Pipeline for Repositioning Anti-Cancer Drugs using Copy Number Variation Data

Nielsen, Jared; Olson, Kaleb and Olpin, Richard

April 26, 2017

Abstract

Drug repositioning research can significantly narrow down the number of drugs that might be considered when trying to find potential treatments for a cancer. By using a recently released dataset of copy number variation covering 19,000 genomes we attempt to find drugs that that are effective against cell lines with similar copy number variation. We present the pipeline used to perform this analysis and the results obtained, namely, pairings between cancer types and candidate anti-cancer drugs.

Contents

Introduction 1

Methods 2
- Overview . . . 2
- Data Exploration and Processing . 2
- Translating GDSC CN Level . . . 2
- Subtyping GENIE . . . 2
- Drug Selection . . . 3
- Profile Generation . . . 3
- Profile Scoring . . . 3

Results 3
- DRPs Display Similarity . . . 3
- Translation Produces Similar Distribution . . . 3
- Scoring Matrix Results . . . 4

Discussion 4
- Implications . . . 4
- Challenges . . . 4
- Future Work . . . 4

Author Contributions 5

Introduction

Cancer research has made significant advances in the last decade leading to more targeted anti-cancer drugs. This benefits patients as the more targeted drugs are less harmful to non-cancerous cells; however, having targeted drugs means that pairing the right drug with a cancer type becomes much more important. Because of the many possible combinations this becomes a significant challenge. Luckily, this challenge can be lessened if we can computationally identify drugs which are likely to be effective in treating certain types of cancer. This process is known as drug repositioning.

This year, January 2017, a collaborative effort between several well known cancer institutions was published. This effort pools data from each institution's clinical cancer data into a repository called Genomic Evidence Neoplasia Information Exchange (GENIE)[1], making their data available for other researchers to study. One of the major aims of GENIE is drug repositioning. Many researchers already use copy number variation data as one of the genomic features to use as a predictive biomarker for drug response[2].

Most of the current research using CN data is limited to using only cell line data because of the scarcity of clinical CN data. With the GENIE release we now have access to over 19,000 clinical records including patient CNV data which provides a valuable insight into the biological basis for many cancers.

We also use a highly respected and cited database, the Genomics of Drug Sensitivity in Cancer (GDSC)[3] to link the cell line CN data associated with a variety of drugs and the clinical CN data in GENIE. Our goal is to create an efficient and flexible pipeline to connect the data in both databases in order to match cancer types with potential drugs. This pipeline scores how effective a drug might be to treat a cancer based on similar CNV data found in both the cell lines and clinical records.

We selected 16 drugs from the GDSC (4), filtered the data accordingly, and then generated a Drug Response Profile (DRP) from the associated cell line CNV data by CN level. We also subset the GENIE CNV data by a variety of

factors resulting in over 500 cancer subsets from GENIE and generated a Cancer Subset Profile (CSP) for each by CN level. Then we ranked the sum similarity between each DRP and CSP combination. We found that our best scores correctly connected cancers with drugs which target pathways currently used in oncology treatments.

Methods

Our pipeline for identifying candidate drugs for re-positioning was built in R and runs on Linux and Windows. We use a variety of R libraries to help process and plot the data including readxl[4], readr[5], synapseclient[6], reshape[7], dplyr[8], data.table[9], ggplot2[10], and treemap[11]. We wrote our pipeline in a literate programming style using knitr[12] to aid readability and better convey the exploratory narrative of our work. We have made this code and several preprocessed data files available on our project's github page. A high level overview of the pipeline can be seen in our process flow diagram (figure 4), but is also described below.

Overview

First we worked with the GENIE data to find the intersection of the genes in the GENIE and the GDSC CNV data. We then filtered the GENIE data to include only those genes. After which we programatically subset the GENIE data by several variables available in the clinical data associated with each sample. As the final step in processing the GENIE data we created a proportional representation of the CNVs for the subset data which we call the Cancer Subset Profile (CSP).

When processing the GDSC data we selected a handful of drugs that were effective for only a few cell lines in the hopes to find more targeted drugs that would be less harmful to non cancerous cells. We then filtered the GDSC data accordingly and translated the CNV data to another schema with five discrete values to match the GENIE data. Finally we created profiles of the drugs, similar to how we created profiles of the GENIE data, from the set of cell lines that were responsive to that drug which we call the Drug Response Profile (DRP).

