Discovering therapeutic activities from venoms using differential gene expression

Joseph D. Romano[1,2,3], Hai Li[2,4], Ronald Realubit[2,4], Charles Karan[2,4], and Nicholas P. Tatonetti[1,2,3,5,*]

July 11, 2019

Abstract: Venoms are a diverse and complex group of natural toxins that have been adapted to treat many types of human disease, but rigorous informatics approaches for discovering new therapeutic activities are scarce. We have designed and validated a new platform—named `VenomSeq`—to systematically generate putative associations between venoms and drugs/diseases via high-throughput transcriptomics and perturbational differential gene expression analysis. In this study, we describe the architecture of `VenomSeq`, and its evaluation using the crude venoms from 25 diverse animal species. By integrating comparisons to public repositories of differential expression, associations between regulatory networks and disease, and existing knowledge of venom activity, we provide a number of new therapeutic hypotheses linking venoms to human diseases supported by multiple layers of preliminary evidence. We are currently performing validation experiments *in vitro* to corroborate these findings.

Contents

[1] Department of Biomedical Informatics, Columbia University, New York, NY, 10032.
[2] Department of Systems Biology, Columbia University, New York, NY, 10032.
[3] Institute for Genomic Medicine, Columbia University, New York, NY, 10032.
[4] Columbia Genome Center, Columbia University, New York, NY, 10032.
[5] Data Science Institute, Columbia University, New York, NY, 10032.
[*] Corresponding author. Email: nick.tatonetti@columbia.edu.

1. Introduction

Venoms are complex mixtures of organic macromolecules and inorganic cofactors that are used for both predatory and defensive purposes. Since the dawn of recorded history, humans have exploited venoms and venom components for treating a wide array of illnesses and conditions, a trend which has continued into modern times [Lewis and Garcia, 2003]. Currently, approximately 20 venom-derived drugs are in use world-wide, with 6 approved by the US Food and Drug Administration for clinical use, and many more currently undergoing clinical trials [Pennington et al., 2018]. As new discovery of small-molecule drugs has slowed considerably in recent decades, venoms and other natural products hold great promise for discovering innovative treatments for disease and injury, especially for diseases that have evaded treatment through conventional medical science.

Furthermore, venoms are incredibly diverse. Depending on the species, a single venom can contain hundreds of distinct compounds [Terlau and Olivera, 2004]. Since current estimates claim that millions of venomous species exist across the tree of life, venom-derived compounds provide an immense library of evolutionarily optimized candidates for drug discovery [von Reumont et al., 2014, Calvete et al., 2009].

Toxinologists have applied modern high-throughput sequencing (HTS) methodologies to the study of venoms (a field that has come to be known as *venomics*) [Calvete et al., 2009]. Venomics generally involves the sequencing and structural identification of multiple types of macromolecules—genomic DNA, venom gland mRNA transcripts, and/or venom proteins—to best evaluate which genes, transcripts, and polypeptides (including post-translational modifications) are present in a venom and convey its activity.

Venomics has become a popular framework for drug discovery in recent years. However, other applications of HTS and biomedical data science beyond discovery/evaluation of venom components can be used for drug discovery. One such application is data-driven analysis of *perturbational gene expression* data, in which human cells are exposed *in vitro* to controlled dosages of candidate compounds and then profiled for differential gene expression via RNA sequencing (RNA-Seq). In this paper, we present `VenomSeq`—a new informatics workflow for discovering associations between venoms and therapeutic avenues of treatment for disease.

Briefly, `VenomSeq` involves exposing human cells to dilute venoms, and then generating differential expression profiles for each venom, comprised of the significantly up- and down-regulated genes in cells perturbed by the venom. We then compare the differential expression profiles to data from public compendia of perturbational gene expression data and gene regulatory data corresponding to disease states. `VenomSeq` works in the absence of any predefined hypotheses, instead allowing the data to suggest hypotheses that can then be explored comprehensively using rigorous traditional approaches.

Figure 1: Graphical abstract outlining the `VenomSeq` workflow.

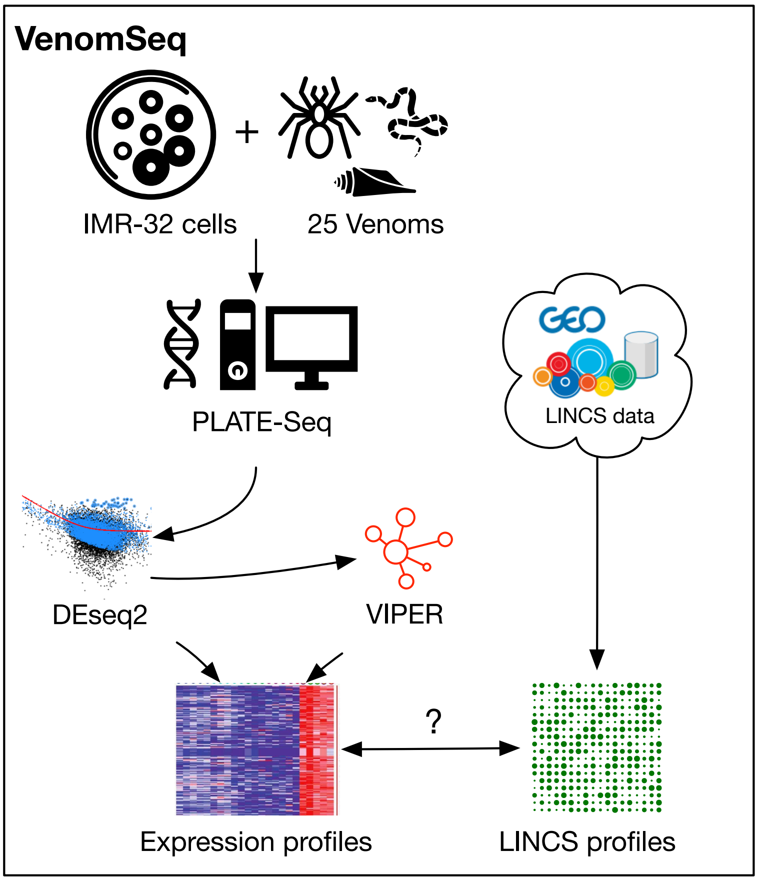


Table 1: Statistics for *S. maurus* growth inhibition data.

***S. maurus* venom vs. IMR-32**		
GI_{20}(µg µl^{-1})		0.0926
R^2		0.991
Hill slope	Bottom	−2.096
	Top	92.572
	log GI_{50}	−0.640
	Slope (h)	−1.928

2. Results

2.1. Venom dosages

In order to optimize the exposure concentrations of each venom, we performed growth inhibition assays on human cells exposed to varying concentrations of the venoms. This is necessary to minimize the impact of toxicity while ensuring the venom is in high enough concentration to exert an effect on the human cells. Since each venom is comprised of many (largely unknown) molecular components, we performed the assays on samples of venom measured in mass per volume, rather than compound concentration (molarity). We used GI_{20}—the concentration of a venom at which it inhibits growth of the human cells by 20%—as the effective treatment dose in all subsequent experiments.

The experimental GI_{20} values and complete dose-response data for each of the 25 venoms are provided in **Appendix A** (**Table 9**), a sample of which is reproduced (for *S. maurus*) in **Table 1**. The resulting growth inhibition curves for all venoms are shown in **Figure 2**. Venoms from *L. colubrina*, *D. polylepis*, *S. verrucosa*, *S. horrida*, *C. marmoreus*, *O. macropus*, and *P. volitans* did not demonstrate substantial growth inhibition at any tested concentration, so for those venoms we instead performed sequencing at 1.0 µg µl^{-1}, which is the highest concentration used in the growth inhibition curves.

2.2. mRNA sequencing of venom-perturbed human cells

After determining appropriate dose concentrations for each venom, we performed RNA-Seq on human IMR-32 cells exposed to the individual venoms. **Table 2** summarizes the experimental conditions used for sequencing. After transforming the raw sequencing reads to gene counts

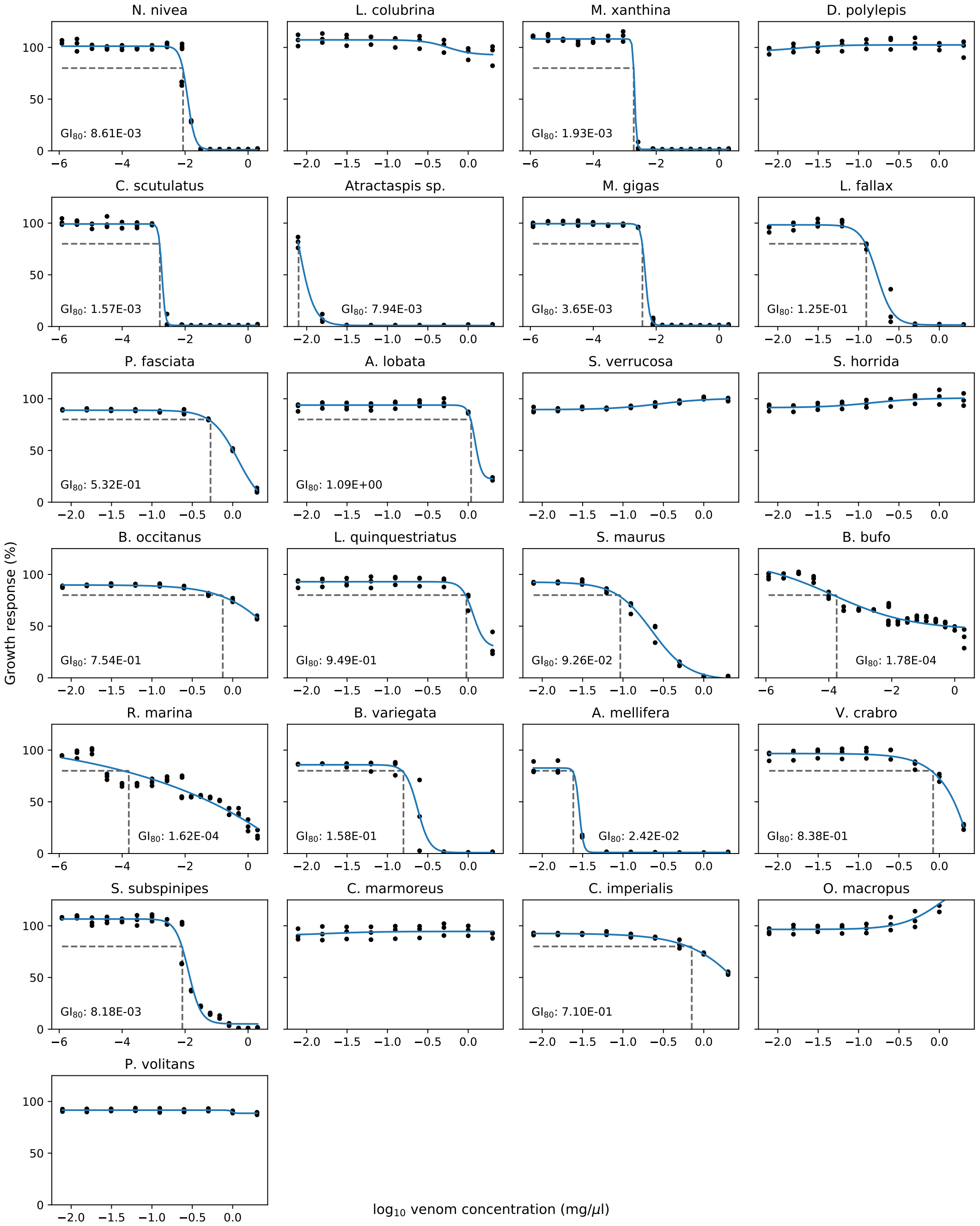


Figure 2: Growth inhibition plots for each of the 25 venoms. GI$_{80}$ values are provided, unless growth inhibition was not observed (in which case sequencing was instead performed at $2\,\mathrm{mg\,\mu l^{-1}}$).

Table 2: Experimental conditions for RNA-Seq.

Venoms	25 species
Cell line	IMR-32 (Human neuroblastoma)
Dosage	GI_{20} for each venom
Time points	6/24/36 hours post-treatment
Replicates	3 per time point per venom
Controls	12 water controls, 9 untreated
Solvent	Water

(see §7.4), we compiled the results into a matrix, where rows represent genes, columns represent samples, and cells represent counts of a gene in a sample. For detailed quality control data, refer to **Appendix A**, which includes links to related files. The raw (i.e., FASTQ files produced by the sequencer) and processed (i.e., gene counts per sample) data files are available for download and reuse on NCBI's Gene Expression Omnibus database; accession GSE126575.

2.3. Differential expression profiles of venom-perturbed human cells

We constructed differential expression signatures for each of the 25 venoms as described in §7.5, where each signature consists of a list (length ≥ 0) of significantly upregulated genes, and a list (length ≥ 0) of significantly downregulated genes. The specific expression signatures are available on FigShare at `https://doi.org/10.6084/m9.figshare.7609160`. An excerpt from the expression signature for *O. macropus* is shown in **Table 3**. The total number of differentially expressed genes for each venom ranges from 2 genes (*Laticauda colubrina* and *Dendroaspis polylepis polylepis*) to 1494 genes (*Synanceia verrucosa*). Note that these signatures are specific to IMR-32 cells—we expect that the same procedure applied to other cell lines would yield substantially different expression signatures.

Gene-wise statistical significance is a function of both $\log_2$ fold change and the number of observed counts. This relationship is illustrated in **Figure 15**, which is derived from the same data shown in **Table 3** (for *O. macropus*).

