technology to Improve Healthcare for Americans: the Path Forward[85]." In this report, the potential to transform healthcare by sharing health and research data and the use of IT and data mining is discussed. This report contained a number of specific recommendations for both the short and mid-term, among which is the recommendation to develop a strategic plan for rapid action to integrate and align information systems through the government's public health agencies (including FDA, NIH, the USA's Agency for Health Care Research and Quality, and the USA's Centers for Disease Control & Prevention), thereby making data and its structure more accessible for mining.

The FDA houses an enormous repository of clinical data including all the safety, efficacy, and performance information that has been submitted to the agency for new products, as well as an increasing volume of post-market safety surveillance data. If the totally of these data can be integrated and analyzed, this could provide many leads for a better understanding of medical and pharmaceutical knowledge, which could lead to the development of many, many new and repurposed drugs. It would also give insight in different disease parameters that would allow a determination of ineffective products earlier in the development process. To open up this potential of information, stimulated by this 2010 report, the FDA is rebuilding its IT and data analytic capabilities and establishing science enclaves that will allow for the analysis of large, complex datasets while maintaining proprietary data protections and protecting patients' information[86].

A governmental initiative more aimed towards drug repurposing was set up in 2014, when the US House of Representatives introduced the Orphan Product Extensions Now Accelerating Cures and Treatments Act (OPEN ACT). This act grants existing pharmaceutical products an additional six months of marketing exclusivity if a company is able to demonstrate the product is able to treat or prevent a

rare disease or condition[87]. This act as such encourages pharmaceutical companies and organizations to 'repurpose' drugs already in the market by adding a rare disease indication. The focus will be on drugs with market exclusivity and not generic drugs as there is little or no incentive to conduct the additional clinical trials required by the FDA. Modeled on the incentive programs of the Best Pharmaceuticals for Children Act (BPCA), the OPEN Act would make available to drug companies an "Orphan Product Exclusivity Extension," so long as the sponsor company establishes that the therapy is designated to treat a rare disease and obtains a rare disease indication from the FDA on the drug label[88].

Non-profit Initiatives for Repurposing

A number of non-profit foundations have decided to aid the repurposing process, in order to speed up the process of finding drugs for specific diseases. One of such initiatives is CureAccelerator, a platform dedicated to repurposing research for rare and unsolved diseases, supporting both philanthropic and commercial routes for repurposing[89]. This platform aims to connect different groups of people, being:

- Funders, who can post a request for proposal or commit to support a specific project, either alone or with other funders
- Researchers, who can post a new research project proposal in search of funding, respond to a funders request for proposal
- Clinicians, who can, just as researchers, post a new project or respond to an existing project. Additionally they can suggest to patients to participate in projects.
- Patients can post their experience if they have been prescribed an off-label drug

Commercial Initiatives for Data Mining/ Repurposing

Several initiatives for data mining and drug repurposing are ongoing in the private sector. Pfizer, besides being part of the NCATS public initiative for drug repurposing, has launched several initiatives. In 2010, Pfizer has launched collaboration with Washington University School of Medicine (WUSM) in St. Louis, in which it gave scientists access to information regarding more than 500 pharmaceutical compounds for data mining. The compounds are or have been in clinical testing, and thus a large quantity of data is available. To encourage the exchange of ideas, an online portal is available through which scientists will have access. To aid on this exchange, Pfizer's Indications Discovery Unit has developed an online portal through which certain WUSTL investigators will have unprecedented access to information about Pfizer's proprietary compounds, including clinical and preclinical data. The compounds have been extensively studied and their mechanisms of action are well-understood. An advisory committee composed of scientists from both Washington University and Pfizer will evaluate proposals for new research that have been co-written by university and Pfizer researchers.

Another collaboration between Pfizer and Boston's Children Hospital was set up in 2011 to identify potential new treatment for Duchenne's Muscular Dystrophy[90]. This collaborative agreement Pfizer, via its Orphan and Genetic Diseases Research Unit, allowed Boston's Children Hospital access to select proprietary compounds as well relevant data about these compounds. Pfizer also committed internal resources to the project such as medicinal chemistry. Boston's Children Hospital will test the compounds provided by Pfizer in the DMD zebra fish model, with an eye toward identifying candidates for further preclinical development.

Other commercial initiatives include THERAMetrics, which has developed among others the DRR 2.0 platform, for the discovery and development of new pharmaceutical product candidates[91]. This platform is based on a hypothesis generating software tool for drug repurposing and repositioning, based on syntactic parsing and semantic analysis of biomedical literature and on mathematical analysis of the resulting knowledge representation. DRR 2.0 is based on the Graph Theory for drug repurposing. SOM Biotech, a company located in Barcelona Science Park, has developed a proprietary virtual screening platform that is based on an in-silico computerized approach that identifies new biological activities of a given drug[92]. A Cambridge-based start-up, Healx, has set up a patient-driven drug development model. This data model uses a combination of machine learning with advanced 'Omic analytics allows the identification of hidden links between drugs and diseases[93]. This allows identifying novel therapeutic solutions for rare diseases by shortlisting effective drug repositioning candidates. It has allowed for the development of a property library of products that have reached clinical or approved status. These approaches selects drug repositioning candidates with are thought to have a higher probability of success, which was validated in a number of different disease areas. This initiative is joined by several patient groups and charities.

Purpose of the Workshop and Questions to be Debated

IRDiRC aims to stimulate and coordinate basic and clinical research, by promoting links between existing resources, fostering the molecular and clinical characterization of rare diseases and encouraging translational, preclinical and clinical research.

The **purpose** of this workshop is to identify what should and can be done to make the most of data mining possibilities to boost the development of new therapies for rare diseases.

The general objectives are to:

- define how to open up new opportunities of drug development with existing data, via data mining;
- create new models for drug development partnership;
- identify areas where funding opportunities should be opened.

These objectives are aligned with IRDiRC goals to maximize resources and coordinate research efforts in the rare diseases field in order to boost the research and development process to help deliver effective therapies as soon as possible.

