Creating a screening set of potential anthelmintic compounds using ChEMBL

CURRENT STATUS: POSTED

Matthew Berriman
Berriman Lab Group, Sanger Institute
mb4@sanger.ac.uk*Corresponding Author*

Avril Coghlan
Berriman Lab Group, Sanger Institute

Prudence Mutowo
European Bioinformatics Institute

Noel O'Boyle
NextMove Software

Jane Lomax
Berriman Lab Group, Sanger Institute

Andrew R. Leach
European Bioinformatics Institute

DOI:
10.1038/protex.2018.053

SUBJECT AREAS
Computational biology and bioinformatics *Biotechnology*

KEYWORDS
drug discovery, drugs, anthemintic, helminths, nematodes, flatworms, parasites, parasitic worms, platyhelminths, Nematoda, Platyhelminthes, compounds, anthelminthic, ChEMBL

Abstract

Many existing anthelmintic compounds are compromised by low efficacy, serious side-effects and/or rising resistance in parasite populations. Thus, to reveal potential new drug targets and drugs, we describe a protocol to identify the most promising nematode/flatworm targets, and compounds likely to interact with them from the ChEMBL database (which includes both targets of known drugs and of other biologically active compounds). This provides a potential screening set of drug-like compounds.

Introduction

Many existing anthelmintic compounds are compromised by low efficacy, serious side-effects and/or rising resistance in parasite populations. Thus, to reveal potential new anthelmintic drug targets and drugs, we describe a protocol to identify the most promising nematode/flatworm targets, and compounds likely to interact with them from the ChEMBL database (Gaulton *et al.* 2017), includes both targets of known drugs and of other biologically active compounds. Thus, this provides a potential screening set of drug-like compounds that can be tested for anthelmintic activity.

Our protocol (Figure 1) starts by using predicted proteins from nematode and platyhelminth genomes to search using BLASTP (E ≤ 1e-10) (Altschul *et al.* 1997) against single-protein targets from ChEMBL. To remove redundant hits, the nematode and platyhelminth proteins with significant BLAST hits are collapsed by gene family using an in-house database of gene families (e.g. built the Ensembl Compara pipeline; Vilella *et al.* 2009). Then, a ‘target score’ is calculated for each worm gene, taking into account similarity to known drug targets; lack of human homologues; and whether *C. elegans* and *D. melanogaster* homologues have lethal phenotypes, amongst other factors.

Figure 1. Flowchart of the methods used for identifying potential new anthelmintic targets and drugs using ChEMBL. The following images were used from Openclipart: ‘prescription-bottle-and-pills’ (by algotruneman), ‘famous-and-infamous-molecules-13’ (by Firkin) and ‘worm-gusano’ (by ainara14).

Next, ChEMBL is used to identify hundreds of thousands of compounds with activities against ChEMBL targets to which worm proteins had BLAST matches. To calculate ‘compound scores’ for those

compounds, we prioritise compounds in high clinical development phases, oral/topical administration, crystal structures, properties consistent with oral drugs, and lacking toxicity. At this stage, we focus on the compounds for the top 15% of highest-scoring worm targets, as these are the most promising novel targets. The set of compounds is filtered by selecting compounds that co-appear in a PDBe (Velankar *et al.*) structure with the ChEMBL target; or have high median pChEMBL score, reflecting high potency/affinity for the ChEMBL target.

Lastly, the candidates are placed into chemical classes, based on molecular fingerprints. Then they are further filtered by (i) checking availability for purchase in ZINC 15 (Sterling *et al.* 2015); and (ii) for each worm target, taking the highest-scoring compound from each chemical class.

This gives a relatively small diverse screening set of a few thousand compounds.

Many previous analyses of nematode and platyhelminth genomes have identified potential novel drug targets among their predicted proteins, and prioritised these based on factors such as lethal knockout phenotypes. However, relatively few previous studies have identified compounds predicted to target those nematode/platyhelminth proteins, and those that have done so (e.g. Berriman *et al.* 2009) have generally been limited to small numbers of compounds. By analysing the huge wealth of data in the ChEMBL database, this protocol produces much larger sets of compounds to screen for anthelmintic activity.

Reagents

ChEMBL database website

PDBe database website

ZINC15 database website

Ensembl Compara pipeline

European Nucleotide Archive website

ArrayExpress website

PubChem database website

KEGG Drugs database website

DrugBank database website

ChEBI database website

WormBase database website

FlyBase database website

Equipment

Computer cluster.

Procedure

Step A: identifying potential nematode/platyhelminth drug targets

The inputs for Step A are gene predictions for a set of nematode/platyhelminth genome assemblies (as well as the corresponding predicted proteins), and the protein sequences for single-protein targets from the ChEMBL database (Gaulton *et al.* 2017).