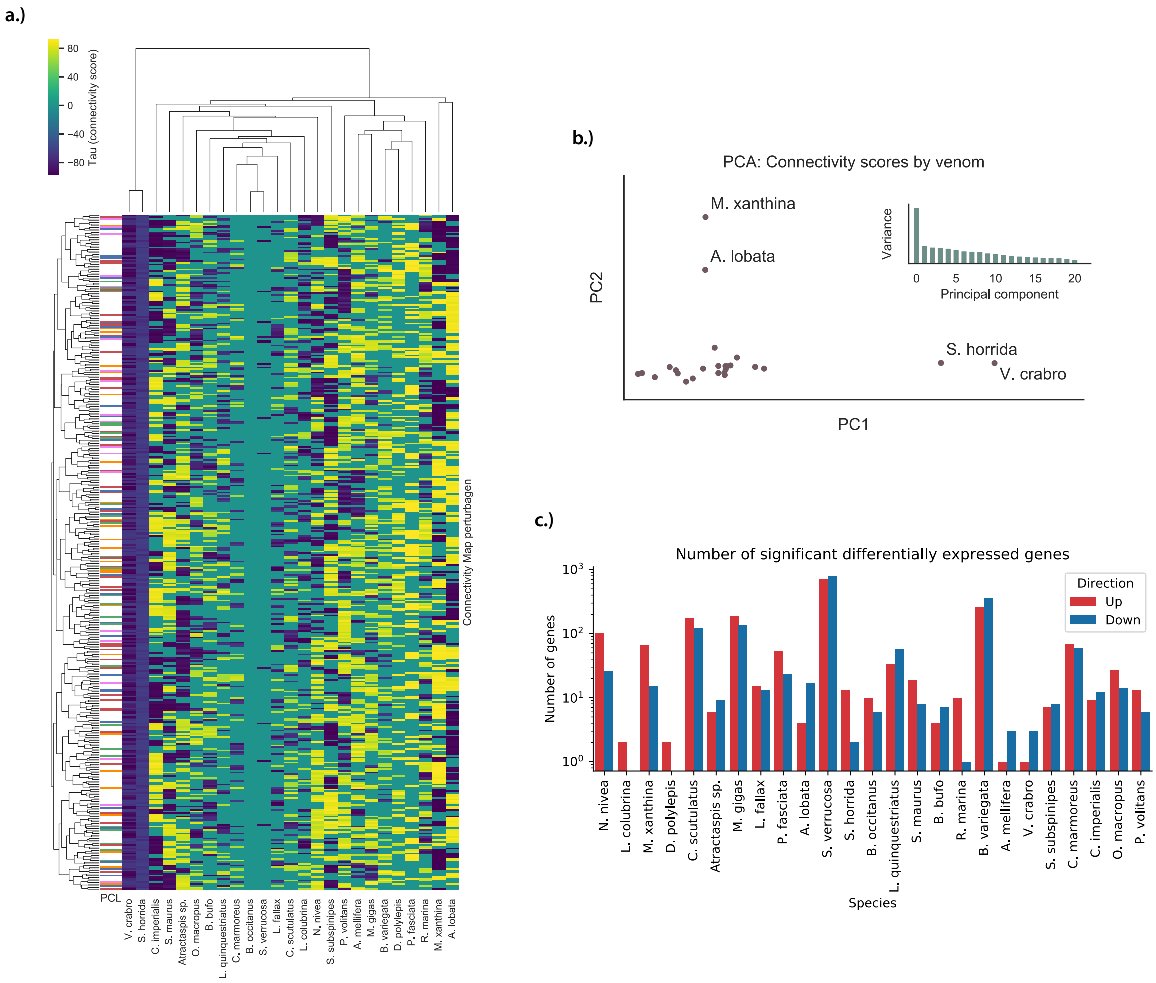


Figure 3: Connectivity analysis results. **a.)** Heatmap of τ-scores between the 25 venom perturbations and the 500 Connectivity Map signatures with the highest variance across all venoms. A distinct hierarchical clustering pattern is evident across the venom perturbations, although it does not conform to any obvious grouping pattern of the venoms. **b.)** Principle component analysis of the 25 venom perturbations, where features are all τ-scores between the venom and signatures from the Connectivity Map reference database. 4 distinct outliers are labeled—these venoms correspond to outliers in the heatmap. Also shown are the ratios of variance explained by each of the first 21 principle components—after the first principle component, the distribution is characterized by a long tail, suggesting that much of the variance is spread across many dimensions, underscoring the complexity of the connectivity score data. **c.)** Barplot showing the number of significant differentially expressed genes for IMR-32 cells exposed to each of the 25 venoms.

Table 3: Partial differential expression signature for *O. macropus*. Most of the significantly differentially expressed genes (35 of 41 total) are omitted for brevity.

Gene	Base mean	$\log_2$-FC	Wald statistic	p-adj
SPRY4	37.38	-2.27534	-3.3084	0.0991
REPIN1	38.30	-0.95256	-4.3326	0.0061
DUSP14	33.88	-0.91311	-3.3327	0.0991
⋮	⋮	⋮	⋮	⋮
BRD3	130.81	1.37645	4.115	0.0096
RSRC1	63.48	1.38140	4.2042	0.0091
BAZ1B	120.05	1.69463	5.0846	0.0003

2.4. Associations between venoms and existing drugs

Using publicly-available differential expression profiles for existing drugs—many with known effects and/or disease associations—we were able to identify statistically significant associations between venoms and classes of drugs. These associations are based on the methods designed by the Connectivity Map (CMap) team [Lamb et al., 2006], and utilize their perturbational differential expression data as the "gold standard" against which to evaluate the venom expression data. In short, this approach uses a Kolmogorov-Smirnov–like signed enrichment statistic to compare a query signature (i.e., venoms) to all signatures in a reference database (i.e., known drugs), normalizing for cell lines and other confounding variables, and finally aggregating scores of 'like' signatures (i.e., drug MoAs) using a maximum-quantile procedure. Complete details of these methods are provided in §7.6.1.

Different venoms yield different profiles of connectivity scores based on the genes present in their differential expression signatures. For example, all connectivity scores between *B. occitanus* and CMap perturbagens are zero, and all connectivity scores between *S. horrida* and CMap perturbagens are negative, which suggest that these venoms either behave like no known perturbagen classes, or that the venoms have no therapeutic activity on IMR-32 cells. Kernel density plots of the connectivity scores for each venom are shown in **Figure 4**. In **Figure 3**, we show several visualizations of the connectivity analysis results that highlight characteristics of the data. Interestingly, when hierarchical clustering is performed on the connectivity scores by venom perturbation, the venom perturbations form robust clustering patterns that persist

Table 4: Venom–drug class associations.

Venom	Drug class (MoA)
Synanceia horrida	ATPase inhibitor CDK inhibitor DNA synthesis inhibitor
Scolopendra subspinipes dehaani	T-type Ca^{2+} channel inhibitor
Pterois volitans	Topoisomerase inhibitor
Argiope lobata	ATPase inhibitor PI3K inhibitor PPARγ agonist
Scorpio maurus	FGFR inhibitor
Rhinella marina	HIV protease inhibitor

across multiple non-overlapping subsets of the connectivity data. This suggests that the clustering corresponds to meaningful characteristics of the venom perturbations in comparison to known drugs, although these characteristics are not readily apparent (i.e., the clustering does not reproduce taxonomy, or other obvious traits of the venoms).

The associations we identified are shown in **Table 4**. As we anticipated, only some venoms show strong associations to any classes of drugs. Interestingly, only one venom (*S. subspinipes dehaani*) was linked to an ion channel inhibition MoA—venoms, in general, tend to have powerful ion channel blocking or activating effects. However, this may be due to a preponderance of non-ion channel MoAs in the CMap data rather than an actual lack of ability to identify ion channel activity.

Many of these MoAs comprise either well-established or emerging classes of cancer drugs. Some that have been used extensively as chemotherapeutic agents include CDK inhibitors (palbociclib, ribociclib, and abemaciclib), topoisomerase inhibitors (doxorubicin, teniposide, and irinotecan, among others), and DNA synthesis inhibitors (mitomycin C, fludarabine, and floxuridine). Meanwhile, PI3K inhibitors and FGFR inhibitors are classes of "emerging" chemotherapy drugs, each recently leading to many high-impact research studies and early-stage clinical trials.

The other classes are indicated for a diverse range of diseases, including circulatory and

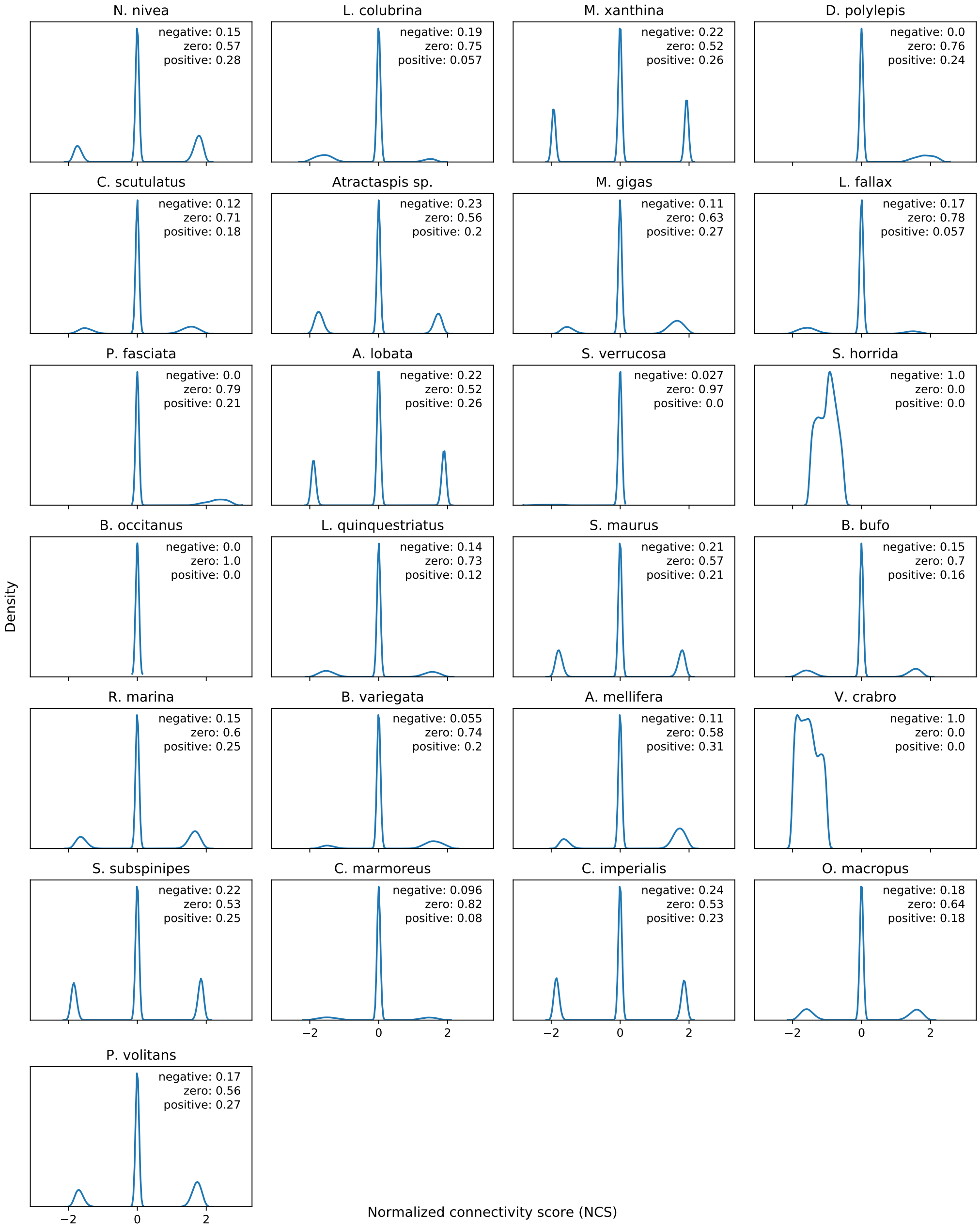


Figure 4: Kernel density plots of normalized connectivity scores (NCSs) for each of the 25 venoms. Note the tendency to introduce sparsity by setting NCS to zero if the quantities a and b have opposite signs (see §7.6.1). Text labels indicate proportion of NCSs for a single venom that are negative, zero, or positive. Each plot is based on 473,647 NCSs (all differential expression profiles in GSE92742 [Subramanian et al., 2017]).

mental conditions (calcium channel blockers), and cardiac abnormalities (ATPase inhibitors). PPAR receptor agonists have been used to treat diabetes, hyperlipidemia, pulmonary inflammation, and cholesterol disorders.

We are in the process of validating several of the associations listed in **Table 4** using targeted, cell-based assays, the results of which will be documented in subsequent publications.