Annex 1: List of Online Tools Available for Data Mining/ Repurposing

Target Fishing Docking (TarFisDock)

Goals	TarFisDock is a web-based tool for automating the procedure of searching for small molecule–protein interactions over a large repertoire of protein structures.
Website	http://www.dddc.ac.cn/tarfisdock/
Active years	2006 to present
Relevant contacts	▪ Hualiang Jiang, School of Pharmacy, East China University of Science and Technology, Shanghai, China, hljiang@mail.shcnc.ac.cn
Status	Academic
Funding	TarFisDock was supported by the Special Fund for the Major State Basic Research Project of China (grants 2002CB512802 and 2002CB512807) from Ministry of Science and Technology of China and the National Natural Science Foundation of China (grant 10572033).
Description	TarFisDock is a web-based tool for automating the procedure of searching for small molecule–protein interactions over a large repertoire of protein structures. It offers PDTD (potential drug target database), a target database containing 698 protein structures covering 15 therapeutic areas and a reverse ligand–protein docking program. In contrast to conventional ligand–protein docking, reverse ligand–protein docking aims to seek potential protein targets by screening an appropriate protein database. The input file of this web server is the small molecule to be tested, in standard mol2 format; TarFisDock then searches for possible binding proteins for the given small molecule by use of a docking approach. The ligand–protein interaction energy terms of the program DOCK are adopted for ranking the proteins.
Reference	▪ Honglin Li, Zhenting Gao, Ling Kang, Hailei Zhang, Kun Yang, Kunqian Yu, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Jianhua Shen, Xicheng Wang & Hualiang Jiang. TarFisDock: a web server for identifying drug targets with docking approach. ***Nucl. Acids Res.*** 34, W219-224 (2006). ▪ Zhenting Gao, Honglin Li, Hailei Zhang, Xiaofeng Liu, Ling Kang, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Xicheng Wang & Hualiang Jiang. PDTD: a web-accessible protein database for drug target identification. ***BMC Bioinformatics*** 9:104 (2008).
Relevance for rare diseases	Not specifically

DrugNet for Drug Repurposing

Goals	Assisting drug repositioning processes is drawing a raising interest, since it an lead to a considerable reduction in cost and time in any drug development process. This tool can help to find new drugs to be repositioned.
Website	http://genome2.ugr.es/drugnet/
Active years	2014 to present
Relevant contacts	▪ Blanco, A. Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain, armando@decsai.ugr.es
Status	Academic
Funding	DrugNet has been carried out as part of projects PI-0710-2013 of the Junta de Andalucia, Sevilla and TIN2013-41990-R of DGICT, Madrid. It was also supported by Plan Propio de Investigacion, University of Granada.
Description	Computational drug repositioning can lead to a considerable reduction in cost and time in any drug development process. Recent approaches have addressed the network-based nature of biological information for performing complex prioritization tasks. In this work, we propose a new methodology based on heterogeneous network prioritization that can aid researchers in the drug repositioning process. As such, DrugNet was developed, a new methodology for drug–disease and disease-drug prioritization. The approach is based on a network-based prioritization method called ProphNet which has recently been developed by the same authors. In this work, a network of interconnected drugs, proteins and diseases was build and applied to different types of tests for drug repositioning.
Reference	▪ MARTíNEZ, V., CANO, C., BLANCO, A.. Network-based gene-disease prioritization using PROPHNET. EMBnet.journal, North America, 18, nov. 2012.
Relevance for rare diseases	Not specifically

Textpresso

Goals	Allowing for text mining of the biological literature for database curation, linking of biological entities and literature searches of model organisms.
Website	http://www.textpresso.org/
Active years	2004 to present
Relevant contacts	▪ James Done, Yuling Li, Hans-Michael Muller, Paul Sternberg, California Institute of Technology, textpresso@caltech.edu

Status	Academic
Funding	This work has been supported by a grant (# P41 HG02223) from the National Human Genome Research Institute at the United States National Institutes of Health
Description	This tool allows for literature searches of model organism research, giving access to the full text, so that the entire content of the article can be investigated and capabilities using categories of biological concepts and classes; relating or indentifying different objects. Textpresso also allows for text mining of biomedical literature for database curation, identifying and extracting biological entities and facts from the full text of research articles. Furthermore, it allows for the linking of biological entities in PDF and online journal articles.
Reference	▪ Muller HM, Kenny EE, Sternberg PW (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. 2(11):e309.
Relevance for rare diseases	Not specifically

BioRAT

Goals	A search engine and IE tool for biological research.
Website	http://bioinf.cs.ucl.ac.uk.gate2.inist.fr/?id=754
Active years	2004 to present
Relevant contacts	▪ David Corney, UCL Department of Computer Sciences, d.corney@cs.ucl.ac.uk
Status	Academic
Funding	This work is partly funded by BBSRC grant BB/C507253/1,"Biological Information Extraction for Genome and Superfamily Annotation
Description	BioRAT is a virtual research assistant, which will find research papers for you, read them, and tell you the key facts it finds. Well, that's the theory. It is part of an ongoing research project, and so is NOT a highly-polished, well-supported, fully-documented system that we might dream of. With suitable cajoling, it will let you find and download biological research papers, principally through PubMed. It will let you specify patterns of interesting words / phrases, and use these to "read" papers. It will put the results into a variety of useful formats, including XML and HTML. And it will continue to change, and I hope, improve with time. At the heart of BioRAT is the GATE system from Sheffield University's Natural Language Processing group. This is a great open-source Java tool for performing various text processing tasks.