1. To identify potential drug targets in ChEMBL, use BLASTP (Altschul *et al.* 1997) to search all predicted nematode/platyhelminth proteins against the sequences of single-protein targets from the ChEMBL database.
2. Take the top BLASTP hit (with E-value ≤ 1e-10) for each putative nematode/platyhelminth gene (including isoforms of the same nematode/platyhelminth gene).

The output for Step A is a set of nematode/platyhelminth genes with BLASTP hits to single-protein ChEMBL targets.

Step B: ranking potential nematode/platyhelminth drug targets

The input for Step B is set of nematode/platyhelminth genes with BLASTP hits to single-protein ChEMBL targets (output from Step A); a database of gene families containing all the genes from all those nematode/platyhelminth species and also human genes (e.g. built using the Compara pipeline, Vilella *et al.* 2009); and a set of chokepoint enzymes predicted in the nematode/platyhelminth genomes (e.g. using the approach of Tyagi *et al.* 2018).

In Step B, to rank potential nematode/platyhelminth targets, target properties are considered that would be most attractive in a potential new drug target, including:

• Similarity to a known drug target in any species;

• Lack of a human homologue, to avoid toxicity issues in humans;

• Whether *C. elegans* or *D. melanogaster* homologues have lethal or sterile phenotypes when disrupted.

Thus, each of the nematode/platyhelminth genes (output from Step A) is assigned an overall target score, by following these steps:

1. A *BLAST score* is set to 1 for high sequence similarity hits to ChEMBL (E ≤e-85, and ≥80% of the residues in the ChEMBL protein covered by the BLAST match), otherwise it is set to zero. Typically, ~15% of the nematode/platyhelminth genes have BLAST score of 1.
2. A *ChEMBL non-human score* of 1 is applied to nematode/platyhelminth proteins that only match non-human targets within ChEMBL, because for these targets, developing drug selectivity to avoid toxicity issues in humans should be easier. The human proteins in the in-house Compara database are searched against ChEMBL 21 proteins using BLASTP. Where the top ChEMBL BLAST hit of a helminth gene has itself a match with human (with E ≤0.05) or belongs to a Compara family containing a human member, it is assigned a score of zero; otherwise 1. Typically, ~0.3% of nematode/platyhelminth genes have a ChEMBL non-human score of 1.
3. A *phenotype score* is calculated. Large-scale mutant phenotype data are not available for parasitic worms so essentiality is inferred from model organisms. A list of *C. elegans* genes with 'lethal', 'sterile', 'L3 arrest', 'molt defect', or 'paralysed' phenotypes (based on knockouts/knockdowns/variants) is downloaded from WormBase (Howe *et al.* 2016). In addition, a list of *D. melanogaster* genes with 'lethal', 'sterile' or 'paralytic' (paralysed) phenotypes is downloaded from FlyBase (Millburn *et al.* 2016). Lethal phenotypes are weighted more heavily than other phenotypes. If a nematode/platyhelminth gene belongs to a Compara family containing both *C. elegans* and *Drosophila* genes with lethal phenotypes it is assigned a phenotype score of 1. Otherwise, nematode genes that belong to families with *C. elegans* genes that have lethal phenotypes score 0.9 and those in families with essential *Drosophila* (but not *C. elegans*) genes score 0.8. Platyhelminth genes

belonging to families with essential *C. elegans* or *Drosophila* genes score 0.8. If a nematode/platyhelminth gene belongs to a Compara family containing both *C. elegans* and *Drosophila* genes that both have 'sterile' phenotypes, or both have 'paralysed' phenotypes, it is assigned a score of 0.7. Otherwise, for a nematode gene, if it belongs to a family with a *C. elegans* gene with a paralysed/sterile/L3 arrest/molt defect phenotype, it is assigned 0.6; and if it belongs to a family with a fly (but no *C. elegans*) gene of paralysed/sterile phenotype, it is assigned 0.5. A platyhelminth gene that belongs to a family with a *C. elegans* or *Drosophila* gene of paralysed/sterile/L3 arrest/molt defect phenotype is assigned 0.5. The distribution of phenotype scores for nematode/platyhelminth genes is typically something like this: 1:5%, 0.9:22%, 0.8:8%, 0.7:0.5%, 0.6:7%, and 0.5:4%.