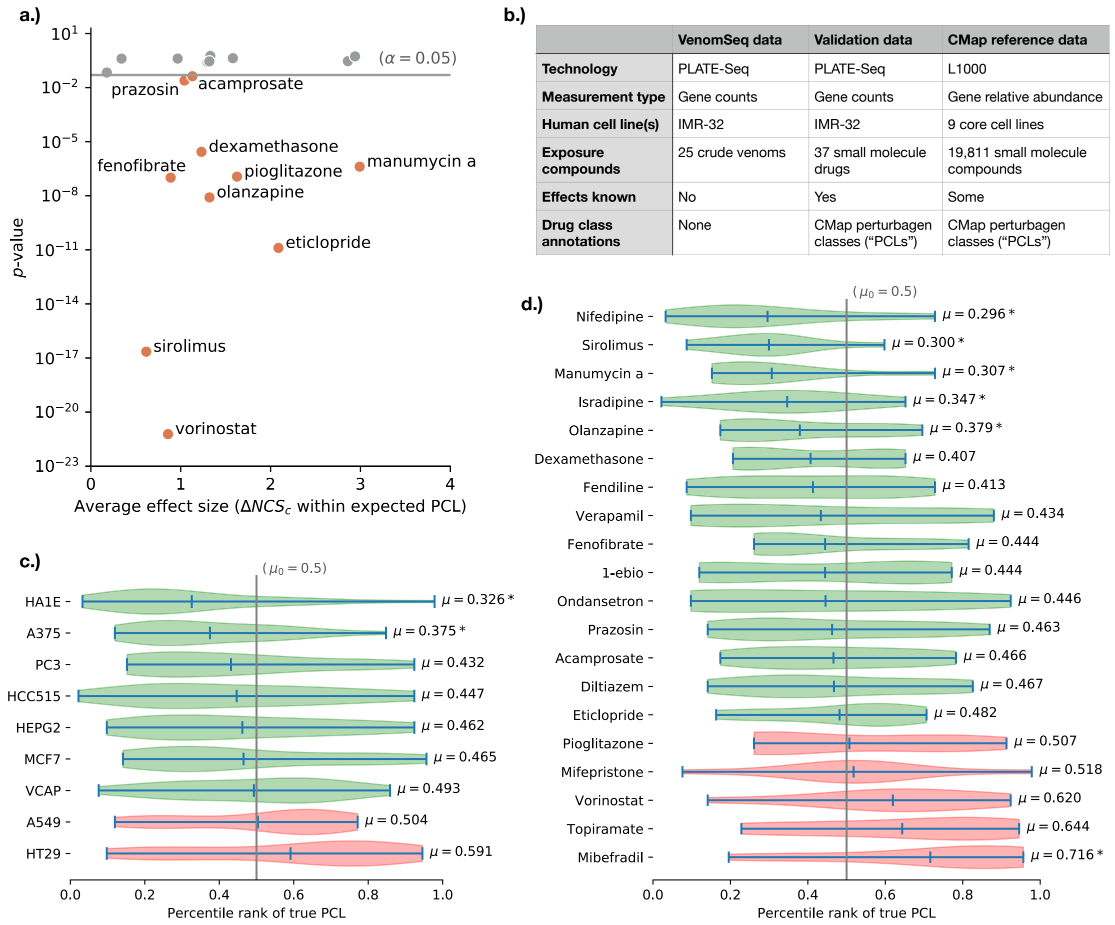
2.5. `VenomSeq` technical validation

Following the procedures described in §7.7, we used a secondary PLATE-Seq dataset of 37 existing drugs (with known effects) tested on IMR-32 cells to assess whether the sequencing technology (PLATE-Seq) and cell line (IMR-32) employed by `VenomSeq` are compatible with connectivity analysis and the CMap reference dataset. In this dataset, we were able to map 20 of the 37 drugs to a single existing CMap perturbational class (PCL). The drugs, their modes of action, and the PCLs of which they are members are listed in **Table 5**.

2.5.1. `VenomSeq` technical validation: Recovering connectivity by integrating cell lines

When we aggregated all connectivity scores between a known drug and members of the same PCL in the CMap dataset, irrespective of cell line, the connectivity scores are significantly greater than those in a null model in 12 out of 20 instances, which indicates that drugs within the same functional class tend to have more similarities in the query and reference datasets than if the compounds are chosen at random. In all 20 cases, the average effect size[1] was positive, regardless of statistical significance. These—and their corresponding measures of significance—are shown in **Figure 5** and **Table 6**. Overall, these data are congruent with those made by the Connectivity Map team in [Subramanian et al., 2017]—namely, that expected connections between query drugs and reference compounds can be recovered for some PCLs, but not for others. Importantly, in both our observations and the observations in [Subramanian et al., 2017], PCLs related to highly conserved core cellular functions perform better under this approach.

[1] Effect size is defined as the average difference between connectivities within the expected PCL and the null model of random connectivities for the same query



	VenomSeq data	Validation data	CMap reference data
Technology	PLATE-Seq	PLATE-Seq	L1000
Measurement type	Gene counts	Gene counts	Gene relative abundance
Human cell line(s)	IMR-32	IMR-32	9 core cell lines
Exposure compounds	25 crude venoms	37 small molecule drugs	19,811 small molecule compounds
Effects known	No	Yes	Some
Drug class annotations	None	CMap perturbagen classes ("PCLs")	CMap perturbagen classes ("PCLs")

Figure 5: Results of applying the VenomSeq sequencing and connectivity analysis workflow to 37 existing drugs with known effects, to validate the compatibility of PLATE-Seq and IMR-32 cells with the connectivity analysis algorithm and dataset. **a.)** Scatter plot showing validation drugs that are members of a CMap PCL and the mean differences between within-PCL connectivity scores and a null distribution of random connectivity scores for the same drug (**Table 6**). Verticle axis shows the *p*-value of a Student's *t*-test comparing the within-PCL and null connectivity score distributions (corrected for multiple testing). Statistically significant drugs are labeled by name. **b.)** Summary of the validation strategy, showing that the validation dataset bridges certain gaps between the VenomSeq data and the CMap reference data. **c.)** Distributions of rank percentiles of expected ("true") PCLs within the list of all PCLs ordered by average connectivity score (**Table 7**), aggregated by CMap dataset cell lines, and **d.)** validation drugs. Green distributions indicate a shift towards the front of the rank ordered list, indicating stronger compatibility with the PLATE-Seq/IMR-32 query data, based on expected connections, and "*" indicates statistically significant shifts.

Table 5: Drugs used to validate PLATE-Seq and the IMR-32 cell line for connectivity analysis. Not all compounds of a given mechanism of action will necessarily map to that mechanism's associated PCL—PCLs consist of compounds that are members of the same functional class and also have high transcriptional impact.

Drug	Mechanism of Action	CMap perturbagen class (PCL)
Mibefradil	T-type Ca^{2+} channel inhibitor	CP_T_TYPE_CALCIUM_CHANNEL_BLOCKER
Isradipine	L-type Ca^{2+} channel inhibitor	CP_CALCIUM_CHANNEL_BLOCKER
Nifedipine	L-type Ca^{2+} channel inhibitor	CP_CALCIUM_CHANNEL_BLOCKER
Diltiazem	Ca^{2+} channel inhibitor	CP_CALCIUM_CHANNEL_BLOCKER
Verapamil	Ca^{2+} channel inhibitor	CP_CALCIUM_CHANNEL_BLOCKER
Fendiline	Ca^{2+} channel inhibitor	CP_CALCIUM_CHANNEL_BLOCKER
Topiramate	Na^{+} and Ca^{2+} channel modulator	CP_SODIUM_CHANNEL_BLOCKER
Ionomycin	Ca^{2+} channel signal inducer	
1-EBIO	Ca^{2+}-gated K^{+} channel activator	CP_POTASSIUM_CHANNEL_ACTIVATOR
Forskolin	Adenylyl cyclase activator	
Pregabalin	Increases GABA biosynthesis	
Gabapentin	Increases GABA biosynthesis	
Baclofen	GABA$_B$-receptor agonist	
Memantine	Glu-receptor inhibitor	
Acamprostate	Glu-receptor inhibitor	CP_GABA_RECEPTOR_ANTAGONIST
MTEP	Glu-receptor inhibitor	
Ivermectin	Glu-gated Cl^{-} channel inhibitor	
Carbenoxolone	Glucocorticoid metabolism inhibitor	
Mifepristone	Glucocorticoid receptor inhibitor	CP_PROGESTERONE_RECEPTOR_ANTAGONIST
Dexamethasone	Glucocorticoid receptor agonist	CP_GLUCOCORTICOID_RECEPTOR_AGONIST
Aldosterone	Mineralocorticoid receptor agonist	
Spironolactone	Mineralocorticoid receptor inhibitor	
Olanzapine	Dopamine receptor inhibitor	CP_DOPAMINE_RECEPTOR_ANTAGONIST
Eticlopride	Dopamine receptor inhibitor	CP_DOPAMINE_RECEPTOR_ANTAGONIST
Ondansetron	5-HT$_3$ serotonin receptor inhibitor	CP_SEROTONIN_RECEPTOR_AGONIST
Naltrexone	Opioid receptor inhibitor	
Disulfiram	Acetaldehyde dehydrogenase inhibitor	
Cerlitinib	ALK inhibitor	
Crizotinib	ALK inhibitor	
Sirolimus	mTOR inhibitor	CP_MTOR_INHIBITOR
Manumycin a	Farnesyltransferase inhibitor	CP_NFKB_PATHWAY_INHIBITOR
Vorinostat	HDAC (I/II/IV) inhibitor	CP_HDAC_INHIBITOR
Prazosin	Adrenergic receptor inhibitor	CP_BETA_ADRENERGIC_RECEPTOR_AGONIST
Rolipram	Phosphodiesterase-4 inhibitor	
Minocycline	NOS inhibitor	
Pioglitazone	PPARγ/α inhibitor	CP_PPAR_RECEPTOR_AGONIST
Fenofibrate	PPARα agonist	CP_PPAR_RECEPTOR_AGONIST

Table 6: Enrichment of strong connections in expected PCL annotations . p-values correspond to independent, two-sample Student's t-tests between "within-PCL" connectivities and a null model of randomly sampled compound connectivities (see text) for the same query drug, and are corrected for multiple testing using the Benjamini-Hochberg procedure. Effect size is the difference of means between those two groups, such that larger effect sizes correspond to higher expected connectivity scores between the query drug and members of its same drug class. Note that effect sizes are relatively small in most cases—this is due in part to the sparsity of connectivity scores.

Drug	PCL	p-value	Effect size
Topiramate	CP_SODIUM_CHANNEL_BLOCKER	**1.018e-31**	13.168
Vorinostat	CP_HDAC_INHIBITOR	**5.952e-22**	1.717
Sirolimus	CP_MTOR_INHIBITOR	**2.240e-17**	1.232
Eticlopride	CP_DOPAMINE_RECEPTOR_ANTAGONIST	**1.278e-11**	4.175
Olanzapine	CP_DOPAMINE_RECEPTOR_ANTAGONIST	**8.117e-09**	2.640
Fenofibrate	CP_PPAR_RECEPTOR_AGONIST	**1.012e-07**	1.775
Pioglitazone	CP_PPAR_RECEPTOR_AGONIST	**1.158e-07**	3.252
Manumycin a	CP_NFKB_PATHWAY_INHIBITOR	**4.124e-07**	5.983
Dexamethasone	CP_GLUCOCORTICOID_RECEPTOR_AGONIST	**2.741e-06**	2.462
Prazosin	CP_BETA_ADRENERGIC_RECEPTOR_AGONIST	**2.476e-02**	2.083
Acamprosate	CP_GABA_RECEPTOR_ANTAGONIST	**4.290e-02**	2.260
Mibefradil	CP_T_TYPE_CALCIUM_CHANNEL_BLOCKER	6.871e-02	0.355
1-EBIO	CP_POTASSIUM_CHANNEL_ACTIVATOR	2.573e-01	2.597
Fendiline	CP_CALCIUM_CHANNEL_BLOCKER	2.854e-01	2.636
Diltiazem	CP_CALCIUM_CHANNEL_BLOCKER	2.929e-01	5.719
Isradipine	CP_CALCIUM_CHANNEL_BLOCKER	4.062e-01	0.683
Nifedipine	CP_CALCIUM_CHANNEL_BLOCKER	4.100e-01	1.932
Mifepristone	CP_PROGESTERONE_RECEPTOR_ANTAGONIST	4.309e-01	3.160
Verapamil	CP_CALCIUM_CHANNEL_BLOCKER	5.404e-01	5.880
Ondansetron	CP_SEROTONIN_RECEPTOR_AGONIST	5.710e-01	2.659

Table 7: Correct PCL ranks aggregated by cell line. Mean rank percentile is the mean rank of the correct ("true") PCL, aggregated over all query drugs and divided by the total number of PCLs (92), reported by cell line.

CMap cell line	Mean rank percentile	FDR-corrected p-value
HA1E	0.326087	0.001663
A375	0.375000	0.004926
PC3	0.431522	0.109226
HCC515	0.446739	0.193877
HEPG2	0.461957	0.258068
MCF7	0.465217	0.279325
VCAP	0.492935	0.443995
A549	0.503804	0.468387
HT29	0.075445	0.591304

2.5.2. `VenomSeq` technical validation: Impact of reference cell lines and query drugs on expected PCL percentile ranks

Since IMR-32 cells are not present in the CMap reference dataset, we were particularly interested in seeing which cell lines present in the reference dataset (if any) performed better than others at the task of recovering expected connections. Using the PCL ranking strategy described in §7.7, 7 of the 9 core cell lines show at least a moderate tendancy to place the true PCL towards the front of the ranked list of all PCLs, indicating that at least some of the ability to recover expected connections is retained when looking at those 7 cell lines individually. PCL rankings stratified by drug (rather than cell line) show a similar pattern—15 of 20 PCL-annotated drugs tend to have the expected PCL ranked towards the front of the list ("enrichment"), while 5 tend to have the expected PCL show up towards the back of the list ("depletion"). Of these 20, the only It should be noted that—due to the rather small number of profiles in the reference dataset that are annotated to PCLs—these two analyses were limited in terms of statistical power, and deserve a follow up analysis in the future, when more PCLs and members of those PCLs are present in the reference database.

2.6. Associations between venoms and disease regulatory networks

Direct observations of expressed genes (via mRNA counts) provide an incomplete image of the regulatory mechanisms present in a cell. To complement the CMap approach that focuses on perturbations at the *gene* level, we designed a parallel approach that uses cell regulatory network data to investigate perturbations at the *regulatory module* (e.g., pathways and metabolic networks) level; an approach we refer to as *master regulator analysis*. In master regulator analysis, the ARACNe algorithm [Margolin et al., 2006] is used to obtain regulatory network data for our cell line of interest (in this case, IMR-32), consisting a list of *regulons*—overlapping sets of proteins whose expression is governed by a master regulator (e.g., a transcription factor). The msVIPER algorithm [Alvarez et al., 2016] is then used to determine the activity of each regulon by computing enrichment scores from observed expression levels of the genes/proteins contained in that regulon (here, using the RNA-Seq results described in §2.2).

We matched the significantly up- and down-regulated master regulators for each venom to diseases using high-confidence TF-disease associations in DisGeNET [Piñero et al., 2016]—a publicly available database of associations between diseases and gene network component. This approach is based on the idea that diseases caused by disregulation of metabolic and signaling networks can be treated by administering drugs that "reverse" the cause (i.e., abnormal master regulator activity) of disregulation. Since we are interested in discovering associations with multiple corroborating pieces of evidence, we specifically filtered for diseases where *two or more* linked TFs are disregulated when perturbed by the venom. The complete list of associations are provided on figshare at `https://doi.org/10.6084/m9.figshare.7609793`; here, we describe a handful of interesting observations.