Once we have both the DRP and CSP we compare them by CN level and create a similarity matrix. With this similarity matrix we can identify the best pairings between a DRP and a CSP. Although this is not conclusive evidence that a drug will be effective against a certain cancer, it does provide a good starting point for further research.

Data Exploration and Processing

We download all of the source data files from both GENIE and GDSC, which is delineated in our public pipeline code. We transformed the GDSC files from .xlsx to .tsv to better meet R standards. Then we converted both GENIE and GDSC files into "tidy" format as described by Hadley Wickham[7]. Next, we found 587 intersecting genes between both databases and filtered each to only have the matching genes. At this stage, we have both databases with files that only have genes which are present in both, so we can make accurate comparisons.

Translating GDSC CN Level

The two databases had different forms of copy number annotation. GENIE used a more popular style with five discrete values (-2, -1, 0, 1, and 2)*. A value of 0 represents a normal, diploid copy level. Values of 1 and 2 represent slight and extreme over expression, respectively. Values of -1 and -2 are slight and extreme under expression.

The GDSC had a different annotation that was difficult to decipher. One value for copy level at a gene was represented by 4 variables; "Max copy level observed", "Minimum copy number observed", "Heteroegeneity", "Disruption". We ignored the last values as they did not contribute to the expression level. From the GDSC definitions, we are extremely confident that a value of "0,0" is equivalent to a -2 in GENIE, a value of "3,3" in GDSC is the same as a 0 in GENIE and a value of "8 or greater,8 or greater" equals a 2 in GENIE. However, we had to make educated guesses to translate GDSC levels to either 1 or -1. We chose "1,1 or 0" to represent a -1 and "8, less than 8" to represent a 1. Due to the fact that the extreme expression levels (-2,2) have greater influence of the copy number profile we feel justified in our translation. A complete table of the translation mapping is found in figure 1.

Subtyping GENIE

We chose to subset the GENIE database by 5 main categories; cancers, cancer subtype, sex, primary race, and age. We used the 25 cancer types with the most records and all their corresponding subtypes (figure 1). We chose both male and female patients. Although GENIE

*GENIE did contain about 500 records that had a CN level of -1.5; however, we choose to ignore these values for the purposes of this study

GENIE	-2	-1	0	1	2
GDSC	0,0	1, <=1	<=3, <3	8, <8	>=8, >=8

Table 1: GDSC to GENIE CN Translation

We do not alter the the GENIE data. We translate the GDSC data which is originally represented by 4 characters, "2,2,H,D". The first two characters give us a maximum and minimum copy number, respectively and range from 0-14. The table shows the conversion, for example, to represent a "-2" found in GENIE, in GDSC, they use "0,0,-,-". To create matching profiles between the two datasets we convert the GDSC to match the annotation in the GENIE. We found that by allowing max and min values less than three and greater than one to resolve to zero to best replicate the distribution found in GENIE.

contains several demographic categories, many of them are highly skewed and limited. Of the 19,000 records, nearly 15,000 are from White patients. The only other two races with significant records are Black and Asian. We decided to look at only patients who were under 40 years of age. We broke the ages into four subcategories: <18, 19-30, 31-40, and 19-40 (see table 2).

Drug Selection

We selected our top sixteen drugs by first filtering out the drug data that had an IC50 value less than -2. This made it so that we were looking only at the drug tests that caused a significant response in the cell line. In addition, we filtered out drug tests that had an AUC value of greater than .5, which kept those cell lines for which we had the greatest certainty. From this filtered data we ranked each drug, from high to low, based on how many cell lines it affected. We chose the drugs affecting between 10 and 100 cell lines as a way to eliminate weak drugs, as well as overly powerful ones. After doing this we had a total of sixteen drugs (4), with each one affecting different numbers of cell lines. To normalize each of the sixteen drugs, so that the CN profiles would be comparable, we selected the top ten cell lines for each one. These were then used in generating our top DRPs.