4. A *chokepoint score* is set to 1 for nematode/platyhelminth genes predicted to encode chokepoint enzymes and belonging to Compara families with ≥3 predicted chokepoints. Because chokepoint predictions are not very accurate, predicted chokepoints that do not belong to such families are assigned a score of 0.1. Non-chokepoint enzymes are assigned a zero score. The score distribution is typically something like this: 1:10%, 0.1:0.4%, 0:90%.
5. A *multi-species score*. To penalise species-specific nematode/platyhelminth gene predictions that could include residual unfiltered contamination (e.g. from bacteria), this score is set to zero for nematode/platyhelminth genes belonging to Compara families with only a single species, but otherwise set to 1. The reason for including this score is that typically 99% of the nematode/platyhelminth genes have multi-species scores of 1.
6. An *expression score*. For the majority of nematode/platyhelminth species, including filarial species, drugs should target the adult stage. However, for some species, targeting other stages is important, for instance: the metacestode stage for cestodes, larvae of *Trichuris* and *Strongyloides*, and somules of *Schistosoma*. In the absence of expression data for many stages from many species, adult and metacestode (for cestodes) expression data are used here. For

each nematode/platyhelminth gene, expression data from the most closely related ‘reference’ species is used: *H. contortus* for clade V; *A. suum* for Ascaridomorpha; *B. malayi* and *O. volvulus* for Spiruromorpha and Oxyuridomorpha; *T. muris* for clade I; *S. ratti* for clade IV; *H. microstoma* for cestodes except *E. multilocularis* for *Taenia* species; and *S. mansoni* for trematodes. If the expression level of any corresponding ‘reference’ gene is ≥ 5.0 RPKM in adults or metacestodes, an expression score of 1 is assigned, otherwise zero is used. Typically, ~70% of the nematode/platyhelminth genes have expression scores of 1.

RNASeq studies and samples of interest are obtained from the European Nucleotide Archive (/www.ebi.ac.uk/ena) and ArrayExpress (www.ebi.ac.uk/arrayexpress): *H. contortus* (run accession SRR928055, SRR928056), *A. suum* (SRR504556, SRR504557, SRR504558, SRR504559, SRR504560, SRR504561), *B. malayi* (ERR048961, ERR048962, ERR048970, ERR048972), *O. volvulus* (ERR225734, ERR485009), *T. muris* (ERR279677, ERR279678, ERR279676), *S. ratti* (ERR299169, ERR299175, ERR299168, ERR299174, ERR299170, ERR299176), *H. microstoma* (adult: ERR225730, ERR225728, ERR225729; metacestode: ERR337915, ERR337928, ERR337940, ERR337952, ERR337964, ERR337976), *E. multilocularis* (metacestode: ERR337932, ERR337944, ERR337956, ERR337968, ERR337906, ERR337919), and *S. mansoni* (ERR022873). The analysis of each sequencing run is performed by the iRAP pipeline of the EMBL-EBI Gene Expression Team and downloaded from their RNA-seq Analysis API (www.ebi.ac.uk/fg/rnaseq/api/doc). This pipeline aligns quality-filtered reads to reference genomes from WormBase ParaSite (Howe *et al.* 2017) using TopHat 2 (Trapnell *et al.* 2012); and then quantifies expression of genes and exons in the corresponding GTF file from WormBase ParaSite using HTSeq (Anders *et al.* 2015) (intersection-non-empty mode) and DEXSeq (Anders *et al.* 2012) respectively.

7. A *species-distribution score* is calculated. To give greater preference to potential nematode/platyhelminth targets that are found in multiple species (so that the same drug may target multiple species), a score of 1 is assigned if a nematode/platyhelminth gene is present in a Compara family containing ≥90% of the species from a major group of helminths. The major groups were defined as (in order of preference): all nematodes and platyhelminths, just

nematodes, just platyhelminths, cestodes, filaria, trematodes, or schistosomatids. Otherwise, it is assigned a score of 0. Typically, ~90% of the nematode and platyhelminth genes have a species distribution score of 1.

8. An *invertebrate biology score* is calculated. If the ChEMBL BLAST hit is from a closely related animal, it is more likely that the nematodes/platyhelminths have conserved biology, for example, nematodes/platyhelminths may share processes involved in moulting and life cycle control with arthropods. Therefore, if the top ChEMBL BLAST hit of a nematode/platyhelminth gene is to a UniProt protein from a non-chordate metazoan, it is assigned an 'invertebrate biology score' of 1, otherwise 0. We do not assign a score of 1 for matches to chordate proteins, in order to downweight targets that may be shared with the vertebrate hosts of nematodes/platyhelminths. Typically ~3% of the nematode/platyhelminth genes have an invertebrate biology score of 1.
9. A *singleton score* is calculated. Developing drugs against simple single-copy targets is likely to be easier than developing drugs for multigene families. The score is therefore set to 1 for nematode/platyhelminth genes that lack within-species paralogues (according to the Compara database), otherwise zero score is applied. Typically ~53% of the nematode/platyhelminth genes have singleton scores of 1.
10. A *PDB score* of 1 is assigned to helminth proteins that match ChEMBL targets with available structures in the PDBe (Velankar *et al.* 2016). This is to reflect the possibility of structure-aided drug development. Typically ~51% of nematode/platyhelminth genes have a PDB score of 1.
11. An *alignment conservation score* is calculated. Drugs that act against multiple species are highly desirable and this is more likely to occur if the target is sufficiently conserved between nematode/platyhelminth species. Across each column of the alignment, for each Compara family, a score is calculated using the approach of Capra & Singh (2007) using the Jensen-Shannon divergence, with a window size of 3 (on either side of the residue) and the BLOSUM62 background distribution. The overall score for a family is taken as the median of the scores for

all columns that have scores of > -1000. If a nematode/platyhelminth gene belongs to a family with an alignment score of ≥0.68, it is assigned an 'alignment conservation score' of 1, and otherwise 0. Typically, ~5% of nematode/platyhelminth genes have alignment conservation scores of 1.