The most prevalent class of illness (comprising 19.7% of all associations across all venoms) is `DISEASES OF THE NERVOUS SYSTEM AND SENSE ORGANS`. This is not surprising, considering many of the 25 venoms have neurotoxic effects, and IMR-32 is a cell line derived from neuroblast cells. One source of bias in these results is that similar diseases tend to be associated with the same regulatory mechanisms [Sun et al., 2011]. For example, associations between a venom and schizophrenia will often be co-reported with associations to other mental conditions, such as bipolar disorder and alcoholism.

3. Discussion

3.1. Venoms versus small-molecule drugs

In the connectivity analysis portion of `VenomSeq`, we demonstrated that these techniques have the ability to identify novel venom–drug class associations, and corroborate known venom activity. One distinct advantage of performing queries against the CMap reference dataset is their inclusion of manually-curated PCLs, which allow for normalization of data gathered from multiple perturbagens and multiple cell lines, aggregated at a class level that corresponds approximately with drug mode of action. For this reason, hypotheses generated by the connectivity analysis portion of `VenomSeq` are often testable at the protein level.

One important caveat is that venom components have a tendency to interact with cell surface receptors (e.g., ion channels or GPCRs), inciting various signaling cascades and therefore acting indirectly on downstream therapeutic targets. While this is certainly the case for many drugs as well (GPCRs are considered the most heavily investigated class of drug targets [Hopkins and Groom, 2002]), small molecules often can be designed to enter the cell and interact directly with the downstream therapeutic target. This has important implications regarding assay selection for *in vitro* validation of associations learned through the connectivity analysis. For example, if the MoA of interest is inhibition of an intracellular protein (e.g., topoisomerase), a cell-based assay should be considered when testing venom hypotheses, since the venom likely is not interacting directly with the topoisomerase (and, therefore, the effect would not occur in non-cell based assays).

3.2. Venoms versus human diseases

The master regulator analysis portion of `VenomSeq` discovers associations between venoms and the diseases they may be able to treat, rather than to drugs. This could be especially useful for discovering treatments to diseases with no or few existing indicated drugs (or drugs that are not present in public differential expression databases). Additionally, since the master regulator approach is sensitive to complex metabolic network relationships, it is (theoretically) more sensitive to patterns, as well as more suited to diseases with complex genetic etiologies that are not explainable by observed gene counts alone.

Currently, the primary drawback to the master regulator approach is that criteria for statistical significance are not well established. Therefore, it is challenging to determine which venom-disease associations are most likely to reflect actual therapeutic efficacy. As a temporary alternative, we used several heuristics to ensure there are multiple corroborating sources of evidence for the reported associations.

As discussed previously, the connectivity analysis produces hypotheses that are relatively straightforward to validate experimentally, using affordable, widely available assay kits and reagents. Since the master regulator workflow gives hypotheses at the disease level (where the underlying molecular etiologies can be unknown), validation instead needs to be performed at the *phenotype* level, either using animal models of disease, or carefully engineered, cell-based phenotypic assays that measure response at multiple points in disease-related metabolic pathways (e.g., DiscoverX's BioMAP® platform [Berg et al., 2003]).

3.3. Specific therapeutic hypotheses

`VenomSeq` contains multiple types of data analysis for two reasons: (1) This allows us to cover diseases with a wider array of molecular etiologies, and (2) it provides a means for obtaining multiple pieces of corroborating evidence for a given hypothesis. If a link between a venom and a drug/disease is suggested by both connectivity analysis and master regulator analysis, and there is additional literature evidence that lends biological or clinical plausibility, this increases our confidence that the suggested therapeutic effect is "real".

3.3.1. Argiope lobata venom versus cardiopulmonary and psychiatric diseases

A. lobata is a species of spider in the same genus as the common garden spider. The species is relatively understudied, largely due to its lack of interaction with humans, in spite of being distributed across Africa and much of Europe and Asia. The venom from species of *Argiope* spiders contain toxins known as *argiotoxins* [Poulsen et al., 2013], which are harmless to humans, in spite of having inhibitory effects on AMPA, NMDA, kainite, and nicotine acetylcholine receptors, which have been implicated in neurodegenerative and cardiac diseases. `VenomSeq` provides supporting evidence for therapeutic activity in each of these classes.

Connectivity analysis links *A. lobata* venom to ATPase inhibitor drugs (see **Figure 13**), which include digoxin, ouabain, cymarin, and other cardiac glycosides, and are used to treat a variety of heart conditions. Another venom-derived compound—bufalin (from the venom of toads in the genus *Bufo*) [Laursen et al., 2015]—is considered an ATPase inhibitor, and has demonstrated powerful cardiotonic effects. Connectivity analysis also links the venom to PPAR agonist drugs, which are used to treat cholesterol disorders, metabolic syndrome, and pulmonary inflammation. Interestingly, PPARγ activation results in cellular protection from NMDA toxicity. Given the known inhibitory effect of argiotoxins on NMDA receptors [Moe et al., 1998], this is striking and biologically plausible evidence for toxin synergism, where two or more venom components target multiple cellular structures with related functions in order to incite a more powerful response [Laustsen, 2016].

Master regulator analysis supports these findings, as well. We found that *A. lobata* venom is associated with a number of circulatory diseases, including hypertension, heart failure, cardiomegaly, myocardial ischemia, and others. Additionally, it reveals strong associations with an array of mental conditions, such as schizophrenia, bipolar disorder, and psychosis. These associations are supported by recent research into argiotoxins (and other polyamine toxins), showing that their affinity for iGlu receptors can be exploited to treat both psychiatric diseases and Alzheimer disease [Poulsen et al., 2013].

3.3.2. Scorpio maurus venom for cancer treatment via FGFR inhibition

S. maurus—the Israeli gold scorpion—is a species native to North Africa and the Middle East. Its venom is not harmful to humans, but it is known to contain a specific toxin, named maurotoxin, which blocks a number of types of voltage-gated potassium channels—an activity that is under investigation for treatment of gastrointestinal motility disorders [Beyder and Farrugia, 2012].

Our connectivity analysis suggests an additional association with FGFR inhibitor drugs. FGFR inhibitors are an emerging class of drugs with promising anticancer activity, and much research focused on them aims to understand and counteract their adverse effects (see **Fig-**

ure 14). Although there is no prior mention of FGFR-related activity from this or related species of scorpions, descriptions of unexpected side effects of *S. maurus* venom on mice provides evidence that such activity could be true. In particular, the venom has been shown to have biphasic effects on blood pressure: when injected, it causes rapid hypotension, followed by an extended period of hypertension. The fast hypotension is known to be caused by a phospholipase A_2 in the venom, but no known components elicit hypertension when administered in purified form [Ettinger et al., 2013]. The observed FGFR inhibitor-like effects on gene expression suggest that an unknown component (or group of components) may cause the hypertensive effect via FGFR inhibition. We are currently performing experimental validation of this link, and will report results in future revisions of this manuscript.

3.4. Accessing and querying VenomSeq data

`VenomSeq` is designed as a general and extensible platform for drug discovery, and we encourage secondary use of both the technology as well as the data produced using the 25 venoms tested on IMR-32 cells described in this manuscript. We maintain the data in two publicly-accessible locations: (1.) a “frozen” copy of the data, as it exists at the time of writing (on figshare, at `https://doi.org/10.6084/m9.figshare.7611662`), and (2.) a copy hosted on `venomkb.org`, available both graphically and programmatically, and designed to be expanded as new data and features are added to VenomKB.

3.5. Transitioning from venoms to venom components

`VenomSeq` is a technology for discovering early evidence that a *venom* has a certain therapeutic effect. However, most successful approved drugs derived from venoms make use of the activity of a single component within that venom, rather than the entire (crude) venom. As previously mentioned, venoms can be comprised of hundreds of unique components, each with a unique function and molecular target. We are in the early stages of applying `VenomSeq` individually to purified samples of each of the peptides from the venom of a snail in the family Terebridae. The goal of this project will be twofold: (1) To demonstrate the use of `VenomSeq` to screen individual venom components rather than crude venoms, and (2) to determine *which* of these venom components actually exerts transcriptomic effects on human cells. Each of these questions provides opportunities to understand better how specific venoms can cause therapeutic changes in human cells.

Even though most existing venom-derived drugs consist of a single component, crude venoms in nature use the synergistic effects of multiple components to cause specific phenotypic effects [Laustsen, 2016]. Therefore, testing each venom component individually using the `VenomSeq` workflow might fail to capture all of the clinically beneficial activities demonstrated by the crude venom. A brute-force solution is to perform `VenomSeq` on all combinations of the isolated venom components, but doing so requires a massive number of experiments ($2^n - 1$, where n is the number of components in the venom). Therefore, it will be necessary to establish a protocol for prioritizing combinations of venom components. One potential solution is to fractionate the venom (i.e., using gel filtration) and perform `VenomSeq` on combinations of the fractions, but this will need to be tested. Alternatively, integrative systems biology techniques could be used to predict which components act synergistically, via similarity to structures with well-established activities.

3.6. Applying the VenomSeq framework to other natural product classes

`VenomSeq` was, obviously, designed for the purpose of discovering therapeutic activities from venoms, but it could be feasibly extended to other types of natural products, including plant and bacterial metabolites, and immunologic components. Venoms provide a number of advantages and simplifying assumptions that were useful in designing the technology, but once `VenomSeq` becomes more proven it should be possible to relax these assumptions with some minor modifications to experimental protocol and data analysis. We foresee a few of these as the following:

- Venoms' targeted nature makes it easy to assume they will have some effect in animals; other natural products may be inert.
- Venom components are intentionally delivered as a mixture; other natural product mixtures might only be easy to collect as a mixture, in spite of unrelated biological activities.
- Venoms are usually soluble in water, while other natural products often are not.

- Non-venom toxins may have less-targeted MoAs, disrupting biological systems indiscriminantly (e.g., by interrupting cell membranes regardless of cell type).
- The kinetics of non-venom natural products may be more subtle than venoms, which tend to have powerful binding and catalytic protperties.

3.7. Interpreting connectivity analysis validation results

In §2.5, we described the results of the connectivity analysis procedure applied to PLATE-Seq expression data from IMR-32 cells treated with 37 existing drugs that have known effects, many of which are members of Connectivity Map perturbagen classes (PCLs). Since `VenomSeq` uses an expression analysis technology that is different from the Connectivity Map's L1000 platform, as well as a cell line that is not present in the Connectivity Map reference dataset, this is crucial for establishing that one can discover meaningful associations between crude venoms and profiles in the reference data within the `VenomSeq` framework.

Overall, the findings of our analysis are congruent with those made by the Connectivity Map team in [Subramanian et al., 2017]. Specifically, PCLs that affect highly conserved, core cellular functions (such as HDAC inhibitors, mTOR inhibitors, and PPAR receptors) tend to form strong connectivities with members of the same class regardless of cell line. Therefore, associations discovered between crude venoms and these drug classes are likely "true associations", even when using IMR-32 cells in the analysis. Furthermore, by virtue of leveraging data corresponding to drugs with known effects, but using a new cell line and different assay technology, we have made the following novel findings:

- Although IMR-32 is not present in the reference dataset, similarities between IMR-32 and cell lines that *are* present in the reference data can be leveraged to select reference expression profiles that are more likely to reproduce true associations. For example, HA1E and A375 cells produce expression profiles that form reasonably strong connectivities between IMR-32 query signatures and members of the same drug classes.
- More cell lines need to be included in the Connectivity Map data in order to better understand correlation structures in cell-specific expression, as well as to better capture therapeutic associations that are specific to cell types underrepresented in current datasets.
- Similarly, continued effort should be devoted to adding new PCL annotations. Currently, only 12.3% of compound signatures in the reference dataset are annotated to at least one PCL, and some PCLs contain only a few signatures. A more rigorous definition of what

specifically comprises a PCL would allow secondary research groups to contribute to this effort, ultimately improving the utility of the CMap data and increasing the sensitivity of the algorithms used to discover new putative therapeutic associations.

In spite of the large degree of corroborating evidence these results provide (e.g., every drug in our validation set produced a positive average effect on within-PCL connectivities versus corresponding null distributions), we cannot confidently predict that the associations discovered for crude venoms are true associations, rather than simply data artifacts. Although our confidence in the novel associations would be greatly improved by more PCL annotations to allow our analyses to attain greater statistical power, the ultimate test is to perform *in vitro* and (eventually) *in vivo* tests for these predicted therapeutic mechanisms of action. Aside from larger quantities of reference data against which to run the validation analyses, we also hope to employ other data science techniques involving network analysis and more advanced applications of master regulator analysis (see, e.g., §2.6) to further understand the dynamic interactions between cell types, gene expression, and perturbational signals that underly therapeutic processes.

4. Conclusions

Venoms provide an immensely valuable opportunity for drug discovery, but it has become necessary to revise the techniques used for identifying new therapeutic activities. Traditional methods—involving rigorous experimental validation and high cost—are still necessary for establishing whether associations between venoms and therapeutic effects actually work in living systems, but data-driven computational approaches stand ready to make this process easier by generating new hypotheses backed by existing evidence and multiple levels of statistical validation. `VenomSeq` is an early example of these.

`VenomSeq` takes a two-pronged approach, combining connectivity analysis and master regulator analysis to provide two orthogonal views of the effects venoms have on human cells, where likely therapeutic effects are validated using publicly available knowledge representations and databases. In this study, we tested the venomseq workflow on 25 diverse venoms applied to human IMR-32 cells, and discovered a number of new therapeutic hypotheses supported by existing literature evidence.

To reinforce the validity of the hypotheses found by `VenomSeq`, we will need to apply the pipeline to new venoms and new human cell lines, and to test the pipeline on venoms, venom fractions, and isolated venom components with well-understood therapeutic modes of action. Like described previously, we are in the process of conducting follow-up validation assays to test specific hypotheses learned via the connectivity analysis, the results of which will be included in a future revision of this manuscript.

5. Supplemental Materials

All relevant supplemental data and materials are available in a .zip archive accompanying this manuscript. Additional figures and tables are available in the appendices of this manuscript, as referred to throughout the text.