Reference	▪ Corney, D. P. A., Buxton, B. F., Langdon W.B. and Jones, D. T. (2004) "BioRAT: Extracting Biological Information from Full-length Papers", Bioinformatics, vol. 20(17); pp.3206-13
Relevance for rare diseases	Not specifically

iHOP

Goals	Information Hyperlinked over Proteins
Website	http://www.ihop-net.org/
Active years	2004 to present
Relevant contacts	▪ Robert Hoffman, National Center of Biotechnology, CNB-CSIC, Cantoblanco Madrid M-28049, Spain, hoffmann@ihop-net.org
Status	Academic
Funding	The development of this tool was supported in part by the ORIEL and TEMBLOR EC projects.
Description	A network of genes and proteins extends through the scientific literature, touching on phenotypes, pathologies and gene function. We report the development of an information system that provides this network as a natural way of accessing the more than ten million abstracts in PubMed. By using genes and proteins as hyperlinks between sentences and abstracts, we convert the information in PubMed into one navigable resource and bring all the advantages of the internet to scientific literature investigation. Moreover, this literature network can be superimposed on experimental interaction data (e.g., yeast-two hybrid data from Drosophila melanogaster1 and Caenorhabditis elegans2) to make possible a simultaneous analysis of new and existing knowledge. The network, called Information Hyperlinked over Proteins (iHOP), contains half a million sentences and 30,000 different genes3 from humans, mice, D. melanogaster, C. elegans, zebrafish, Arabidopsis thaliana, yeast and Escherichia coli.
Reference	▪ A Gene Network for Navigating the Literature. Hoffmann, R., Valencia, A. *Nature Genetics* 36, 664 (2004)
Relevance for rare diseases	Not specifically

Annex 2: List of Open-source Databases Available for Data Mining/ Repurposing

The Orphanet database of orphan drugs

Goals	The list of Orphan Drugs in the Orphanet database includes all the substances which have been granted an orphan designation for disease(s) considered as rare in Europe, whether they were further developed to become drugs with marketing authorisation (MA) or not.
Website	http://www.orpha.net/consor/cgi-bin/Drugs_Search.php?lng=EN
Active years	1997 to present
Relevant contacts	▪ Orphanet, INSERM US-14, Paris, France. contact.orphanet@inserm.fr
Status	Public
Funding	INSERM
Description	Orphanet is a database of rare diseases and orphan drugs which was established jointly by the French Ministry of Health and the National Institute of Health and Medical Research (INSERM). It started as a national initiative and progressed into a European project after 2000. The concept was to provide all stakeholders with compiled information on rare diseases through a directory of expert services with the assumption that not only rare diseases were rare, but also experts. All information is freely accessible at the website www.orpha.net and all data are accessible at www.orphadata.org
Reference	▪ Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Hum Mutat. 2012 May;33(5):803-8.
Relevance for rare diseases	Specific for rare diseases

Supertarget: an extensive web resource for analyzing 332828 drug-target interactions.

Goals	Supertarget is a database developed to collect information about drug-target relations. It consists mainly of three different types of entities, being drugs, proteins and side-effects. Additionally, it contains information about pathways and ontologies. These three entities are connected between each other through drug-protein, protein-protein and drug-side-effect relations and include rich annotation about the source, ID's, physical properties, references and more.

Website	http://insilico.charite.de/supertarget
Active years	2007 to present
Relevant contacts	▪ Robert Preissner, Institute for Physiology, Structural Bioinformatics Group, Berlin, Germany (robert.preissner@charite.de)
Status	Academic
Funding	The development of SuperTarget was supported by BMBF (Quantpro), Deutsche Forschungsgemeinschaft (DFG: SFB 449), IRTG Berlin-Boston-Kyoto, Investitionsbank Berlin (IBB) and Deutsche Krebshilfe.
Description	**SuperTarget** integrates drug-related information about medical indication areas, adverse drug effects, drug metabolization, pathways and Gene Ontology terms of the target proteins. An easy-to-use query interface enables the user to pose complex queries, for example to find drugs that target a certain pathway, interacting drugs that are metabolized by the same cytochrome P450 or drugs that target the same protein but are metabolized by different enzymes. Furthermore, it provides tools for 2D drug screening and sequence comparison of the targets. The database contains more than 2500 target proteins, which are annotated with about 7300 relations to 1500 drugs; the vast majority of entries have pointers to the respective literature source.
Reference	▪ Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ, Schneider R, Skoblo R, Russell RB, Bourne PE, Bork P, Preissner R. SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res. 2008 Jan;36(Database issue):D919-22. Epub 2007 Oct 16.
Relevance for rare diseases	Not specifically

MATADOR: Manually Annotated Targets and Drugs Online Resource

Goals	MATADOR is a resource for protein-chemical interactions.
Website	http://matador.embl.de/
Active years	2007 to present
Relevant contacts	▪ Robert Preissner, Institute for Physiology, Structural Bioinformatics Group, Berlin, Germany (robert.preissner@charite.de)
Status	Academic
Funding	The development of MATADOR was supported by BMBF (Quantpro), Deutsche Forschungsgemeinschaft (DFG: SFB 449), IRTG Berlin-Boston-Kyoto, Investitionsbank

	Berlin (IBB) and Deutsche Krebshilfe.
Description	The manually annotated list of direct (binding) and indirect interactions between proteins and chemicals was assembled by automated text-mining followed by manual curation. Each interaction contains links to PubMed abstracts or OMIM entries that were used to deduce the interaction.
Reference	▪ Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ, Schneider R, Skoblo R, Russell RB, Bourne PE, Bork P, Preissner R. SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res. 2008 Jan;36(Database issue):D919-22. Epub 2007 Oct 16.
Relevance for rare diseases	Not specifically