12. Thus, each of the nematode/platyhelminth genes (output from Step A) is assigned an overall target score based on the sum of scores above, weighted as follows: (50 x BLAST score) + (40 x ChEMBL non-human score) + (30 x phenotype score) + (15 x chokepoint score) + (10 x multi-species score) + (10 x expression score) + (10 x species-distribution score) + (10 x invertebrate-biology score) + (5 x singleton score) + (5 x PDB score) + (5 x alignment-conservation score)

The output from Step B is a 'target score' for each of the nematode/platyhelminth genes with BLASTP hits to single-protein ChEMBL targets (output from Step A).

Step C: collapsing the list of potential helminth drug targets by gene family

The input for Step C is the set of nematode/platyhelminth genes with BLASTP hits to single-protein targets, along with their target scores from Step B.

These nematode/platyhelminth proteins usually include many with BLAST matches to the same ChEMBL protein, so the following steps are taken to reduce redundancy.

1. The list is collapsed to just take the top-scoring nematode/platyhelminth protein (using the 'target score'; see above) matching each ChEMBL protein, leaving fewer nematode/platyhelminth proteins.
2. The list of nematode/platyhelminth proteins usually still contains redundancy, since many nematode/platyhelminth proteins are from the same Compara family and match homologous ChEMBL proteins (e.g., from human, mouse, rat). The list is further collapsed to just the highest-scoring nematode/platyhelminth protein from each Compara family.

The output from Step C is a reduced list of nematode/platyhelminth proteins with BLASTP matches to single-protein ChEMBL targets (*n* nematode/platyhelminth proteins matching *n* ChEMBL targets).

Step D: identifying potential new anthelmintic drugs in ChEMBL

The input to Step D is the reduced list of nematode/platyhelminth proteins with BLASTP matches to single-protein ChEMBL targets, from Step C.

In Step D, The ChEMBL database (Gaulton *et al.* 2017) is first used to identify compounds (including approved drugs, compounds in clinical development and bioactive compounds from the medicinal chemistry literature) that have activities against ChEMBL single-protein targets to which our nematode/platyhelminth proteins have significant (E ≤ 1e-10) BLAST matches (ie. the helminth proteins from Step C). For each of the nematode/platyhelminth genes in the reduced list (from Step C), we take the compounds with activities to its best ChEMBL BLAST hit and also the corresponding compounds for all nematode/platyhelminth genes in the same Compara family (including those that were discarded by Step C and/or have a different top BLAST hit in ChEMBL).

To assess the likely suitability of each compound as a potential new anthelmintic drug, an overall 'compound score' is generated. Compound properties considered are those that would be most advantageous for a neglected disease drug (Oprea & Overington 2015) including:

• Compounds that could be more quickly and cheaply developed into drugs: we prioritise compounds in high clinical development phases and those where a crystal structure is available to help inform molecule design;

• Compound properties: by focussing on compounds with properties consistent with those of oral drugs (Quantitative Estimate of Drug-likeness) and lacking known toxicity issues using Black Box warning information and toxic effect predictors;

• Preferred route of administration for the ideal anthelmintic: compounds with oral or topical administration are considered most desirable.

Thus, each of the ChEMBL compounds with activities against the ChEMBL targets to which our nematode/platyhelminth proteins (output from Step C) have BLAST matches is assigned an overall compound score, by following these steps:

1. The *Quantitative Estimate of Drug-likeness* (QED, Bickerton *et al.* 2012) is retrieved from the ChEMBL database. QED values can range from 0-1, and the closer a QED value is to 1, the more

oral-drug-like is the compound under consideration. Typically here QED values range from 0.01-0.95.