6. Acknowledgements and funding

This work is supported by the National Institute of General Medical Sciences (R01 GM107145; PI: Tatonetti).

7. Methods

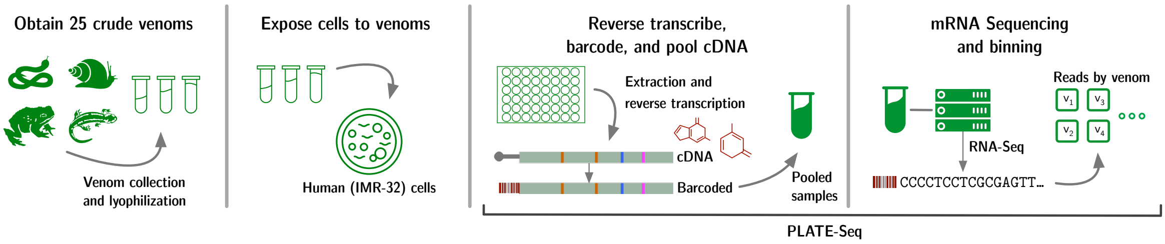


Figure 6: RNA-Seq strategy for `VenomSeq`. Crude venoms are extracted and lyophilized. IMR-32 cells in culture are then treated with predetermined dosages of reconstituted venoms, and the PLATE-Seq method [Bush et al., 2017] is used to isolate, sequence, and count reads corresonding to cellular mRNA.

Table 8: 25 venoms used to validate the VenomSeq workflow. Numbers in the right column are used as placeholder names for the venoms in data files.

Species name	Common name	Venom number
Naja nivea	Cape cobra	1
Laticauda colubrina	Banded sea krait	2
Montivipera xanthina	Ottoman viper	3
Dendroaspis polylepis polylepis	Black mamba	4
Crotalus scutulatus scutulatus	Mojave rattlesnake	5
Atractaspis sp.	Burrowing asp	6
Macrothele gigas	Japanese funnel web spider	7
Linothele fallax	Tiger spider	8
Poecilotheria fasciata	Sri Lanka ornamental spider	9
Argiope lobata	—	10
Synanceia verrucosa	Reef stonefish	11
Synanceia horrida	Estuarine stonefish	12
Buthus occitanus	Common yellow scorpion	13
Leiurus quinquestriatus	Deathstalker	14
Scorpio maurus	Large-clawed scorpion	15
Bufo bufo	Common toad	16
Rhinella marina	Cane toad	17
Bombina variegata	Yellow-bellied toad	18
Apis mellifera	Western honey bee	19
Vespa crabro	European hornet	20
Scolopendra subspinipes dehaani	Vietnamese centipede	21
Conus marmoreus	Marbled cone snail	22
Conus imperialis	Imperial cone snail	23
Octopus macropus	Atlantic white-spotted octopus	24
Pterois volitans	Red lionfish	25

7.1. Reagents and materials

We performed growth inhibition assays and perturbation experiments using IMR-32 cells—an adherent, metastatic neuroblastoma cell line used in previous applications of PLATE-Seq and VIPER—grown in FBS-supplemented Eagle's Minimum Essential Medium (EMEM). All venoms were provided in lyophilized form and stored at -20 C. Since venoms naturally exist in aqueous solution, we reconstituted them in ddH_2O at ambient temperature.

7.2. Obtaining 25 venoms

`VenomSeq` is designed to apply to all venomous species across all taxonomic clades. Accordingly, we validated the workflow using 25 venoms sampled from a diverse range of species distributed across the tree of life. We selected the 25 species based on availability and compliance with international law, and sought to balance maximal cladistic diversity with minimal expected cytotoxicity (e.g., snakes in the genus *Bitis* are known for inducing tissue death and necrosis, and are therefore challenging to use for drug discovery applications [Ponte et al., 2010]). We purchased the 25 venoms from Alpha Biotoxine in lyophilized form, and obtained prior approval from the US Centers for Disease Control (CDC) through the Federal Select Agent Program [Gonder, 2005] for importing venoms containing α-conotoxins. The 25 venoms we selected are shown in **Table 8**. Note that we assigned a numeric identifier to each venom for convenience—these numbers show up numerous places in the data for `VenomSeq`. We also have included a rooted cladogram of the 25 species in **Figure 7**.

7.3. Growth inhibition assays

A major challenge in generating differential gene expression data for discovery purposes is finding appropriate dosages for the compounds being tested. This is done to ensure the compound is in sufficient concentration to be exerting an observable effect on the cells, while also mitigating processes that result from toxicity (e.g., apoptosis). In practice, determining an appropriate dosage concentration usually makes use of previous experimental evidence and/or biochemical constants, but since these are generally not available for crude venoms, we instead determined dosages based on growth inhibition.

We prepared 2-fold serial dilutions of each venom, starting from $2.0\,\mathrm{mg}\,\mu\mathrm{l}^{-1}$. We seeded 96-well plates with IMR-32 cells and exposed them to the serial dilutions of the venoms after 24 hours of incubation. 48 hours after exposure, we quantified growth inhibition of the IMR-32 cells via cell viability luminesence assays.

For each venom, we fit these data to the Hill equation:

$$y = \mathrm{Bottom} + \frac{(\mathrm{Top} - \mathrm{Bottom})}{1 + 10^{(\log \mathrm{GI}_{50} - x) \times h}}$$

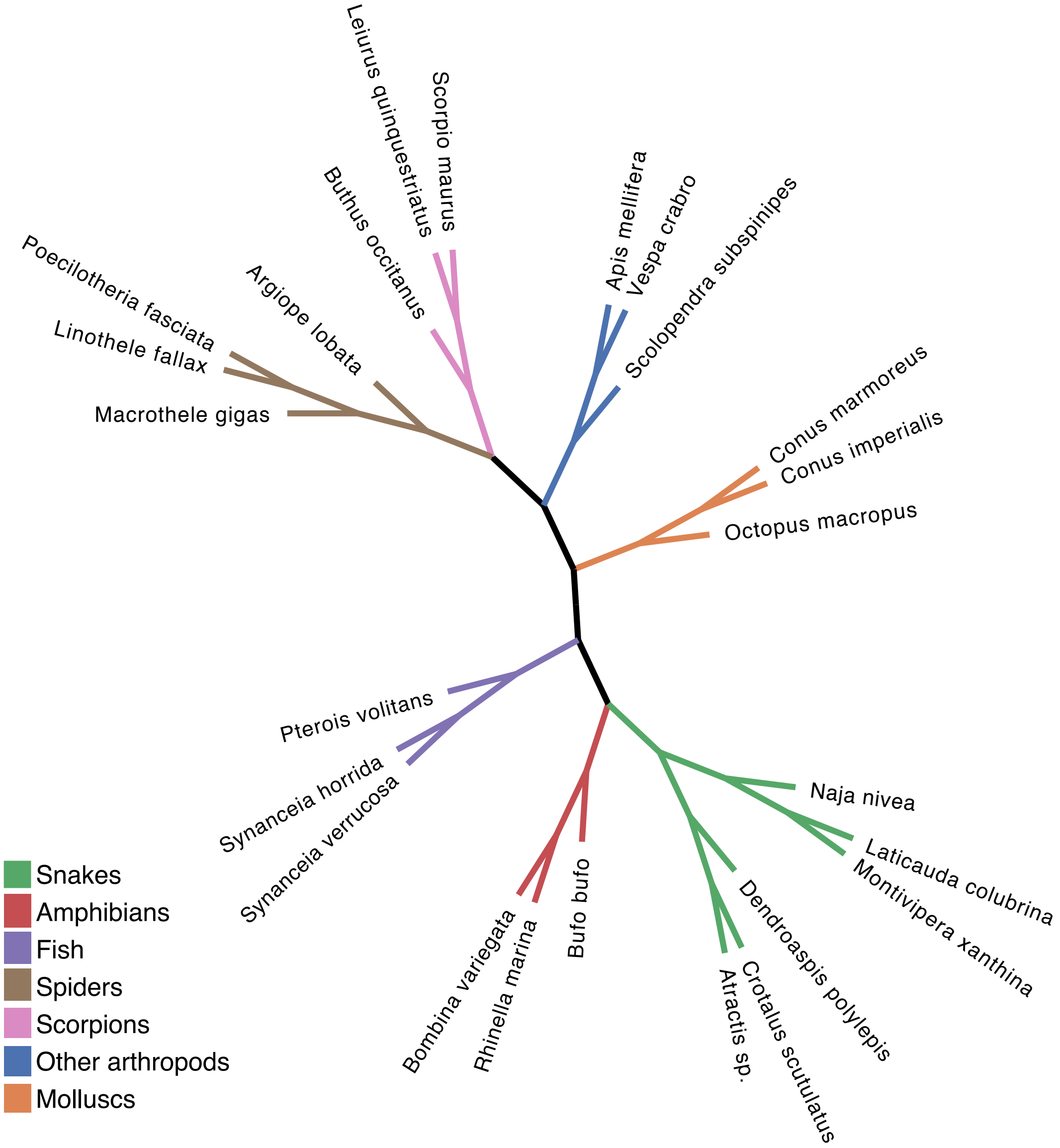


Figure 7: Rooted cladogram showing the 25 species used in `VenomSeq`. Clades corresponding to major taxonomic groups are labeled as indicated.

where x is venom concentration, y is response (i.e., percent growth compared to untreated cells), Top and Bottom are the maximum and minimum values of y, respectively, and h is a constant that controls the shape of the sigmoidal curve. We used the resulting GI_{20} values (i.e., the value of x such that $y = 100\% - 20\% = 80\%$) as the venom exposure concentrations for the following sequencing experiments. Since some of the curves had very steep slopes (indicating rapid loss of total cell viability after miniscule changes in venom concentration), we confirmed the accuracy of the GI_{20} concentrations via secondary viability assays using the exact GI_{20} values extrapolated from the growth inhibition curves.

7.4. mRNA Sequencing

We prepared samples of human IMR-32 cells in 96-well cell culture plates, allowing for 3 replicates at each of 3 time points (6, 24, and 36 hours post-treatment) for each of the 25 venoms. The layout of the samples across 2 96-well plates is available in **Appendix A**. We reconstituted the crude venoms in water, and treated the samples with corresponding venoms at the previously determined GI_{20} values. We additionally prepared 12 control samples treated with water only, and 9 control samples that were untreated. Following total mRNA extraction, we carried out the PLATE-Seq protocol [Bush et al., 2017] to obtain gene counts for each sample. All sequencing was performed on the Illumina HiSeq platform. We used STAR [Dobin et al., 2013] to (1) map the demultiplexed reads to the human genome (build GRCh38 [Schneider et al., 2017]) and (2) count the reads mapping to known genes. For detailed quality control data for the sequencing experiments, refer to **Appendix A**.

7.5. Constructing expression signatures

We constructed differential gene expression signatures using the DESeq2 [Love et al., 2014] library for the R programming language. DESeq2 fits observed counts for each gene to a negative binomial distribution with mean μ_{ij} and dispersion (variance) α_i, which we find to be a more robust model than traditional approaches based on the Poisson distribution (i.e., by allowing for unequal means and dispersions). In practice, users can substitute any method for determining significantly up- and down-regulated genes from count data. We filtered for genes

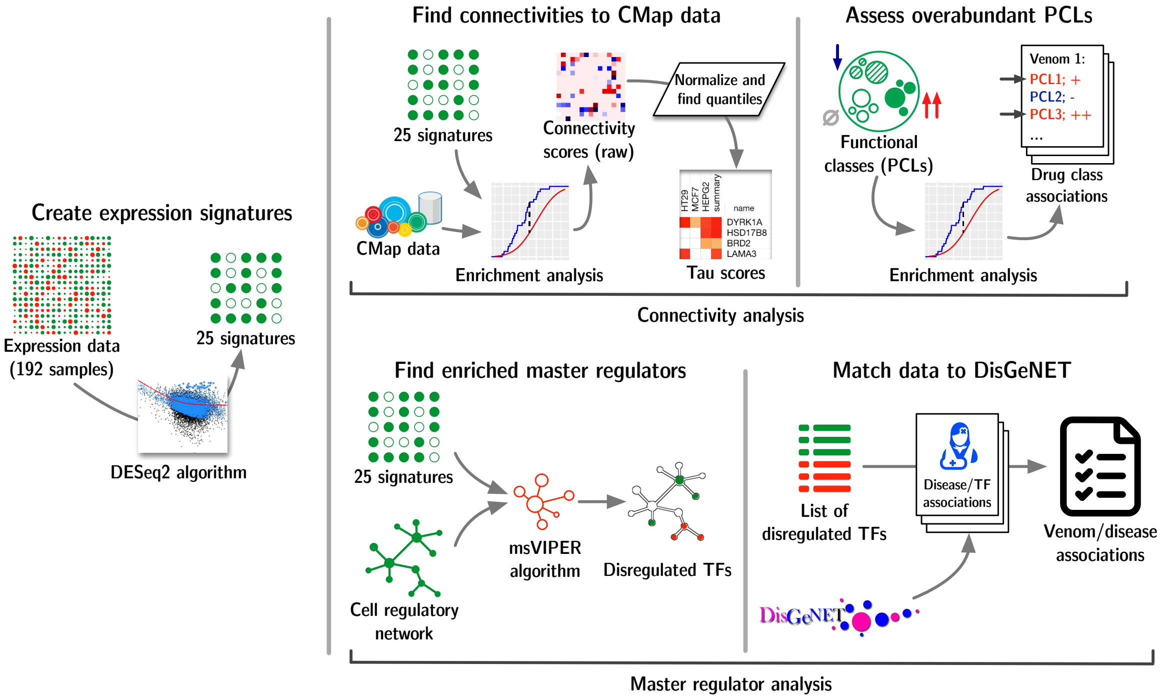


Figure 8: Strategy for discovering new associations from VenomSeq data. After obtaining processed gene counts per sample, we generated differential expression signatures for each venom, and then used the signatures in two parallel analyses: connectivity analysis, and master regulator analysis.

with an FDR-corrected p-value < 0.05, and recorded their respective mean $\log_2$-fold change values, noting whether expression increased (up-regulated) or decreased (down-regulated).