DrugBank: Open Data Drug & Drug Target Database

Goals	The DrugBank database is a bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information.
Website	http://www.drugbank.ca/
Active years	2006 to present
Relevant contacts	▪ David S. Wishart, Department of Computing Science and Department of Biological Sciences, University of Alberta, Edmonton, Canada (david.wishart@ualberta.ca)
Status	Academic
Funding	The development of DrugBank was supported by Genome Prairie, a division of Genome Canada.
Description	DrugBank is a comprehensive, web-accessible database that brings together quantitative chemical, physical, pharmaceutical and biological data about thousands of well-studied drugs and drug targets. DrugBank is primarily focused on providing the kind of detailed molecular data needed to facilitate drug discovery and drug development. This includes physical property data, structure and image files; pharmacological and physiological data about thousands of drug products as well as extensive molecular biological information about their corresponding drug targets. DrugBank is unique, not only in the type of data it provides but also in the level of integration and depth of coverage it achieves. In addition to its extensive small molecule drug coverage, DrugBank is certainly the only public database we are aware of that provides any significant information about the 110+ approved biotech drugs. DrugBank also supports an extensive array of visualizing, querying and search options including a structure similarity search tool and an easy-to-use relational data extraction

	system.
Reference	▪ DrugBank 4.0: shedding new light on drug metabolism.Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS.Nucleic Acids Res. 2014 Jan 1;42(1):D1091-7. ▪ DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS.Nucleic Acids Res. 2011 Jan;39(Database issue):D1035-41. ▪ DrugBank: a knowledgebase for drugs, drug actions and drug targets. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M.Nucleic Acids Res. 2008 Jan;36(Database issue):D901-6. ▪ DrugBank: a comprehensive resource for in silico drug discovery and exploration. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J.Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72.
Relevance for rare diseases	Not specifically

Potential Drug Target Database (PDTD)

Goals	PDTD is a dual function database that associates an informatics database to a structural database of known and potential drug targets
Website	http://www.dddc.ac.cn/pdtd/
Active years	2006 to present
Relevant contacts	▪ Honglin Li, Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China, hlli@mail.shcnc.ac.cn ▪ Xicheng Wang, Department of Engineering Mechanics, State Key Laboratory of Structural Analysis for Industrial Equipment, Dalian University of Technology, Dalian, China, guixum@dlut.edu.cn ▪ Hualiang Jiang, School of Pharmacy, East China University of Science and Technology, Shanghai, China, hljiang@mail.shcnc.ac.cn
Status	Academic
Funding	PDTD was partly supported by the Special Fund for Major State Basic Research Project (grant 2002CB512802), the National Natural Science Foundation of China (grants 20721003 and 10572033), and the 863 Hi-Tech Program of China (grant 2007AA02Z304).
Description	PDTD is a web-accessible protein database for *in silico* target identification. It currently contains >1100 protein entries with 3D structures presented in the Protein Data Bank. The data are extracted from the literatures and several online databases such as TTD, DrugBank and Thomson Pharma. The database covers diverse information of >830 known or potential drug targets, including protein and active sites structures in both

	PDB and mol2 formats, related diseases, biological functions as well as associated regulating (signaling) pathways. Each target is categorized by both nosology and biochemical function. PDTD supports keyword search function, such as PDB ID, target name, and disease name. Data set generated by PDTD can be viewed with the plug-in of molecular visualization tools and also can be downloaded freely. Remarkably, PDTD is specially designed for target identification. In conjunction with TarFisDock, PDTD can be used to identify binding proteins for small molecules. The results can be downloaded in the form of mol2 file with the binding pose of the probe compound and a list of potential binding targets according to their ranking scores.
Reference	▪ Zhenting Gao, Honglin Li, Hailei Zhang, Xiaofeng Liu, Ling Kang, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Xicheng Wang & Hualiang Jiang. PDTD: a web-accessible protein database for drug target identification. ***BMC Bioinformatics*** 9:104 (2008) ▪ Honglin Li, Zhenting Gao, Ling Kang, Hailei Zhang, Kun Yang, Kunqian Yu, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Jianhua Shen, Xicheng Wang & Hualiang Jiang. TarFisDock: a web server for identifying drug targets with docking approach. ***Nucl. Acids Res.*** 34, W219-224 (2006)
Relevance for rare diseases	Not specifically

Therapeutic Target Database (TTD)

Goals	TTD is a database to provide information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs directed at each of these targets.
Website	http://bidd.nus.edu.sg/group/cjttd/
Active years	2002 to present
Relevant contacts	▪ Dr. Chen Yuzong, Department of Computational Science, National University of Singapore, Singapore. (phacyz@nus.edu.sg)
Status	Academic
Funding	TTD was funded by fundamental research funds for the Central Universities of China and the Singapore Academic Research Fund R-148-000-141-750 and R-148-000-141-646.
Description	TTD has been developed to provide comprehensive information about efficacy targets and the corresponding approved, clinical trial and investigative drugs. In addition to the significant increase of data content (from 1894 targets and 5028 drugs to 2025 targets and 17,816 drugs), target validation information was added (drug potency against target, effect against disease models and effect of target knockout, knockdown or genetic variations) for 932 targets, and 841 quantitative structure activity relationship models for active compounds of 228 chemical types against 121 targets. Moreover,

	data from previous drug studies including 3681 multi-target agents against 108 target pairs, 116 drug combinations with their synergistic, additive, antagonistic, potential or reductive mechanisms, 1427 natural product-derived approved, clinical trial and pre-clinical drugs and cross-links to the clinical trial information page in the ClinicalTrials.gov database for 770 clinical trial drugs was added.
Reference	▪ Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, Zhang L, Song Y, Liu XH, Zhang JX, Han BC, Zhang P, Chen YZ. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic Acids Res. 40(D1): D1128-1136, 2012. ▪ Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. Nucleic Acids Res. 30(1):412-415, 2002. ▪ Zhu F, Han BC, Pankaj Kumar, Liu XH, Ma XH, Wei XN, Huang L, Guo YF, Han LY, Zheng CJ, Chen YZ. Update of TTD: Therapeutic Target Database. Nucleic Acids Res. 38(suppl 1):D787-91, 2010.
Relevance for rare diseases	Not specifically

Promiscuous

Goals	An exhaustive resource of protein-protein and drug-protein interactions with the aim of providing a uniform data set for drug repositioning and further analysis. The database contains three different types of entities: drugs, proteins and side-effects as well as relations between them.
Website	http://bioinformatics.charite.de/promiscuous/
Active years	2011 to present
Relevant contacts	▪ Joachim von Eichborn, Institute of Physiology, Charité Universitatsmedicin, Berlin, joachim.eichborn@charite.de ▪ Robert Preissner, Experimental and Clinical Research Center (ECRC), Charité Universitatsmedicin, Berlin, robert.preissner@charite.de
Status	Academic
Funding	Promiscuous was funded by the International Research Training Group on Genomics and Systems Biology of Molecular Networks (GRK1360); German Federal Ministry of Education and Research (MedSys); European Commission (SynSys).
Description	PROMISCUOUS delivers complex relations among drugs, their respective targets and side-effects of the drugs. For each entity detailed information is given. To enable the user to explore and handle the data in a scientific yet intuitive way, we developed a novel interface that offers a "natural" way of exploring the network. Here database entities (drugs, targets and side effects) are represented as nodes in a network with edges, which represent the relations between them.