2. The *maximum development phase score* for a compound is set equal to its maximum development phase taken from the ChEMBL database. Typically here the distribution of scores is: 0:99.7%, I:0.02%, II:0.03%, III:0.05%, and IV:0.22%.
3. A *route of administration score* is calculated based on the route of administration, based on information obtained from the FDA orange book, and retrieved here from the ChEMBL database. Compounds with oral or topical routes are given a score of 1. Typically ~0.17% of compounds have a route of administration score of 1.
4. A *black box warning score* is calculated based on Black Box Warnings. The FDA provides Black Box Warnings for approved drugs where use of the drug is associated with serious or life threatening side effects. Here Black Box Warnings are retrieved from the ChEMBL database, and a score of -1 is used to penalise any compounds with Black Box Warnings. Typically ~0.07% of compounds have a black box warning score of -1.
5. A *molecule structural information availability score* is calculated. A score of 1 is assigned to compounds with at least one structure deposited in the Protein Data Bank in Europe (PDBe) (www.ebi.ac.uk/pdbe), using PDBe information retrieved from the ChEMBL database. Typically here ~0.8% of compounds have a molecule structural information availability score of 1.
6. A *toxicology target interaction prediction score* is calculated based on predictions (retrieved from the ChEMBL database) of compounds predicted to interact with known toxicology targets. These predictions originate from the ChEMBL toxicology target prediction pipeline, with prediction models created using activity information in the database collected as part of the HeCatos project (www.hecatos.eu). Compounds predicted to interact with known toxicology targets are penalised with a score of -1. Typically here ~0.004% of compounds have toxicology target interaction scores of -1.
7. The overall compound score is calculated as the weighted score: (5 x QED) + (5 x

Maximum development phase) + (5 x Route of administration) + (5 x Black box warning information) + (5 x Molecule structural information availability) + (2.5 x Toxicology target interaction prediction)

The output from Step D is the nematode/platyhelminth proteins with BLASTP matches to single-protein ChEMBL targets (from Step C) and corresponding compounds from ChEMBL that have activities against those ChEMBL targets, and a 'compound score' for each of those compounds.

Step E: filtering the list of compound-target pairs, by target score

The input to step E is the list of ChEMBL compound-target pairs output from Step D.

To reduce down the number of compounds and targets considered, this list is now filtered to just retain the top 15% of highest-scoring nematode/platyhelminth targets (using the target scores from Step B).

The output from Step E is a reduced list of ChEMBL compound-target pairs.

Step F: filtering the list of compounds, using information on compound-target pairs

The input to step F is the reduced list of ChEMBL compound-target pairs output from Step E.

In Step F, to filter these ChEMBL compound-target pairs, two additional parameters were retrieved from ChEMBL and used for filtering. The first is pCHEMBL, a parameter calculated by ChEMBL that provides a consistent measure of the affinity of a compound for its (ChEMBL) target and is defined as: −log10 (molar IC50, XC50, EC50, AC50, Ki, Kd or Potency). For example, pChEMBL of 5 corresponds to an activity of 10 μM. The second is whether the compound and target co-appear in a structure: where possible, structures are retrieved from the Protein Data Bank in Europe (PDBe) (www.ebi.ac.uk/pdbe) that contained both the compound and a ChEMBL target.

The compounds (from Step E) are filtered by selecting compounds that (i) co-appear in a PDBe (Velankar *et al.* 2016) structure with the ChEMBL target; or (ii) have a high pChEMBL score (median >5), reflecting high potency/affinity for the ChEMBL target).

The output from step F is a further reduced list of compound-target pairs.

Step G: curating a list of known anthelmintic drugs and compounds

In order to collapse the list of compounds output from Step F by chemical class (see below), we first curate a list of known anthelmintic drugs and compounds.

A list of known anthelmintic drugs (human and veterinary) and compounds, including nematicidal compounds is collated by:

(a). listing compounds with WHO ATC (Anatomical Therapeutic Chemical) code P02 (anthelmintics);

(b). listing compounds with WHO ATCvet code QP52 (anthelmintics) or QP54 (avermectins and milbemycins). Note that halodone is listed as a veterinary anthelmintic, but we do not find any other evidence of that, so we exclude it;

(c). finding compounds listed with anthelmintic activity in ‘The use of stems in the selection of International Nonproprietary Names (INN) for pharmaceutical substances’ (WHO, 2013, www.who.int/medicines/services/inn/stembook/e);

(d). finding compounds described as anthelmintic in the Merck Medical Manual or Merck Veterinary Manual (www.merckmanuals.com);

(e). listing compounds described as anthelmintic drugs or compounds for human or veterinary use from the scientific literature (Mehlhorn 2008; Anand & Sharma 1997; Elks 1990; Marr & Komuniecki 2003; Moffat *et al.* 2011; Lewis 1998; Oliver-Bever 1983; Grieve 2015; Allegretti 2012; Holden-Dye & Walker 2014);

(f). listing compounds tagged with MeSH categories 'anthelmintic', ‘anticestodal’ or ‘antinematodal’ in PubChem (Kim *et al.* 2016);

(g). finding compounds listed as having ‘anthelmintic’, ‘anticestodal’, ‘antischistosomal’, ‘antitrematodal’, or ‘fasciolicide’ activity in KEGG Drugs (Kanehisa *et al.* 2013);