Profile Generation

Profile generation is at the core of our pipeline. We define a profile as the proportion that a specific copy level number appears each gene across a group of either cell lines or cancer patients. Using profiles allows us to avoid complicated and costly comparisons between individual cell lines and tumor samples. The process for generating a profile is the same for both a subset of cancer samples and a set of cell lines that respond to a drug. We start by using a table where each row represents the copy number for each gene represented by a sample, or cell line in our data. With this table we then convert it to a logical matrix (e.g., does an entry have a CN level of -2). From this matrix we create a vector whose entries are the proportion of entries in a column that are set to $TRUE$. These profiles, along with the subset or drug id and CN level that they correspond with, are then stored in tables that are used during scoring.

Profile Scoring

For determining the overall similarity between a drug and a cancer subset we select all DRPs and CSPs for set of CN levels (e.g., -2 and 2). Then for each CN level we find the euclidean distance between the corresponding DRP and CSP and add it to a running total.

The total distance is placed in a score matrix representing how well each drug matched with each cancer subset. We then reshape the matrix to a tidy form in order to graph a heatmap of the distance scores.

Results

DRPs Display Similarity

As a way to test our methodology for producing DRP's, we compared drugs that affect the same pathway to each other, to see if they matched. We did this in two ways first we calculated the overall distance score that they generate when compared to one another. If the overall distance score was low we knew that they were more similar, than when the score was high. The other way we confirmed that they were similar was by plotting the two drugs as a way to visualize their profiles (see figure 3). From this figure we can see that these profiles are similar, which is what we would expect given that they affect the same target pathway.

Translation Produces Similar Distribution

One measure of how well our translation of the GDSC copy number data worked is by comparing the translated CN level distribution in GDSC with that of GENIE. In other words we would expect a realistic distribution to be heavily weighted towards a CN value of 0, followed

Filter Category	Filter Values
Cancer Type	25 Most Common
Cancer Subtype	All subtypes of 25 cancers
Sex	Male & Female
Primary Race	White, Black, Asian
Age	<18, 19-30, 31-40, 19-40

Table 2: Filter string values
Illustrates which filters we applied to generate our 520 CSP's. We define the 25 most common cancers in GENIE by how many records are present in the database.

by -1 and 1, and then by -2 and 2. In our first attempts of finding a suitable translation method we saw that distribution was considerably skewed. However, as we refined the translation by incorporating the min value as well as the max value and tweaking the parameters we were able to reach a similar distribution (see figure 2).

Scoring Matrix Results

Our best scores between the DRP and CSP are supported by current research. We found high similarity between drugs that target the PI3K pathway a highly targeted pathway linked to several different types of cancer like Glioma, Bladder cancer and Mesothelioma[13–16]. Figure 5 visualizes the similarity of our top 60 scoring CSP's and each of the 16 DRP's. To see a summary of just the top 12 scores see 3. These correlations validate our pipeline because they prove that we are making biologically sound connections. Had our top results returned only unknown connections it would reveal fatal errors in our pipeline most likely in either our translation of the GDSC or in our scoring algorithm. When we consider the make up of the GENIE database (see figure 1) we see that if our results were random then well represented cancers in GENIE, like Non-Small lung cancer, would show up much more often in our results. While these results do not show exactly how well our method works, they do serve as a proof-of-concept. We delineate several different options to further validate our pipeline in our discussion.

Discussion

Implications

With our results we have demonstrated that Copy Number provides a link between Clinical and Cell Line data. We have also shown that by using this link it is possible to reposition drugs for specific subtypes of cancer. The pipeline we developed provides a process future researchers will be able to build from. We do not claim that our methods are optimal, however they provide a general methodology that can be applied when using copy number in relating these two types of data. We believe that this pipeline will be an effective tool to aid in the repositioning of cancer drugs.

Challenges

The documentation explaining the copy number data annotation was sparse, particularly when we were trying to determine which values in GDSC that corresponded to either a 1 or -1 in GENIE. There was no clear way to align the different levels so we estimated as best we could. By plotting the different distributions of each database we modified our translation of the GDSC to best match the distribution of the GENIE (see table 2). Acknowledging the inherent bias in the GENIE is important. Minorities are severely underrepresented among the patients, the majority of the patients are over 40 and there are several other biases. These biases must be taken into consideration before using GENIE as a primary source of data for a certain project. Another challenge we faced came when we were calculating the distance between two profiles. We realized that when comparing profiles that represented different numbers of Cell Lines the Copy Number signal at the gene was being washed out. This led to drugs that were hyper-effective being represented more heavily than those that were affecting less Cell Lines. As a way to correct for this we needed to come up with a way to normalize how we generated our profiles from the DRP. We did this by only looking at the top ten cell lines for every drug we were testing, allowing for more accurate comparisons.

