Using linear regression and PathFX networks identifies genes associated with psychiatric drug effects in zebrafish screens

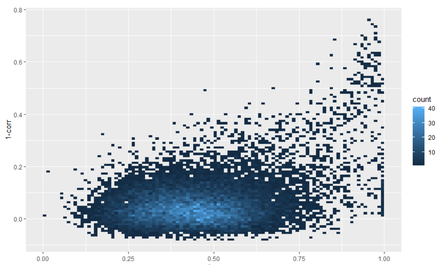
Jarod Le and Jennifer Wilson

Lune Lab. Department of Bioengineering. University of California Los Angeles

Introduction

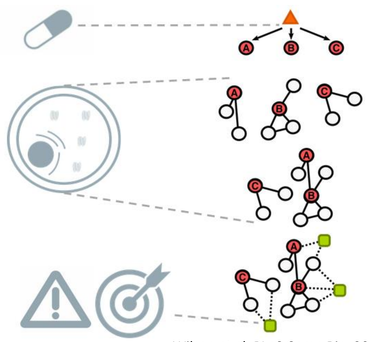
- 25% of U.S. adults have serious mental illness (NIH, 2020)
- No new therapeutic mechanisms or new drug targets have been identified within the last few decades (Mathur & Guo, 2010)
- Many psychiatric diseases are polygenic and discovering new therapies requires a systems level approach

Zebrafish Screening Data



We used the phenoscores published in McCarroll, et al. The phenoscore is a composite metric reflecting how each drug affected live zebrafish behavior. We mapped the chemical names to a corresponding name to DrugBank for later analysis.

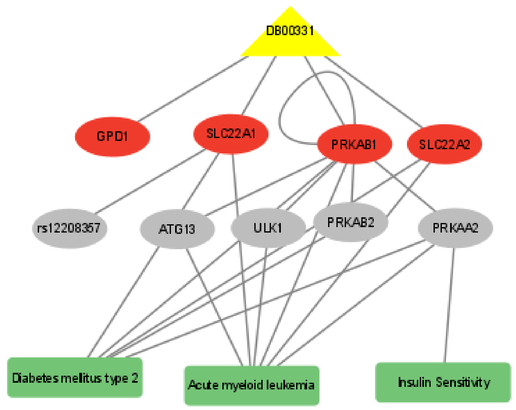
PathFX for getting Networks



Wilson et al, PLoS Comp Bio, 2018
Wilson et al, Bioinformatics, 2019

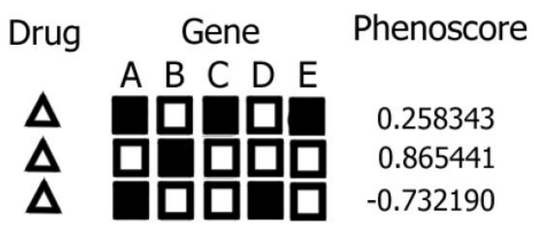
We modeled all drugs using PathFX (Wilson et al, 2018). PathFX is a program that uses protein protein interactions (ppi) to connect drugs to downstream genes/proteins and disease phenotypes.

PathFX Example Network



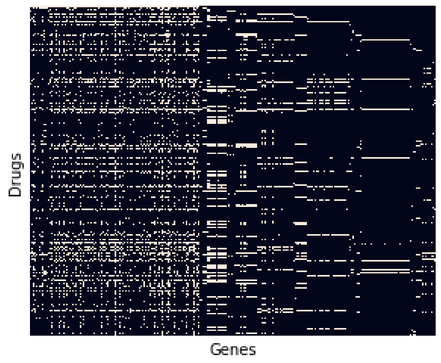
PathFX connected DB00331 (Metformin, and anti-diabetic) to diabetes-related phenotypes. The drug is represented as a triangle, the proteins are represented as ellipses, and phenotypes as rectangles. The drug targets the target proteins, which targets the intermediate proteins, and finally the phenotypes.