(h). finding compounds listed as ‘anthelmintic’ in the ChEBI database (Hastings *et al.* 2016);

(i). finding compounds listed in ChEMBL (Gaulton *et al.* 2017) with the keywords ‘anthelmintic’ OR ‘anthelminthic’ in any of the following fields: ATC code description, mechanism of action, USAN (United States Adopted Name) stem definition, or indication class;

(j). finding compounds tagged with MeSH categories 'anthelmintics’, ‘antinematodal’, ‘filaricides’, ‘antiplatyhelmintic’ or ‘schistosomicides’ in DrugBank (Law *et al.* 2014), or identified by a text search

for ‘anthelmintic OR antihelminthic OR antinematodal OR antitrematodal’ in DrugBank;

(k). finding compounds from PubChem identified by searching for ‘anthelmintics OR anthelmintic OR nematocide OR nematicide’, and with low structural similarity to the list of compounds from (a)–(j) above (similarity ≤0.4 when calculated using the Tanimoto coefficient of ECPF4 fingerprints as implemented in the RDKit toolkit (www.rdkit.org)).

Where the source (see (a)-(k) above) only gives a name (not structure) for the compound, the OPSIN software (Lowe *et al.* 2011) is used to convert IUPAC names to SMILES strings that are then used to search PubChem and ChEMBL for specific compounds. The resulting list of anthelmintic drugs (human and veterinary) and compounds typically includes >260 compounds.

The output from Step G is a curated list of known anthelmintic compounds and their corresponding SMILES strings.

Step H: classifying known anthelmintic compounds into clusters

The input for Step H is the curated list of known anthelmintic compounds from Step G, and their SMILES strings.

To classify these compounds into clusters:

1. To calculate a dendrogram of known anthelmintic compounds, the SMILES strings for known anthelmintic compounds are used. For several compounds, multiple SMILES may be used (for example, because of different stereoisomers). The pairwise similarity between the SMILES strings in this data set is calculated based on ECFP4 fingerprints of length 16384 as implemented in RDKit 2016.09.4 (rdkit.org). The Tanimoto similarity between pairs of fingerprints is calculated using ChemFP 1.1 (Dalke 2011).
2. The similarities have range 0.0-1.0, and are converted to distances using 1.0 - similarity. The distance matrix is read into R, and hierarchical clustering performed using Ward's minimum variance method (using ‘ward.D’ in the ‘hclust’ function). The dendrogram is cut at a height which separates the well known chemical classes of anthelmintic compounds avermectins, milbemycins, and imidazothiazoles from other compounds. If the members of the clusters defined in this way have little in common with respect to chemical structure, the cluster is

further split up. This typically results in ~44 chemical classes.

The output from Step H is a chemical class for each of the known anthelmintic compounds.

Step I: classifying the top drug candidates into chemical classes

To create a ‘diverse screening set’, we perform a diversity analysis to classify the compounds from Step F (which we refer to as our 'top drug candidates') into chemical classes.

The data set consisted of the compounds from Step F for which a SMILES string is available in ChEMBL22 or CHEMBL16, along with the SMILES strings for known anthelmintic compounds used for the dendrogram of anthelmintic compounds (Step H). ChEMBL16 is used to obtain SMILES strings for some metal-containing compounds that are lacking SMILES strings in ChEMBL22. There are typically a small number (usually <100) of the ‘top drug candidate’ compounds (for example, polymers) for which there are no SMILES in ChEMBL, so these are not included in the diversity analysis.

The pairwise similarity between the SMILES strings is calculated, and the similarities converted to distances, as for the dendrogram of known anthelmintic compounds (Step H). This number of compounds is too large to construct a dendrogram using hierarchical clustering. Instead, by manually examining the clusters in the dendrogram of known anthelmintic compounds (from Step H), we find that compounds in the same cluster generally have distances of $\leq x$ (e.g. 0.65) from each other, and those in different clusters usually have distances of $> x$. Thus, a set of clusters is found by first constructing a graph in which each compound is represented by a node, and each pair of nodes is joined if the distance between the compounds is $\leq x$; and finding connected components in this graph.

Some of the connected components in this graph may be very large and include relatively dissimilar compounds among the known anthelmintic compounds. Therefore, we further split the connected components by using a community (cluster) detection algorithm to find clusters within the subgraph corresponding to each connected component. To find the communities, we use the ‘community’ Python module (Aynaud 2009) to find optimal communities in terms of the modularity measure (as described previously by Blondel *et al.* 2008), using the similarities between compounds as the edge weights. When the community detection algorithm is applied to the connected components in the

graph of top drug candidates and anthelmintic compounds, it splits the components into smaller communities (clusters). Upon manual examination, it is usually found that some of these clusters still contains relatively dissimilar compounds among the known anthelmintic compounds. Therefore, for each cluster, we perform hierarchical clustering in R using Ward's minimum variance method, and cut the dendrogram at a height of *y* (e.g. 0.85). The value of *y* is chosen by trying various heights, and finding the height at which anthelmintic compounds known to have similar structures (based on Step H) are placed in the same cluster.