Reference	▪ PROMISCUOUS: a database for network-based drug-repositioning. von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. Nucleic Acids Res. 2011 Jan;39(Database issue):D1060-6. Epub 2010 Nov 10
Relevance for rare diseases	Not specifically

Disease Manifestation Network (DMN)

Goals	DMN is a network build from 50,543 highly accurate disease-manifestation semantic relationships in UMLS.
Website	http://nlp.case.edu/public/data/DMN/
Active years	2015 to present
Relevant contacts	▪ Rong Xu, Division of Medical Informatics, School of Medicine, Case Western Reserve University, Cleveland, United States, rxx@case.edu
Status	Academic
Funding	The development of DMN was funded by Case Western Reserve University/Cleveland Clinic CTSA Grant (UL1TR000439) and partially supported by US National Science Foundation IIS-1162374 and IIS-1218036.
Description	In this study, we built a large-scale DMN from 50,543 highly accurate disease-manifestation semantic relationships in the UMLS. Our new phenotype network contains 2305 nodes and 373,527 weighted edges to represent the disease phenotypic similarities. We first compared DMN with the networks representing genetic relationships among diseases, and demonstrated that the phenotype clustering in DMN reflects common disease genetics. Then we compared DMN with a widely-used disease phenotype network in previous gene discovery studies, called mimMiner, which was extracted from the textual descriptions in OMIM. We demonstrated that DMN contains different knowledge from the existing phenotype data source. Finally, a case study on Marfan syndrome further proved that DMN contains useful information and can provide leads to discover unknown disease causes. Integrating DMN in systems approaches with mimMiner and other data offers the opportunities to predict novel disease genetics.
Reference	▪ Chen Y, Zhang X, Zhang GQ, Xu R. (2015). Comparative analysis of a novel disease phenotype network based on clinical manifestations. J Biomed Inform. 2015 Feb;53:113-20.
Relevance for rare diseases	Not specifically

Collaborative Drug Discovery (CDD)

Goals	CDD Vault is a hosted biological and chemical database that securely manages your private and external data.
Website	https://www.collaborativedrug.com/
Active years	2004 to present
Relevant contacts	▪ Collaborative Drug Discovery, Inc., Burlingame, USA, info@collaborativedrug.com
Status	Company
Funding	
Description	Research collaborations are seen as important for drug discovery to speed up biomedical research, reduce costs, and prevent unnecessary repetition of experiments. There are however considerable IP concerns to be overcome when sharing data. Increasingly, pharmaceutical companies are involved in multi-organization collaborations and public-private partnerships. To address these issues, CDD created a software which enables researchers to have their own private vault for storing chemistry and biology data, which can be securely shared and mined while maintaining IP status.
Reference	▪ Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. Hohman M, Gregory K, Chibale K, Smith PJ, Ekins S, Bunin B. Drug Discov Today. 2009 Mar;14(5-6):261-70.
Relevance for rare diseases	Not specifically

Annex 3: List of Initiatives to Boost Data Mining/ Repurposing

Rare Disease Repurposing Database (RDRD)

Goals	The U.S. Food and Drug Administration's Office of Orphan Products Development (OOPD) has established a valuable resource for drug developers----a database of products that: ▪ have received orphan status designation (i.e. they've been found "promising" for treating a rare disease) AND; ▪ are already market-approved for the treatment of some other diseases up through June 2010
Website	http://www.fda.gov/ForIndustry/DevelopingProductsforRareDiseasesConditions/HowtoapplyforOrphanProductDesignation/ucm216147.htm
Active years	2010 to present
Relevant contacts	▪ Timothé R Coté, US Food and Drug Administration, Silver Springs, USA, timothy.cote@fda.hs.gov
Status	FDA
Funding	
Description	While the data included in the RDRD is a re-configuration/cross-indexing of already FDA-released information, it offers sponsors a useful tool for finding special opportunities to develop niche therapies that are already well-advanced through development. For example, these drugs have already been subjected to pre-clinical (e.g., pharmacokinetic and toxicologic) testing and are already deemed to be pharmacologically active, effective and safe in some clinical context. The opportunities tabulated in the RDRD represent a far "easier lift" to drug developers than beginning with an untested new therapy compound. The RDRD has three sections: ▪ Orphan-designated products with at least one marketing approval for a common disease indication ▪ Orphan-designated products with at least one marketing approval for a rare disease indication ▪ Orphan-designated products with marketing approvals for both common and rare disease indication
Reference	▪ Xu, K. And Coté, T.R. (2010). Database identifies FDA-approved drugs with potential to be repurposed for treatment of orphan diseases, Briefings in Bioinformatics, Jul;12(4):341-5.