Converting network data to linear regression inputs



Of the 10479 unique chemical names published in McCarrroll et al, we mapped 327 to DrugBank identifiers. We modeled the 327 drugs using PathFX and analyzed the network genes of every drug. We created a dataframe by one-hot encoding network proteins and used this matrix with sklearn linear regression. We removed network genes connected to 5 or fewer drugs. This process reduced 2219 gene columns to 995 gene columns for further analysis.

Dataframe of the Drugs to Genes



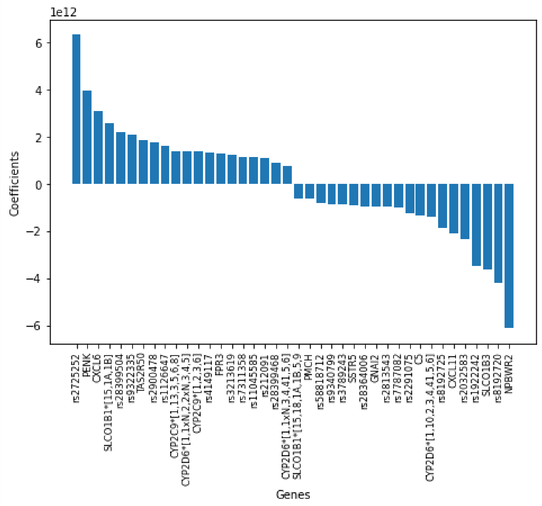
This heatmap shows the 327 drugs (rows) and 995 genes (columns). A "1" or "0" in the dataframe represents whether the drug does or does not have the gene in its PathFX network, respectively. As shown, drug networks share many genes, and some genes are unique to a handful of drugs.

Testing predictability of network proteins

Method	TR R2/MSE	TS R2/MSE
TR/TS Split	0.312/0.006	-7.002e+23/ 2.839e+21
5-Fold Nest Cross Valid	-3.377e+21/ 2.825e+19	-7.162e+22/ 2.904e+20
No Split	0.345/0.005	N/A

We tested the data with training testing split of 20%, 5-Fold Nested Cross Validation, and doing no split. Looking at the R2 and mean squared error scores for the split data, we can see high scores. This would indicate that we do not have enough drugs to accurately find the phenoscore. Another possibility would be that genes might not be enough to predict the phenoscore.

Linear regression with full dataset prioritized biological pathways



Gene	Description	Coefficient
rs2725252	Xenobiotic Transporter with a role in multi-drug resistance	6356558309595.174
PENK	Related to Opioid Receptor to modulate the perception of pain	3956831116161.347
CXCL6	Is a chemotactic for neutrophil granulocytes which is related to chemokine receptor	3113710291853.38
SLCO1B1	Transmembrane receptor that is related to the removal of drug compounds	2594069569488.2866
rs28399504	Monooxygenases which catalyze many reactions involved in drug metabolism	2177691556212.037

Looking at the waterfall graph of the top 20 and bottom 20 prioritized genes based on their coefficient from the linear regression from not splitting the data, we can see which genes the model prioritized. Looking more closely at the top 5 genes, we can see a prioritization of xenobiotic transporter with some importance towards opioid and chemokine receptor.
For further research we would need more drugs. As the original data had 10479 unique chemical names, we could find out the target proteins to push into PathFX for the other drugs we could use.

Citations

Mathur P, Guo S. Use of zebrafish as a model to understand mechanisms of addiction and complex neurobehavioral phenotypes. Neurobiol Dis. 2010 Oct;40(1):66-72. doi: 10.1016/j.nbd.2010.05.016. Epub 2010 May 20. PMID: 20493262; PMCID: PMC3021971.
McCarroll, M.N., Gendelev, L., Kinser, R. et al. Zebrafish behavioural profiling identifies GABA and serotonin receptor ligands related to sedation and paradoxical excitation. Nat Commun 10, 4078 (2019). https://doi.org/10.1038/s41467-019-11936-w
Wilson JL, Racz R, Liu T, Adeniyi O, Sun J, et al. (2018) PathFX provides mechanistic insights into drug efficacy and safety for regulatory review and therapeutic development. PLOS Computational Biology 14(12): e1006614. https://doi.org/10.1371/journal.pcbi.1006614