We consider that the resultant clusters represent the chemical classes to which the top candidate compounds belong.

Step J: further filtering the list of compounds, using information on compound-target pairs

At this stage, some of the 'top drug candidates' are further filtered, by discarding the medicinal chemistry compounds that do not co-appear in a PDBe structure with the ChEMBL target or have a median pCHEMBL score of >7, which leaves a smaller number of top drug candidates.

The same chemical classes identified above for the compounds in Step I are still used for the smaller set of candidates from Step J.

Step K: further filtering the list of compounds, using information on availability for purchase, and chemical class

This set of candidates is further filtered by:

(a). just taking compounds listed as available for purchase in ZINC (see below), and

(b). for each nematode/platyhelminth target, just taking the highest-scoring compound (according to our compound score; see Step D above) from each chemical class.

ZINC 15 (Sterling & Irwin 2015) is used to identify compounds available for purchase in the data set. The subset of ZINC 15 with the highest level of availability ('in stock') is used. Identity-matching should use the parent compound (a single largest component) and standard InChIs, first directly and then after removing the charge and atom-based stereochemistry layers.

This gives the final 'diverse screening set' (typically several thousand compounds).

Anticipated Results

The output from the procedure is a diverse screening set of several thousand compounds from ChEMBL (including both approved drugs and medicinal chemistry compounds), and their putative targets in nematodes/platyhelminths.

References

Allegretti, S. M. *et al.* The Use of Brazilian Medicinal Plants to Combat *Schistosoma mansoni*. In *Schistosomiasis*, InTech. (2012).

Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids res.* **25**, 3389-3402 (1997).

Anand, N. and S. Sharma, Eds. *Approaches to Design and Synthesis of Antiparasitic Drugs*, Elsevier Science (1997).

Anders, S. *et al.* Detecting differential usage of exons from RNA-seq data. *Genome res.* **22**, 2008-2017 (2012).

Anders, S. *et al.* HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).

Aynaud, T. Community detection for NetworkX’s documentation, perso.crans.org/aynaud/communities/index.html (2009).

Berriman *et al.* The genome of the blood fluke *Schistosoma mansoni*. *Nature*. **460**, 352-358 (2009).

Bickerton, G.R. *et al.* Quantifying the chemical beauty of drugs. *Nat Chem.* **4**, 90-98 (2012).

Blondel, V. D. *et al.* Fast unfolding of communities in large networks. *J. Stat. Mechanics-Theory and Experiment* **2008 (10)**, P10008 (2008).

Capra, J.A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875-1882 (2007).

Dalke, A. chemfp - fast and portable fingerprint formats and tools. *Journal of Cheminformatics* **3(Suppl 1)** (2011).

Elks, J., Ed. *The Dictionary of Drugs: Chemical Data*, Springer US (1990).

Gaulton, A. *et al*. The ChEMBL database in 2017. *Nucleic acids res.* **45**, D945-D954 (2017).

Grieve, M. *A Modern Herbal: The Complete Edition*, Stone Basin Books (2015).

Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids res.* **44**, D1214-D1219 (2016).

Holden-Dye, L. & Walker, R.J. Anthelmintic drugs and nematicides: studies in *Caenorhabditis elegans. WormBook*, 1-29 (2014).

Howe, K.L. *et al.* WormBase 2016: expanding to enable helminth genomic research. *Nucleic acids res.* **44**, D774-D780 (2016).

Howe, K.L. *et al.* WormBase ParaSite - a comprehensive resource for helminth genomics. *Molecular and biochemical parasitology* **215**, 2-10 (2017).

Kanehisa, M. Molecular network analysis of diseases and drugs in KEGG. *Methods Mol. Biol.* **939**, 263-275 (2013).

Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic acids res.* **44**, D1202-1213 (2016).

Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids res.* **42**, D1091-1097 (2014).

Lewis, R. A. *Lewis' Dictionary of Toxicology*, CRC Press (1998).

Lowe, D.M. *et al.* Chemical Name to Structure: OPSIN, an Open Source Solution. *Journal of Chemical Information and Modeling* **51**, 739-753 (2011).

Marr, J. J. *et al.*, Eds. *Molecular Medical Parasitology*, Academic Press (2003).

Mehlhorn, H., Ed. *Encyclopedia of Parasitology*, Springer (2008).

Millburn, G.H. *et al.* FlyBase portals to human disease research using *Drosophila* models. *Dis. Model Mech.* **9**, 245-252 (2016).

Moffat, A. *et al.*, Eds. *Clarke's Analysis of Drugs and Poisons*, Pharmaceutical Press (2011).