Relevance for rare diseases	Specific for rare diseases

Therapeutics for Rare and Neglected diseases (TRND)

Goals	The TRND program supports pre-clinical development of therapeutic candidates intended to treat rare or neglected disorders, with the goal of enabling an Investigational New Drug application.
Website	http://www.ncats.nih.gov.gate2.inist.fr/trnd
Active years	2012 to present
Relevant contacts	▪ trnd@mail.nih.gov
Status	NIH-NCATS
Funding	NIH
Description	NCATS' TRND program provides collaborators with access to significant in kind resources and expertise to develop new therapeutics for rare and neglected diseases. No monetary funds are awarded. Academic, nonprofit foundation, industry and other government agency representatives from within and outside of the United States are eligible for TRND program support. In general, pre-clinical expertise and regulatory resources are available to support the development of promising, well-validated therapeutic candidates from as early as lead compound optimization through submission of the IND application
Reference	▪ McKew JC, Pilon AM. (2013). NIH TRND program: successes in preclinical therapeutic development. Trends Pharmacol Sci. 34(2):87-9.
Relevance for rare diseases	Specific for rare and neglected diseases

Center for World Health Medicine at St Louis University (CWHM)

Goals	The CWHM program supports pre-clinical development of therapeutic candidates intended to treat rare or neglected disorders, by providing compounds for high-throughput screenings for collaborations.
Website	http://www.cwhm.org/index.php?page=high-throughput-assay-development-and-screening
Active years	2012 to present
Relevant contacts	▪ CWHM, Seattle, USA, info@bvgh.org

Status	NIH-NCATS
Funding	NIH
Description	The CWHM at Saint Louis University, USA is a not-for-profit research center dedicated to the discovery and development of medicines to treat rare and neglected diseases. The CWHM consists of a multidisciplinary team of former pharmaceutical company scientists possessing the necessary skill sets required for small molecule drug discovery. The CWHM will consider collaborative proposals to develop and optimize target-based or phenotypic assays to identify compounds as useful probes for neglected disease programs. Partners may request screening of available compound collections or provide their own. Lead optimization resources are also available.
Reference	▪
Relevance for rare diseases	Specific for rare and neglected diseases

The NCGC Pharmaceutical Collection (NPC)

Goals	NPC is a comprehensive, publically-accessible collection of approved and investigational drugs for high-throughput screening that provides a valuable resource for both validating new models of disease and better understanding the molecular basis of disease pathology and intervention
Website	https://tripod.nih.gov/npc/
Active years	2011 to present
Relevant contacts	▪ Ajit Jadhav, NIH Chemical Genomics Center, National Institutes of Health, Bethesda, USA, ajadhav@mail.nih.gov
Status	Non-Profit
Funding	The development of NP was supported by the Intramural Program of the National Human Genome Research Institute, National Institutes of Health.
Description	NPC is a comprehensive, publically-accessible collection of approved and investigational drugs for high-throughput screening that provides a valuable resource for both validating new models of disease and better understanding the molecular basis of disease pathology and intervention. The NPC has already generated several useful probes for studying a diverse cross section of biology, including novel targets and pathways. NCGC provides access to its set of approved drugs and bioactives through the TRND program and as part of the compound collection for the Tox21 initiative, a collaborative effort for toxicity screening among several government agencies including the US Environmental Protection Agency, the National Toxicology Program NTP, the US' FDA and the NCGC. Of the nearly 2750 small molecular entities that have been approved for clinical use by US (FDA), EU (EMA), Japanese (National Health Insurance),

	and Canadian (Health Canada) authorities and that are amenable to HTS screening, we currently possess 2,400 as part of our screening collection.
Reference	▪ R. Huang, N. Southall, Y. Wang, A. Yasgar, P. Shinn, A. Jadhav, D.-T. Nguyen, C. P. Austin, The NCGC Pharmaceutical Collection: A Comprehensive Resource of Clinically Approved Drugs Enabling Repurposing and Chemical Genomics. *Sci. Transl. Med.* **3**, 80ps16 (2011).
Relevance for rare diseases	Not specifically