7.6. Comparing venoms to known drugs and diseases

7.6.1. Comparing to known drugs using the Connectivity Map

We retrieved the most recently published Connectivity Map dataset from the Clue.io Data Library (GSE92742), which contains 473,647 perturbational signatures, each consisting of robust Z-scores for 12,328 genes, along with relevant metadata. We then used the procedure described by the Connectivity Map team [Subramanian et al., 2017] to generate connectivity scores between each of the VenomSeq gene expression signatures and each of the reference expression profiles in the Connectivity Map database. This procedure, adapted for VenomSeq, is summarized below.

Let a query q_i be the two lists of up- and down-regulated genes corresponding to the differential expression signature for venom i, and $\mathbf{r}_j \in \mathbf{R}$ be a vector of gene-wise Z-scores in reference expression signature j. We first generate a *Weighted Connectivity Score* (WCT) (or *Raw Connectivity Score*) between q_i and $\mathbf{r}_j$:

$$w_{qr} = \begin{cases} (ES_{\text{up}}^{q,r} - ES_{\text{down}}^{q,r})/2 & \text{if } \operatorname{sgn}(ES_{\text{up}}^{q,r}) \neq \operatorname{sgn}(ES_{\text{down}}^{q,r}) \\ 0 & \text{otherwise} \end{cases}$$

where sgn denotes the sign function $\frac{d}{dx}|x|$, and $ES^{qr}_{\cdot}$ is the signed enrichment score for either the up- or down-regulated genes in the signature, calculated separately (see Appendix 7.6.1 for details).

Although we validated `VenomSeq` on only a single human cell line, the reference database provided by the Connectivity Map provides expression profiles on 9 core cell lines, across multiple classes of perturbagens. Therefore, we compute normalized versions of WCS called Normalized Connectivity Scores (NCSs):

$$NCS_{q,r} = \begin{cases} w_{q,r}/\mu_{c,t}^{+} & \text{if } \operatorname{sgn}(w_{q,r}) > 0 \\ w_{q,r}/\mu_{c,t}^{-} & \text{otherwise} \end{cases}$$

where $\mu_{c,t}^{+}$ and $\mu_{c,t}^{-}$ are the means of all positive or negative WCTs (respectively) for the given cell line and perturbagen type.

The final step in computing connectivity scores between a venom q and a reference r is to convert $NCS_{q,r}$ into a value named τ, which represents the signed quantile score in the context of all positive or negative NCSs:

$$\tau_{q,r} = \operatorname{sgn}(NCS_{q,r}) \frac{100}{N} \sum_{i=1}^{N} [|NCS_{i,r}| < |NCS_{i,r}|]$$

where N is the number of all expression signatures in the reference database and $|NCS|$ is the absolute magnitude of an NCS.

Enrichment Score computation For a venom q and reference expression signature r, the enrichment score $ES_{\cdot}^{qr}$ is a signed Kolmogorov–Smirnov-like statistic indicating whether the subset of up- or down-regulated genes in q tend to occur towards the beginning or the end of a list of all genes ranked by expression level in r. We follow a procedure similar to that described by Lamb *et al.* in [Lamb et al., 2006]. Specifically, we compute the following two values:

$$a = \max_{j=1}^{t} \left[\frac{j}{t} - \frac{\mathbf{V}_{qr}(j)}{n} \right]$$

$$b = \max_{j=1}^{t} \left[\frac{\mathbf{V}_{qr}(j)}{n} - \frac{(j-1)}{t} \right]$$

where $\mathbf{V}_{qr}$ is the vector of nonnegative integers that gives the indexes of the genes in q within the list of all genes ordered corresponding to their assumed values in r, t is the number of genes in q, and n is the number of genes reported in the reference database (in practice, $t \ll n$). We then set ES as follows:

$$ES_{\cdot}^{qr} = \begin{cases} a & \text{if } a > b \\ -b & \text{if } a < b \end{cases}$$

Since each query q consists of two lists—one of up-regulated and one of down-regulated genes—we compute both ES_{up}^{qr} and ES_{down}^{qr}, respectively, and use these two values to compute w_{qr}, as described above.

7.6.2. Comparing to known diseases using master regulator analysis

We discovered associations between the venom expression profiles and known diseases (coded as UMLS concept IDs) as the result of two sequential steps: (1) algorithmic determination of substantially perturbed cell regulatory modules (called *regulons*), and (2) mapping master regulators to diseases using high-confidence associations distributed in the DisGeNET database. These took as input the same differential expression data used in the connectivity analysis. IMR-32 regulon data (in the form of an adjacency matrix, where nodes are genes and edges are measures of mutual information with respect to their coexpression) were provided by the authors of the ARACNe algorithm.

In order to identify perturbed regulons, we first performed a 2-tailed Student's t-test between the genes' expression in the 'test' set (samples perturbed by venoms) and the 'reference' set (control samples). To make the final expression signatures, we then converted the results of the t-tests to Z-scores, to make them consistent with the models used by downstream algorithms. We generated null scores by performing the same test on the expression data with permuted sample labels, to account for correlation structures between genes. Once we had computed Z-scores, we ran the msVIPER algorithm, which derives enrichment statistics for each regulon based on the expression levels of the genes contained in the regulon. The result of msVIPER is a table of regulons (labeled by their master regulator), with enrichment scores, p-values, and FDR-corrected adjusted p-values.

We then compared the significantly upregulated regulons to the manually curated subset of TF–disease associations from the DisGeNET database. To do so, we mapped the statistically significant master regulator TFs for each venom to TFs reported in DisGeNET, and then mapped those TFs to their associated diseases. To help with filtering venom–disease associations with low evidence, we only retained diseases where *at least two* of the regulons that were significantly disregulated by the venom are associated with the same disease. Accordingly, we considered diseases with the highest number of significantly disregulated master regulators to comprise the associations with the greatest amount of evidence.

Similarly to how we mapped drugs to drug classes, we mapped diseases to disease categories. To do so, we identified the set of ICD-9 codes for each disease, based on the diseases' entries in the UMLS (UMLS CUIs were provided by DisGeNET). We then identified the disease category as the top-level ICD-9 'chapter' corresponding to that ICD-9 code (e.g., `NEOPLASMS`, `MENTAL DISORDERS`, `DISEASES OF THE RESPIRATORY SYSTEM`, etc.). In rare instances where a disease or condition was present in two locations (e.g., 'hypertension' is found in 2 chapters: `DISEASES OF THE CIRCULATORY SYSTEM` (401), and `INJURY AND POISONING` (997.91)), we opted for the more specific of the two (e.g., avoiding entries containing "not elsewhere classified").

7.7. Assessing sequencing technology and cell type compatibility

Since `VenomSeq` uses a sequencing technology (PLATE-Seq) and a cell line (IMR-32) that have not been used previously with the connectivity analysis approach, we evaluated their compatibility using a secondary dataset consisting of IMR-32 cells perturbed with 37 drugs and sequenced using PLATE-Seq. Since these drugs have known effects—and since many are present in the L1000 reference dataset—we sought to determine the extent to which connectivity analysis captures functional similarities between these drug data and the L1000 reference expression profiles. The 37 drugs are listed in **Table 5**. For the purposes of this discussion, a "query signature" is an expression signature corresponding to one of the 37 drugs in the validation dataset, and a "reference profile" is an L1000 expression profile from the dataset (GSE92742) published by the Connectivity Map team and used in the crude venom connectivity analysis.

Using these data (consisting of gene count matrices with several technical replicates per drug), we constructed differential expression signatures and performed the connectivity analysis algorithm in the same manner as we had for IMR-32 cells exposed to the 25 crude venoms. We annotated each of the 37 drugs (where possible) with perturbagen classes (PCLs) defined by the Connectivity Map team, which allowed us to identify L1000 expression profiles that come from the same drug classes as the drugs in our validation dataset. We then evaluated connectivity scores among members of the same PCL from two perspectives: (1) By aggregating all τ scores for reference profiles corresponding to a given compound, integrating evidence from all cell lines, and (2) by aggregating τ scores within individual cell lines, allowing us to assess the degrees to which specific cell lines are compatible with IMR-32/PLATE-Seq query signatures.

For the first of these two approaches, we collected all values of τ connecting query signatures in a PCL to reference profiles in the same PCL, and constructed null models by retrieving τ scores between the same query signature and all reference profiles that are members of any PCL. We defined the "effect size" of each PCL annotation as the difference of the mean of the scores within the true PCL and the mean of the scores in the null model. Additionally, we determined statistical significance using independent two-sample Student's t-tests. To correct for multiple testing, we adjusted p-values using the Benjamini-Hochberg procedure ($\alpha = 0.05$).

For the second approach—in which we evaluated each of the 9 core L1000 cell lines separately for each query signature—we retrieved τ scores between query signatures and each of the 92 PCLs in the reference dataset. Then, for each of the 9 cell lines and each of the query signatures annotated to a PCL, we constructed ordered lists of all PCLs ranked by their mean τ score in descending order (highest to lowest connectivity). In each of those lists, we determined the rank corresponding to the expected ("true") PCL—which we call the *rank percentiles*—and aggregated these ranks separately by (a) the drug corresponding to the query signature and (b) cell line of the reference profile. These two strategies allow us to separately assess the effects of *drugs* and *cell lines* on the behavior of connectivity scores. Under the null hypothesis that there is no selective preference for the true PCL in the connectivity data, the mean rank percentiles would follow a continuous uniform distribution in the range $[0, 1]$. Alternatively, if there is a selective preference for the expected PCL in the connectivity data, this rank will tend to occur towards the front of the list of ranks (and vice-versa).

References

[Alvarez et al., 2016] Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H., and Califano, A. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature genetics*, 48(8):838.

[Berg et al., 2003] Berg, E. L., Butcher, E. C., and Melrose, J. (2003). Biomap characterization of biologically active agents. US Patent 6,656,695.

[Beyder and Farrugia, 2012] Beyder, A. and Farrugia, G. (2012). Targeting ion channels for the treatment of gastrointestinal motility disorders. *Therapeutic advances in gastroenterology*, 5(1):5–21.

[Bush et al., 2017] Bush, E. C., Ray, F., Alvarez, M. J., Realubit, R., Li, H., Karan, C., Califano, A., and Sims, P. A. (2017). Plate-seq for genome-wide regulatory network analysis of high-throughput screens. *Nature communications*, 8(1):105.

[Calvete et al., 2009] Calvete, J. J., Sanz, L., Angulo, Y., Lomonte, B., and Gutiérrez, J. M. (2009). Venoms, venomics, antivenomics. *FEBS letters*, 583(11):1736–1743.

[de Bono et al., 2007] de Bono, B., Rothfels, K., Castagnoli, L., Williams, M., and Jassal, B. (2007). Signaling by fgfr [homo sapiens]. *Reactome*.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

[Ettinger et al., 2013] Ettinger, K., Cohen, G., Momic, T., and Lazarovici, P. (2013). The effects of a chactoid scorpion venom and its purified toxins on rat blood pressure and mast cells histamine release. *Toxins*, 5(8):1332–1342.

[Fabregat et al., 2015] Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2015). The reactome pathway knowledgebase. *Nucleic acids research*, 44(D1):D481–D487.

[Gonder, 2005] Gonder, J. C. (2005). Select agent regulations. *ILAR journal*, 46(1):4–7.

[Hopkins and Groom, 2002] Hopkins, A. L. and Groom, C. R. (2002). The druggable genome. *Nature reviews Drug discovery*, 1(9):727.

[Kühlbrandt, 2004] Kühlbrandt, W. (2004). Biology, structure and mechanism of p-type atpases. *Nature reviews Molecular cell biology*, 5(4):282.

[Lamb et al., 2006] Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935.

[Laursen et al., 2015] Laursen, M., Gregersen, J. L., Yatime, L., Nissen, P., and Fedosova, N. U. (2015). Structures and characterization of digoxin-and bufalin-bound na+, k+-atpase compared with the ouabain-bound complex. *Proceedings of the National Academy of Sciences*, 112(6):1755–1760.

[Laustsen, 2016] Laustsen, A. H. (2016). Toxin synergism in snake venoms. *Toxin Reviews*, 35(3-4):165–170.

[Lewis and Garcia, 2003] Lewis, R. J. and Garcia, M. L. (2003). Therapeutic potential of venom peptides. *Nature reviews drug discovery*, 2(10):790.

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

[Margolin et al., 2006] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(1):S7.

[Moe et al., 1998] Moe, S. T., Smith, D. L., Chien, Y. E., Raszkiewicz, J. L., Artman, L. D., and Mueller, A. L. (1998). Design, synthesis, and biological evaluation of spider toxin (argiotoxin-636) analogs as nmda receptor antagonists. *Pharmaceutical research*, 15(1):31–38.

[Pennington et al., 2018] Pennington, M. W., Czerwinski, A., and Norton, R. S. (2018). Peptide therapeutics from venom: Current status and potential. *Bioorganic & medicinal chemistry*, 26(10):2738–2758.

[Piñero et al., 2016] Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L. I. (2016). Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, page gkw943.

[Ponte et al., 2010] Ponte, C. G., Nóbrega, E. L., Fernandes, V. C., da Silva, W. D., and Suarez-Kurtz, G. (2010). Inhibition of the myotoxic activities of three african bitis venoms (b. rhinoceros, b. arietans and b. nasicornis) by a polyvalent antivenom. *Toxicon*, 55(2-3):536–540.

[Poulsen et al., 2013] Poulsen, M. H., Lucas, S., Bach, T. B., Barslund, A. F., Wenzler, C., Jensen, C. B., Kristensen, A. S., and Strømgaard, K. (2013). Structure–activity relationship studies of argiotoxins: selective and potent inhibitors of ionotropic glutamate receptors. *Journal of medicinal chemistry*, 56(3):1171–1181.

[Schneider et al., 2017] Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., et al.

(2017). Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*.