Oliver-Bever, B. Medicinal plants in tropical West Africa. III. Anti-infection therapy with higher plants. *J. Ethnopharmacol.* **9**, 1-83 (1983).

Oprea, T.I. & Overington, J.P. Computational and Practical Aspects of Drug Repositioning. *Assay Drug Dev. Technol.* **13**, 299-306 (2015).

Sterling, T. & Irwin, J. J. ZINC 15--Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324-2337

(2015).

Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562-578 (2012).

Tyagi, R. *et al.* Comparative analysis of metabolism in parasitic worms. *Protocol Exchange* (2018).

Velankar, S. *et al.* PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic acids res.* **44**, D385-395 (2016).

Vilella, A.J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome res.* **19**, 327-335 (2009).

Acknowledgements

We would like to thank the WTSI Pathogen Informatics team, especially Jacqueline Keane and Andrew Page.

Figures

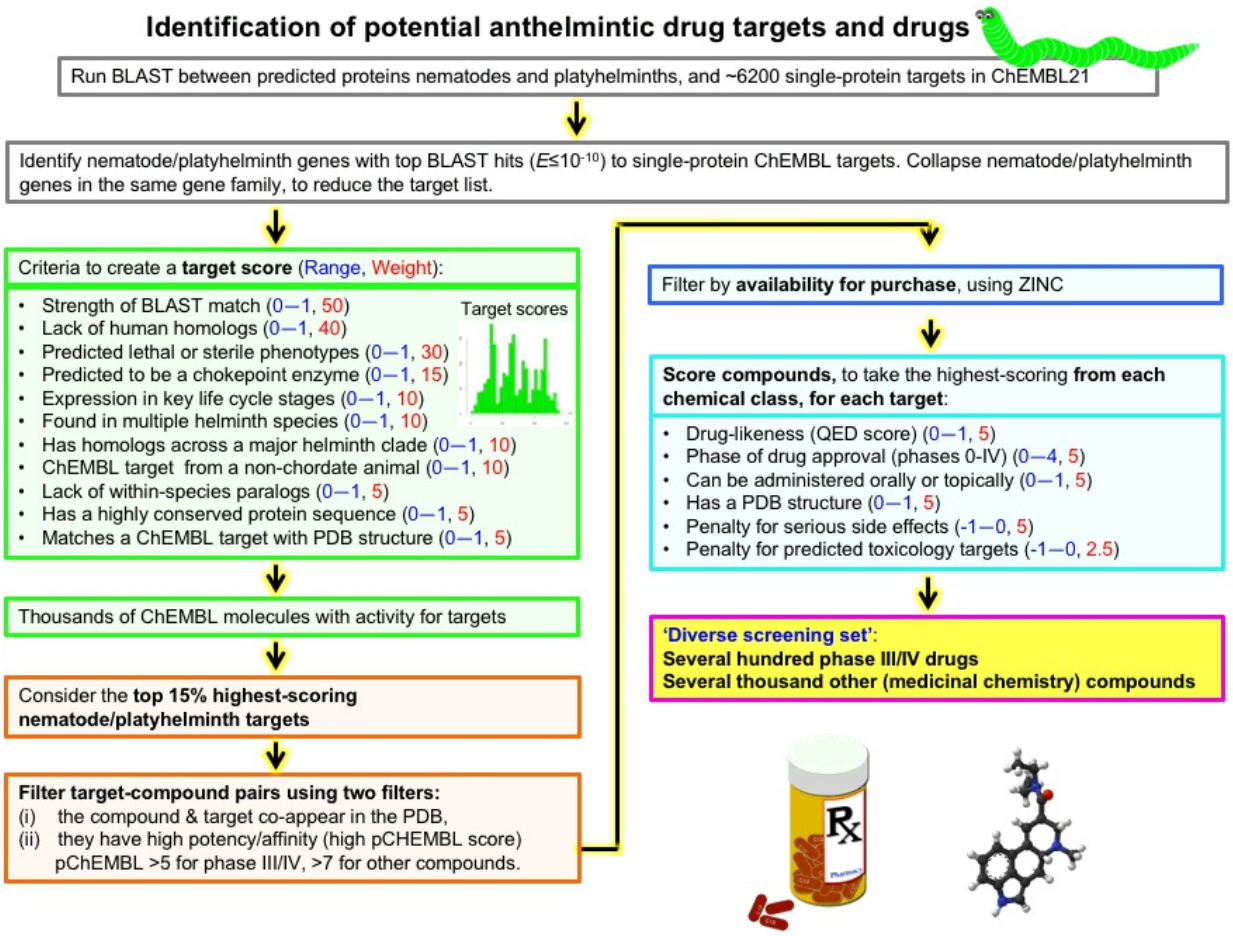

Figure 1

Workflow