Bibliography

1. Wood, J., Sames, L., Moore, A. & Ekins, S. Multifaceted roles of ultra-rare and rare disease patients/parents in drug discovery. *Drug Discov. Today* **18,** 1043–1051 (2013).
2. Coté, T. R., Xu, K. & Pariser, A. R. Accelerating orphan drug development. *Nat. Rev. Drug Discov.* **9,** 901–902 (2010).
3. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3,** 711–715 (2004).
4. USA Public Law 97-414 'Orphan Drug Act'. (1983). at <https://history.nih.gov/research/downloads/PL97-414.pdf>
5. Regulation (EC) No 141/2000 of the European Parliament and of the Council on Orphan Medical Products. (1999). at <http://ec.europa.eu/health/files/eudralex/vol-1/reg_2000_141_cons-2009-07/reg_2000_141_cons-2009-07_en.pdf>
6. European Medicines Agency - Find medicine - European public assessment reports. at <http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/landing/epar_search.jsp&mid=WC0b01ac058001d124>
7. Allarakhia, M. Open-source approaches for the repurposing of existing or failed candidate drugs: learning from and applying the lessons across diseases. *Drug Des. Devel. Ther.* **7,** 753–766 (2013).
8. Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3,** 673–683 (2004).
9. Padhy, B. M. & Gupta, Y. K. Drug repositioning: re-investigating existing drugs for new therapeutic indications. *J. Postgrad. Med.* **57,** 153–160 (2011).
10. Boguski, M. S., Mandl, K. D. & Sukhatme, V. P. Drug discovery. Repurposing with a difference. *Science* **324,** 1394–1395 (2009).
11. Kaiser, J. Biomedicine. NIH's secondhand shop for tried-and-tested drugs. *Science* **332,** 1492 (2011).
12. Jin, G. & Wong, S. T. C. Towards better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discov. Today* **19,** 637–644 (2014).
13. Shineman, D. W. *et al.* Overcoming obstacles to repurposing for neurodegenerative disease. *Ann. Clin. Transl. Neurol.* **1,** 512–518 (2014).
14. Smith, R. B. Repositioned drugs: integrating intellectual property and regulatory strategies. *Today Ther. Strateg.* **8,** 131–137 (2011).
15. Yang, Y., Adelstein, S. J. & Kassis, A. I. Target discovery from data mining approaches. *Drug Discov. Today* **14,** 147–154 (2009).
16. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **162,** 1239–1249 (2011).
17. Sardana, D. *et al.* Drug repositioning for orphan diseases. *Brief. Bioinform.* **12,** 346–356 (2011).
18. Boran, A. D. W. & Iyengar, R. Systems approaches to polypharmacology and drug discovery. *Curr. Opin. Drug Discov. Devel.* **13,** 297–309 (2010).
19. Metz, J. T. & Hajduk, P. J. Rational approaches to targeted polypharmacology: creating and navigating protein-ligand interaction networks. *Curr. Opin. Chem. Biol.* **14,** 498–504 (2010).
20. Pujol, A., Mosca, R., Farrés, J. & Aloy, P. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci.* **31,** 115–123 (2010).
21. Li, Y. Y. & Jones, S. J. Drug repositioning for personalized medicine. *Genome Med.* **4,** 27 (2012).
22. Bradley, D. Why big pharma needs to learn the three 'R's. *Nat. Rev. Drug Discov.* **4,** 446 (2005).
23. Zeder-Lutz, G. *et al.* Validation of surface plasmon resonance screening of a diverse chemical library for the discovery of protein tyrosine phosphatase 1b binders. *Anal. Biochem.* **421,** 417–427 (2012).
24. Karaman, M. W. *et al.* A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **26,** 127–132 (2008).
25. Zhang, L. *et al.* Small molecule regulators of autophagy identified by an image-based high-throughput screen. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 19023–19028 (2007).
26. Antczak, C. *et al.* Revisiting old drugs as novel agents for retinoblastoma: in vitro and in vivo antitumor activity of cardenolides. *Invest. Ophthalmol. Vis. Sci.* **50,** 3065–3073 (2009).
27. Iljin, K. *et al.* High-throughput cell-based screening of 4910 known drugs and drug-like small molecules identifies disulfiram as an inhibitor of prostate cancer cell growth. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **15,** 6070–6078 (2009).
28. Iorio, F. *et al.* Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 14621–14626 (2010).
29. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935 (2006).
30. Brooks, P. J., Tagle, D. A. & Groft, S. Expanding rare disease drug trials based on shared molecular etiology. *Nat. Biotechnol.* **32,** 515–518 (2014).
31. Introduction to Oracle Data Mining. at <http://docs.oracle.com/html/B14339_01/1intro.htm#sthref10>
32. Neves, B. J., Braga, R. C., Bezerra, J. C. B., Cravo, P. V. L. & Andrade, C. H. In Silico Repositioning-Chemogenomics Strategy Identifies New Drugs with Potential Activity against Multiple Life Stages of Schistosoma mansoni. *PLoS Negl. Trop. Dis.* **9,** (2015).
33. Lauria, A., Tutone, M., Barone, G. & Almerico, A. M. Multivariate analysis in the identification of biological targets for designed molecular structures: the BIOTA protocol. *Eur. J. Med. Chem.* **75,** 106–110 (2014).
34. Chiu, Y.-Y. *et al.* Homopharma: a new concept for exploring the molecular binding mechanisms and drug repurposing. *BMC Genomics* **15 Suppl 9,** S8 (2014).
35. Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* **30,** 317–320 (2012).
36. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome Networks and Human Disease. *Cell* **144,** 986–998 (2011).
37. Menche, J. *et al.* Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347,** 1257601 (2015).
38. Gramatica, R. *et al.* Graph theory enables drug repurposing--how a mathematical model can drive the discovery of hidden mechanisms of action. *PloS One* **9,** e84912 (2014).
39. Frijters, R. *et al.* Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases. *PLoS Comput Biol* **6,** e1000943 (2010).
40. Shoichet, B. K., McGovern, S. L., Wei, B. & Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **6,** 439–446 (2002).
41. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **303,** 1813–1818 (2004).
42. Taylor, R. D., Jewsbury, P. J. & Essex, J. W. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* **16,** 151–166 (2002).
43. Abagyan, R. & Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **5,** 375–382 (2001).