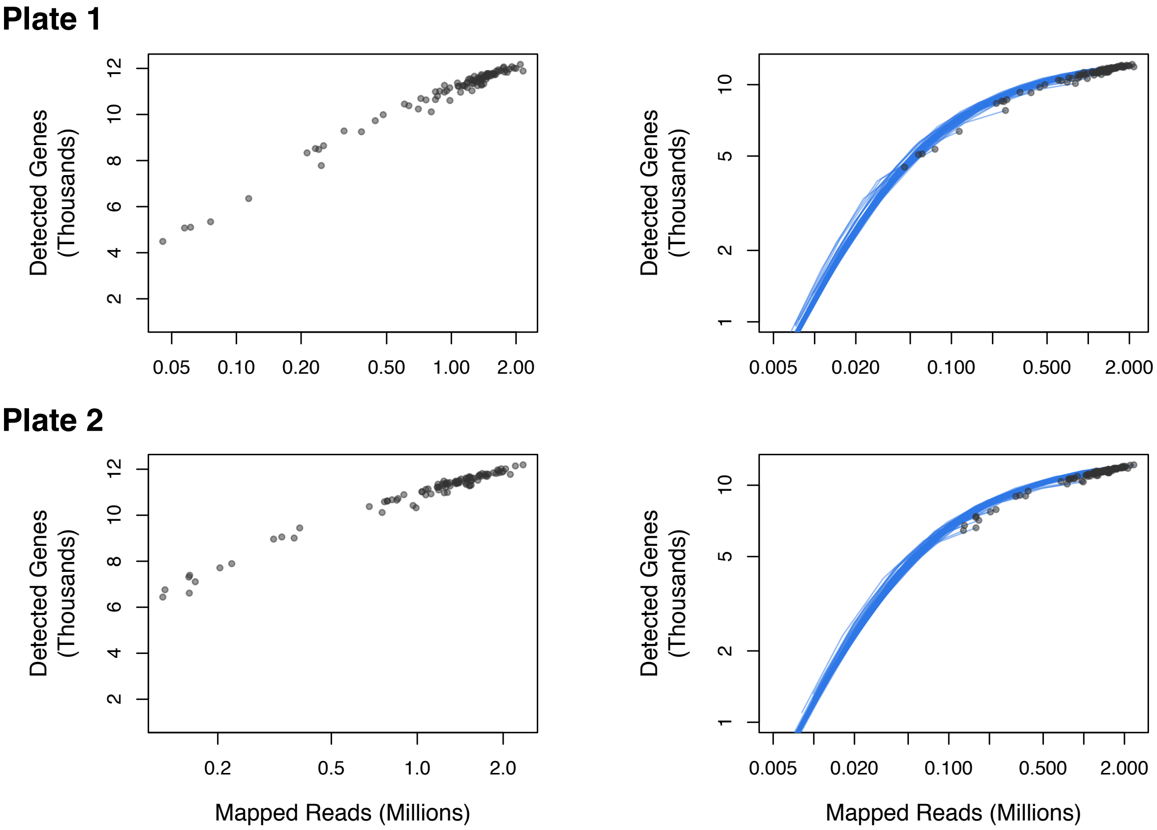
[Subramanian et al., 2017] Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.

[Sun et al., 2011] Sun, P. G., Gao, L., and Han, S. (2011). Prediction of human disease-related gene clusters by clustering analysis. *International journal of biological sciences*, 7(1):61.

[Terlau and Olivera, 2004] Terlau, H. and Olivera, B. M. (2004). Conus venoms: a rich source of novel ion channel-targeted peptides. *Physiological reviews*, 84(1):41–68.

[von Reumont et al., 2014] von Reumont, B., Campbell, L., and Jenner, R. (2014). Quo vadis venomics? a roadmap to neglected venomous invertebrates. *Toxins*, 6(12):3488–3551.

Appendix A PLATE-Seq quality control data



(a) Library Complexity

(b) Saturation Analysis

Figure 9: Quality control plots. (a.) Number of detected genes (mapped reads ≥ 2) as a function of the total number of mapped reads per sample. (b.) Saturation analysis by *in silico* subsampling. Original data points are indicated by the black dots.

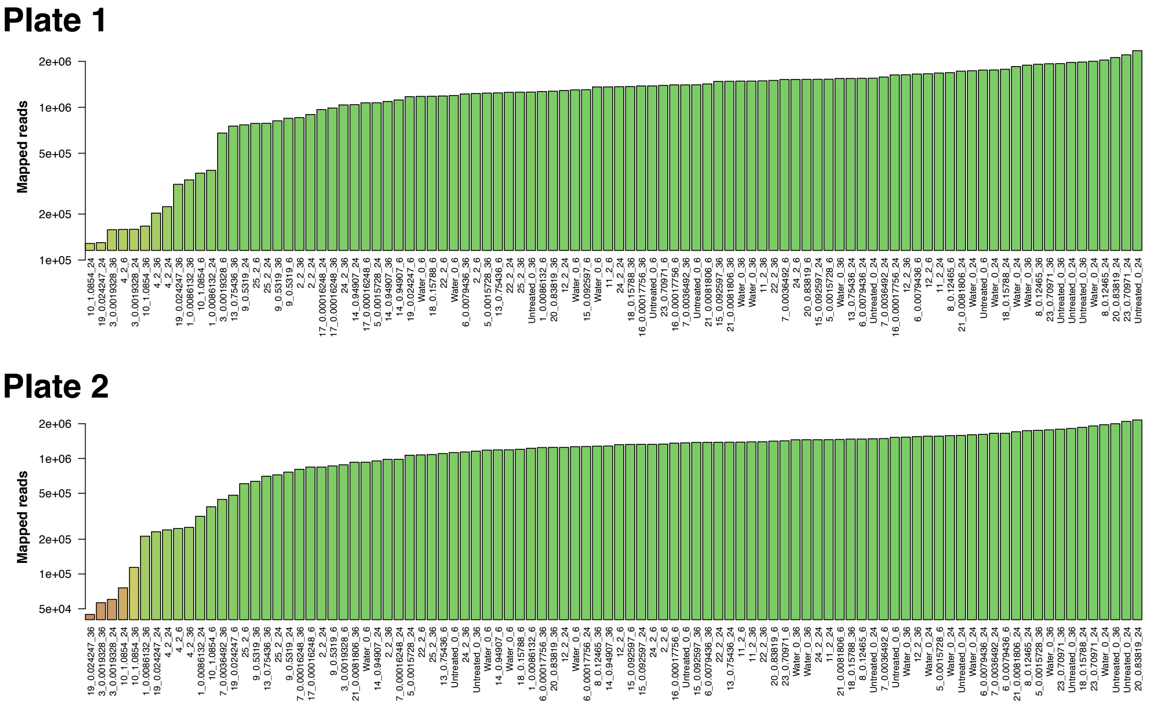


Figure 10: Barplot showing the number of mapped reads per sample.

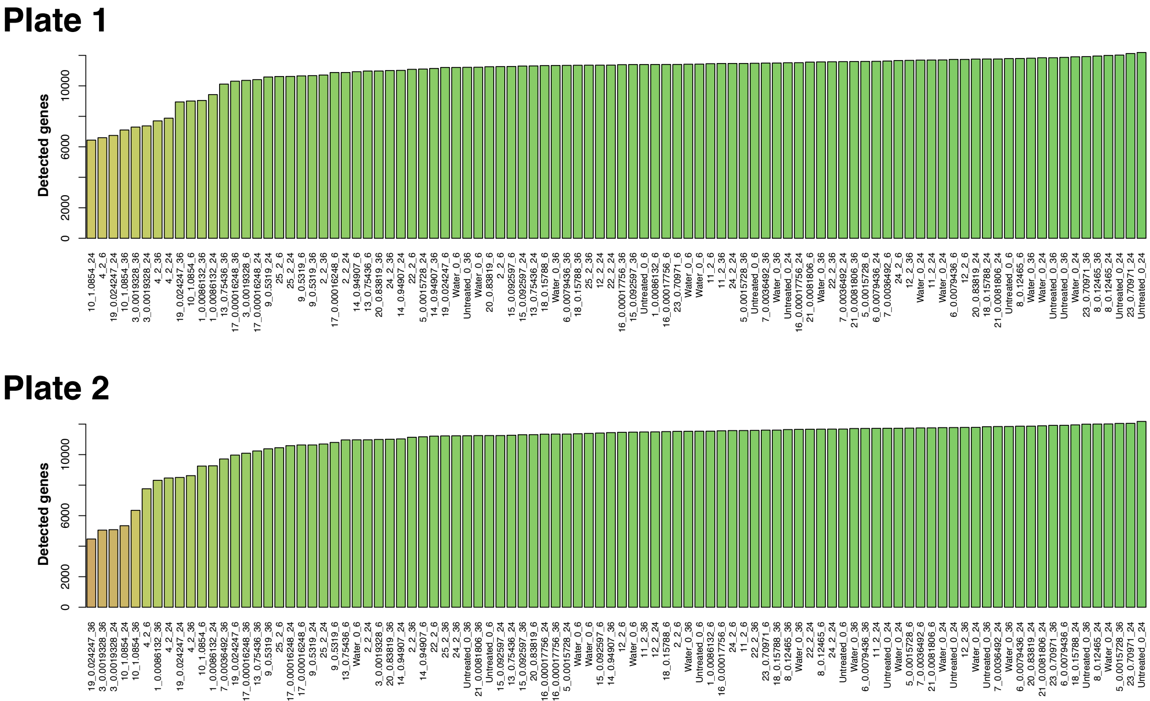


Figure 11: Barplot showing the number of detected genes per sample.

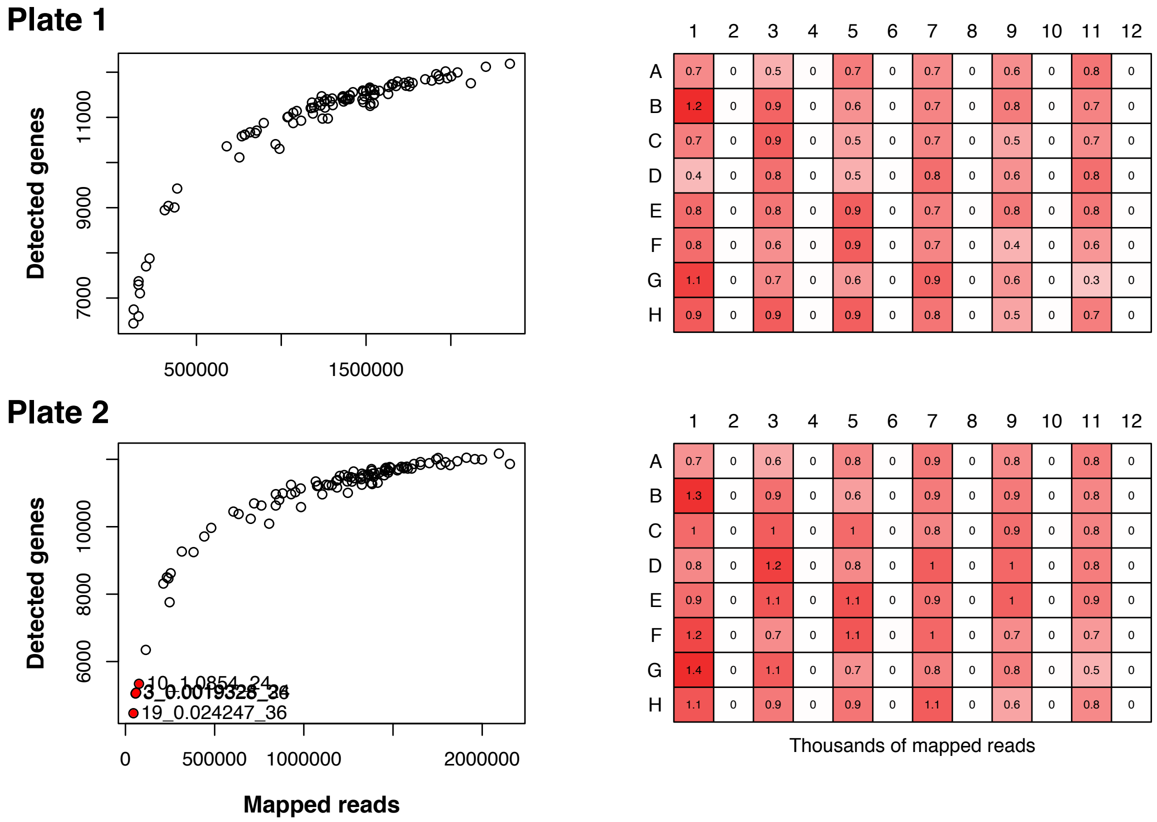
Plate 1:

Well	Venom	Conc. (uG/uL)	Time (Hrs)
A1	1	0.008 6	6
B1	2	2.000 0	6
C1	3	0.001 9	6
D1	4	2.000 0	6
E1	5	0.001 6	6
F1	6	0.007 9	6
G1	7	0.003 6	6
H1	8	0.124 7	6
A2	9	0.531 9	6
B2	10	1.085 4	6
C2	11	2.000 0	6
D2	12	2.000 0	6
E2	13	0.754 4	6
F2	14	0.949 1	6
G2	15	0.092 6	6
H2	16	0.000 2	6
A3	17	0.000 2	6
B3	18	0.157 9	6
C3	19	0.024 2	6
D3	20	0.838 2	6
E3	21	0.008 2	6
F3	22	2.000 0	6
G3	23	0.709 7	6
H3	24	2.000 0	6
A4	25	2.000 0	6
B4	Water	–	6
C4	Water	–	6
D4	Water	–	6
E4	Water	–	6
F4	Untreated	–	6
G4	Untreated	–	6
H4	Untreated	–	6
A5	1	0.008 6	24
B5	2	2.000 0	24
C5	3	0.001 9	24
D5	4	2.000 0	24
E5	5	0.001 6	24
F5	6	0.007 9	24
G5	7	0.003 6	24
H5	8	0.124 7	24
A6	9	0.531 9	24
B6	10	1.085 4	24
C6	11	2.000 0	24
D6	12	2.000 0	24
E6	13	0.754 4	24
F6	14	0.949 1	24
G6	15	0.092 6	24
H6	16	0.000 2	24
A7	17	0.000 2	24
B7	18	0.157 9	24
C7	19	0.024 2	24
D7	20	0.838 2	24
E7	21	0.008 2	24
F7	22	2.000 0	24
G7	23	0.709 7	24
H7	24	2.000 0	24
A8	25	2.000 0	24
B8	Water	–	24
C8	Water	–	24
D8	Water	–	24
E8	Water	–	24
F8	Untreated	–	24
G8	Untreated	–	24
H8	Untreated	–	24
A9	1	0.008 6	36
B9	2	2.000 0	36
C9	3	0.001 9	36
D9	4	2.000 0	36
E9	5	0.001 6	36
F9	6	0.007 9	36
G9	7	0.003 6	36
H9	8	0.124 7	36
A10	9	0.531 9	36
B10	10	1.085 4	36
C10	11	2.000 0	36
D10	12	2.000 0	36
E10	13	0.754 4	36
F10	14	0.949 1	36
G10	15	0.092 6	36
H10	16	0.000 2	36
A11	17	0.000 2	36
B11	18	0.157 9	36
C11	19	0.024 2	36
D11	20	0.838 2	36
E11	21	0.008 2	36
F11	22	2.000 0	36
G11	23	0.709 7	36
H11	24	2.000 0	36
A12	25	2.000 0	36
B12	Water	–	36
C12	Water	–	36
D12	Water	–	36
E12	Water	–	36
F12	Untreated	–	36
G12	Untreated	–	36
H12	Untreated	–	36