44. Li, Y. Y., An, J. & Jones, S. J. M. A computational approach to finding novel targets for existing drugs. *PLoS Comput. Biol.* **7,** e1002139 (2011).
45. Lussier, Y. A. & Chen, J. L. The Emergence of Genome-Based Drug Repositioning. *Sci. Transl. Med.* **3,** 96ps35 (2011).
46. Sirota, M. *et al.* Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* **3,** 96ra77 (2011).
47. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934 (2001).
48. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34,** 166–176 (2003).
49. Jin, G. *et al.* A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy. *Cancer Res.* **72,** 33–44 (2012).
50. Jin, G., Zhao, H., Zhou, X. & Wong, S. T. C. An enhanced Petri-net model to predict synergistic effects of pairwise drug combinations from gene microarray data. *Bioinforma. Oxf. Engl.* **27,** i310–316 (2011).
51. Iskar, M. *et al.* Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol. Syst. Biol.* **9,** 662 (2013).
52. Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L. & Vidal, M. Drug-target network. *Nat. Biotechnol.* **25,** 1119–1126 (2007).
53. Zhao, S. & Li, S. Network-based relating pharmacological and genomic spaces for drug target identification. *PloS One* **5,** e11764 (2010).
54. Cheng, F. *et al.* Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **8,** e1002503 (2012).
55. Alaimo, S., Pulvirenti, A., Giugno, R. & Ferro, A. Drug-target interaction prediction through domain-tuned network-based inference. *Bioinforma. Oxf. Engl.* **29,** 2004–2008 (2013).
56. Kinnings, S. L. *et al.* Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **5,** e1000423 (2009).
57. Patchala, J. & Jegga, A. G. Concept Modeling-based Drug Repositioning. *AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci.* **2015,** 222–226 (2015).
58. Natarajan, J. *et al.* Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. *BMC Bioinformatics* **7,** 373 (2006).
59. Li, S., Wu, L. & Zhang, Z. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. *Bioinforma. Oxf. Engl.* **22,** 2143–2150 (2006).
60. de Chassey, B. *et al.* Hepatitis C virus infection protein network. *Mol. Syst. Biol.* **4,** 230 (2008).
61. DrugNet: Drugs repositioning tool. at <http://genome2.ugr.es/drugnet/>
62. Martínez, V., Navarro, C., Cano, C., Fajardo, W. & Blanco, A. DrugNet: network-based drug-disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.* **63,** 41–49 (2015).
63. Textpresso: an ontology-based information retrieval and extraction system for biological literature. - PubMed - NCBI. at <http://www.ncbi.nlm.nih.gov.gate2.inist.fr/pubmed/?term=Textpresso%3A+an+ontology-based+information+retrieval+and+extraction+system+for+biological+literature+Muller+HM%2C+Kenny+EE%2C+Sternberg+PW>
64. Corney, D. P. A., Buxton, B. H., Langdon, W. B. & Jones, D. T. BioRAT: Extracting Biological Information from Full-length Papers. *Bioinformatics* **20,** 3206–13 (2004).
65. Günther, S. *et al.* SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* **36,** D919–D922 (2008).
66. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34,** D668–672 (2006).
67. Gao, Z. *et al.* PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics* **9,** 104 (2008).
68. Chen, X., Ji, Z. L. & Chen, Y. Z. TTD: Therapeutic Target Database. *Nucleic Acids Res.* **30,** 412–415 (2002).
69. Eichborn, J. von *et al.* PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Res.* **39,** D1060–D1066 (2011).
70. Chen, Y., Zhang, X., Zhang, G.-Q. & Xu, R. Comparative analysis of a novel disease phenotype network based on clinical manifestations. *J. Biomed. Inform.* **53,** 113–120 (2015).
71. Disease Manifestation Network: Index of Public Data. at <http://nlp.case.edu/public/data/DMN/>
72. Litterman, N. K., Rhee, M., Swinney, D. C. & Ekins, S. Collaboration for rare disease drug discovery research. *F1000Research* **3,** 261 (2014).
73. Ekins, S., Williams, A. J. & Hupcey, M. A. Z. in *Collaborative Computational Technologies for Biomedical Research* (eds. Ekins, S., Hupcey, ggie A. Z. & Williams, A. J.) 201–208 (John Wiley & Sons, Inc., 2011). at <http://onlinelibrary.wiley.com/doi/10.1002/9781118026038.ch13/summary>
74. Ekins, S. & Bunin, B. A. The Collaborative Drug Discovery (CDD) database. *Methods Mol. Biol. Clifton NJ* **993,** 139–154 (2013).
75. Ekins, S., Hohman, M. M. & Bunin, B. A. in *Collaborative Computational Technologies for Biomedical Research* (eds. Ekins, S., Hupcey, ggie A. Z. & Williams, A. J.) 335–361 (John Wiley & Sons, Inc., 2011). at <http://onlinelibrary.wiley.com/doi/10.1002/9781118026038.ch21/summary>
76. Hohman, M. *et al.* Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discov. Today* **14,** 261–270 (2009).
77. Xu, K. & Coté, T. R. Database identifies FDA-approved drugs with potential to be repurposed for treatment of orphan diseases. *Brief. Bioinform.* **12,** 341–345 (2011).
78. McKew, J. C. & Pilon, A. M. NIH TRND program: successes in preclinical therapeutic development. *Trends Pharmacol. Sci.* **34,** 87–89 (2013).
79. Huang, R. *et al.* The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci. Transl. Med.* **3,** 80ps16 (2011).
80. The EMBL-EBI Industry Programme. at <http://www.ebi.ac.uk/industry>
81. Technology Platform Drug Discovery. at <http://www.pcb.ub.edu/portal/documents/10181/317054/fitxa_pt_drugdiscovery_EN.pdf/befb4812-6b1a-4d67-8d48-793c69077ea3>
82. Allison, M. NCATS launches drug repurposing program. *Nat. Biotechnol.* **30,** 571–572 (2012).
83. Philippidis, A. NCATS, MRC Take Aim at Teaching Old Drugs New Tricks. *GEN* at <http://www.genengnews.com/insight-and-intelligence/ncats-mrc-take-aim-at-teaching-old-drugs-new-tricks/77899618/>
84. Medical Research Council, M. R. C. Alzheimer's, cancer and rare disease research to benefit from landmark MRC-AstraZeneca compound collaboration. (2014). at <http://www.mrc.ac.uk/news/news/alzheimere28099s-cancer-and-rare-disease-research-to-benefit-from-landmark-mrc-astrazeneca-compound-collaboration/>
85. Report to the President Realizing the Full Potential of Health Information Technology to Improve Technology to Improve Healthcare for Americans: the Path Forward. (2010). at <https://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf>

86. Reports - Harnessing the Potential of Data Mining and Information Sharing. at <http://www.fda.gov/AboutFDA/ReportsManualsForms/Reports/ucm274442.htm>
87. Text of H.R. 5750 (113th): Orphan Product Extensions Now Accelerating Cures and Treatments Act of 2014 (Introduced version). at <https://www.govtrack.us/congress/bills/113/hr5750/text>
88. Food and Drug Administration Amendments Act of 2007. (2007). at <http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DevelopmentResources/UCM049870.pdf>
89. Bloom, B. CureAccelerator: Discovering Treatments through Repurposing Research. (2015). at <http://cureaccelerator.org/>
90. Children's Hospital Boston and Pfizer enter into novel collaborative research agreement for muscular dystrophy therapeutics. *http://www.pfizer.com/* (2011). at <http://www.pfizer.com/sites/default/files/partnering/recent_partnership/childrens_hospital_boston.pdf>
91. Therametrics - Welcome. at <http://www.therametrics.com/>
92. SOM BIOTECH, The drug repositioning company. at <http://www.sombiotech.com/>
93. Healx – drug repurposing for rare diseases. at <http://healx.io/>