Plate 2:

Well	Venom	Conc. (uG/uL)	Time (Hrs)
A1	1	0.008 6	6
B1	2	2.000 0	6
C1	3	0.001 9	6
D1	4	2.000 0	6
E1	5	0.001 6	6
F1	6	0.007 9	6
G1	7	0.003 6	6
H1	8	0.124 7	6
A2	9	0.531 9	6
B2	10	1.085 4	6
C2	11	2.000 0	6
D2	12	2.000 0	6
E2	13	0.754 4	6
F2	14	0.949 1	6
G2	15	0.092 6	6
H2	16	0.000 2	6
A3	17	0.000 2	6
B3	18	0.157 9	6
C3	19	0.024 2	6
D3	20	0.838 2	6
E3	21	0.008 2	6
F3	22	2.000 0	6
G3	23	0.709 7	6
H3	24	2.000 0	6
A4	25	2.000 0	6
B4	Water	–	6
C4	Water	–	6
D4	Water	–	6
E4	Water	–	6
F4	Untreated	–	6
G4	Untreated	–	6
H4	Untreated	–	6
A5	1	0.008 6	24
B5	2	2.000 0	24
C5	3	0.001 9	24
D5	4	2.000 0	24
E5	5	0.001 6	24
F5	6	0.007 9	24
G5	7	0.003 6	24
H5	8	0.124 7	24
A6	9	0.531 9	24
B6	10	1.085 4	24
C6	11	2.000 0	24
D6	12	2.000 0	24
E6	13	0.754 4	24
F6	14	0.949 1	24
G6	15	0.092 6	24
H6	16	0.000 2	24
A7	17	0.000 2	24
B7	18	0.157 9	24
C7	19	0.024 2	24
D7	20	0.838 2	24
E7	21	0.008 2	24
F7	22	2.000 0	24
G7	23	0.709 7	24
H7	24	2.000 0	24
A8	25	2.000 0	24
B8	Water	–	24
C8	Water	–	24
D8	Water	–	24
E8	Water	–	24
F8	Untreated	–	24
G8	Untreated	–	24
H8	Untreated	–	24
A9	1	0.008 6	36
B9	2	2.000 0	36
C9	3	0.001 9	36
D9	4	2.000 0	36
E9	5	0.001 6	36
F9	6	0.007 9	36
G9	7	0.003 6	36
H9	8	0.124 7	36
A10	9	0.531 9	36
B10	10	1.085 4	36
C10	11	2.000 0	36
D10	12	2.000 0	36
E10	13	0.754 4	36
F10	14	0.949 1	36
G10	15	0.092 6	36
H10	16	0.000 2	36
A11	17	0.000 2	36
B11	18	0.157 9	36
C11	19	0.024 2	36
D11	20	0.838 2	36
E11	21	0.008 2	36
F11	22	2.000 0	36
G11	23	0.709 7	36
H11	24	2.000 0	36
A12	25	2.000 0	36
B12	Water	–	36
C12	Water	–	36
D12	Water	–	36
E12	Water	–	36
F12	Untreated	–	36
G12	Untreated	–	36
H12	Untreated	–	36

Table 9: Layout of samples in 2 96-well plates for PLATE-Seq.



	1	2	3	4	5	6	7	8	9	10	11	12
A	0.7	0	0.5	0	0.7	0	0.7	0	0.6	0	0.8	0
B	1.2	0	0.9	0	0.6	0	0.7	0	0.8	0	0.7	0
C	0.7	0	0.9	0	0.5	0	0.7	0	0.5	0	0.7	0
D	0.4	0	0.8	0	0.5	0	0.8	0	0.6	0	0.8	0
E	0.8	0	0.8	0	0.9	0	0.7	0	0.8	0	0.8	0
F	0.8	0	0.6	0	0.9	0	0.7	0	0.4	0	0.6	0
G	1.1	0	0.7	0	0.6	0	0.9	0	0.6	0	0.3	0
H	0.9	0	0.9	0	0.9	0	0.8	0	0.5	0	0.7	0

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.7	0	0.6	0	0.8	0	0.9	0	0.8	0	0.8	0
B	1.3	0	0.9	0	0.6	0	0.9	0	0.9	0	0.8	0
C	1	0	1	0	1	0	0.8	0	0.9	0	0.8	0
D	0.8	0	1.2	0	0.8	0	1	0	1	0	0.8	0
E	0.9	0	1.1	0	1.1	0	0.9	0	1	0	0.9	0
F	1.2	0	0.7	0	1.1	0	1	0	0.7	0	0.7	0
G	1.4	0	1.1	0	0.7	0	0.8	0	0.8	0	0.5	0
H	1.1	0	0.9	0	0.9	0	1.1	0	0.6	0	0.8	0

(a) Mapped reads x Genes

(b) Spike-ins

Figure 12: Detected genes and spike-ins. (a.) Association between the number of mapped reads and detected genes for each of the 96 analyzed samples. (b.) Heatmap showing the number of reads (thousands) mapping to spike-ins for each of the samples.

Appendix B Mechanism diagrams

The following mechanisms—from the Reactome web resource—describe the molecular functions for ATPase inhibitor and FGFR inhibitor drugs, which have similar effects on global gene expression as *A. lobata* and *S. maurus* venom, respectively (see §3.3).

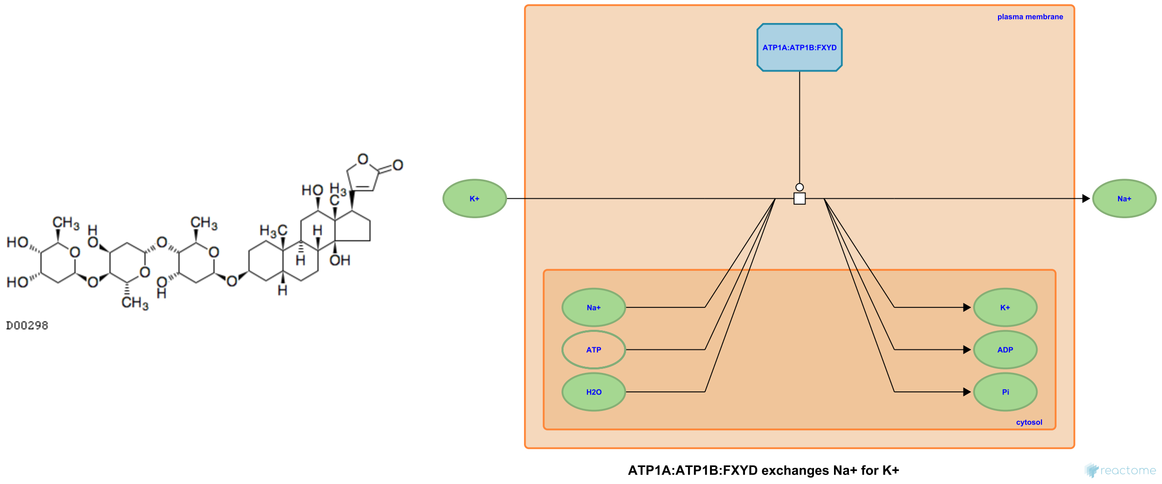


Figure 13: Structure of digoxin (left), a cardiac glycoside that inhibits the function of the Na+/K+ ATPase (ATP1A; right) in the myocardium, which causes a decrease in heart rate [Kühlbrandt, 2004]. *A. lobata* venom has similar differential expression effects to those of digoxin and other ATPase inhibitor drugs, based on connectivity analysis. Diagram from Reactome [Fabregat et al., 2015].

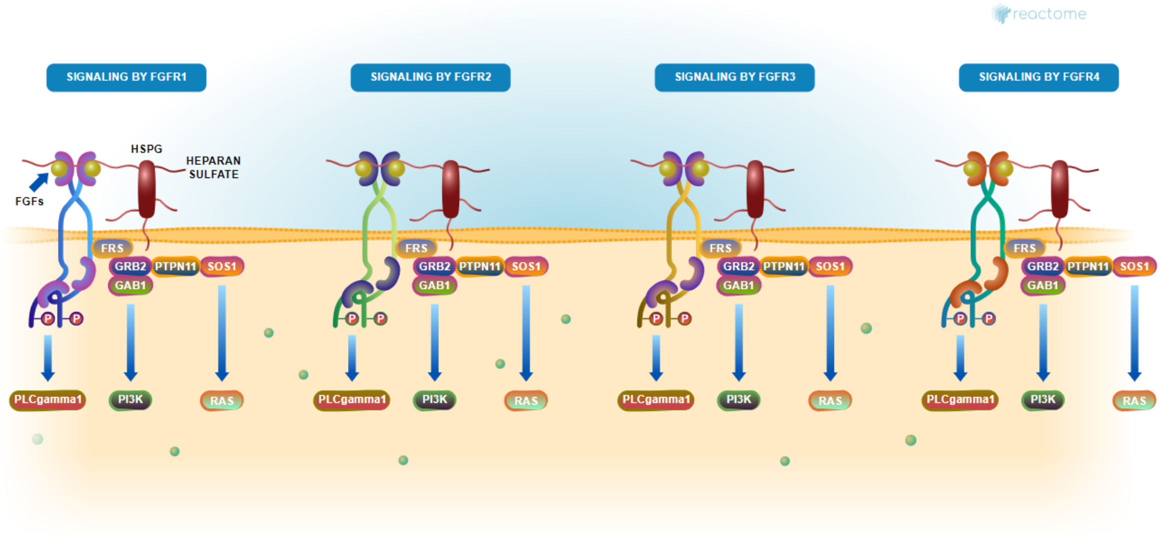


Figure 14: Diagram of FGFR signaling pathways. FGFR inhibitors target 1 of the 4 types of FGFR complexes, abnormal activity of which are involved in angiogenesis. `VenomSeq` suggests therapeutic similarity between *S. maurus* venom and existing FGFR inhibitor drugs. Pathway diagram from Reactome [de Bono et al., 2007].

Appendix C Miscellaneous supplemental figures

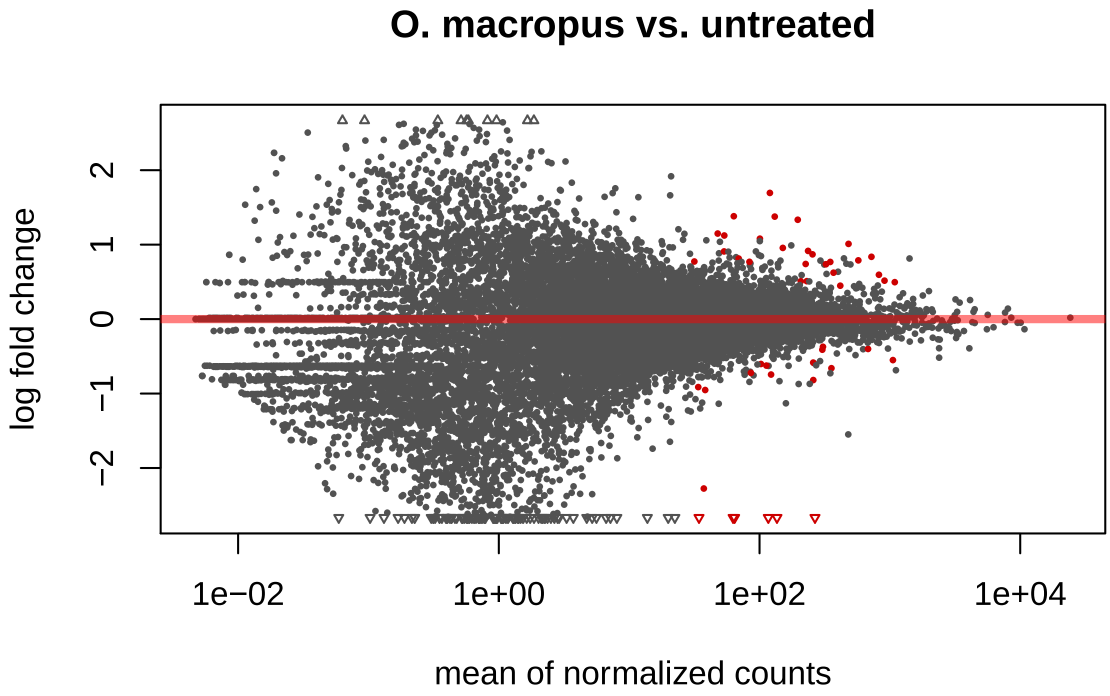


Figure 15: MA plot showing genewise relationship between $\log_2$ fold change and mean of normalized counts in samples corresponding to *O. macropus* venom. Each point represents one gene. Points in red indicate statistically significant genes with regard to differential expression.