Future Work

We need further validation to refine our pipeline. To better test our translation and scoring methods we would pick a new set of drugs that all target the same pathway. Then we would research that pathway to find out which cancers are as-

sociated with that pathway as good candidates for drug intervention. We would need to verify that our current GENIE subsetting process including these cancers. If not, we would adjust the parameters so that we were filtering GENIE appropriately. Then we would rerun our pipeline testing these new drugs against the GENIE subsets. We would expect our results to have low scores for the known good responding cancers and high scores for the unrelated cancer types. It would be ideal to test this with many different drug sets targeting many different pathways. It is possible that our methodology works well for only certain pathways and not others, further validation tests would reveal the strength of our pipeline across the different pathways and cancers. Statistical testing would then be necessary to evaluate the significance of our correlations.

Once our methods are validated we can then push through with high throughput comparisons or create narrow filters to look at rarer forms of cancer. That was part of our design process. We wanted to develop a pipeline to allow future researchers to quickly make comparisons however they please. We also need to better define our different thresholds of how many records or samples we need in order to build a CSP or DRP. We used a lower bound of 25 records to build a CSP with no upper bound and we limited our DRP based on drugs which only had between 25 and 100 cell line responses. These thresholds should be revisited and refined.

If we were to make this truly a pipeline for other researchers, we would need to refactor some aspects of our code to make it more user friendly. It could be adapted into a Shiny app where the user chooses how to filter both the drugs to test and cancer groups to create as well as different thresholds. With our pipeline as the back end code, the app would return the heatmap and score matrix table allowing for quick analysis.

Author Contributions

An estimation of time spent by each team member on the project can be found listed in Table 5. We decided to begin counting hours since the 5th of April as that is when we feel we really settled on the project; however, this is still a rather rough estimate as we did not track time spent on the different parts of the project so far. The time spent becoming familiar with the GENIE data is also not included in the total.

References

1. AACR makes public 19,000 cancer genomes. *Nature Biotechnology* **35,** 104–104. ISSN: 1087-0156 (Feb. 2017).
2. Chen, B. & Butte, A. Leveraging big data to transform target selection and drug discovery. *Clinical Pharmacology & Therapeutics* **99,** 285–297. ISSN: 00099236 (Mar. 2016).
3. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research* **41,** D955–D961. ISSN: 0305-1048 (Jan. 2013).
4. Wickham, H. & Bryan, J. *readxl: Read Excel Files* (2017). <https://cran.r-project.org/package=readxl>.
5. Wickham, H., Hester, J. & Francois, R. *readr: Read Rectangular Text Data* (2017). <https://cran.r-project.org/package=readr>.
6. Furia, M. *synapseClient: Synapse R Client from Sage Bionetworks* (2017). <http://www.sagebase.org>.
7. Wickham, H. Reshaping Data with the {reshape} Package. *Journal of Statistical Software* **21,** 1–20 (2007).
8. Wickham, H. & Francois, R. *dplyr: A Grammar of Data Manipulation* (2016). <https://cran.r-project.org/package=dplyr>.
9. Dowle, M. & Srinivasan, A. *data.table: Extension of 'data.frame'* (2017). <https://cran.r-project.org/package=data.table>.
10. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-0-387-98140-6. <http://ggplot2.org> (Springer-Verlag New York, 2009).
11. Tennekes, M. *treemap: Treemap Visualization* (2017). <https://cran.r-project.org/package=treemap>.
12. Xie, Y. in *Implementing Reproducible Computational Research* (eds Stodden, V., Leisch, F. & Peng, R. D.) ISBN 978-1466561595 (Chapman and Hall/CRC, 2014). <http://www.crcpress.com/product/isbn/9781466561595>.
13. Fan, Q.-W. & Weiss, W. A. Targeting the RTK-PI3K-mTOR axis in malignant glioma: overcoming resistance. *Current topics in microbiology and immunology* **347,** 279–96. ISSN: 0070-217X (2010).

14. Weigelt, B. & Downward, J. Genomic Determinants of PI3K Pathway Inhibitor Response in Cancer. *Frontiers in oncology* **2,** 109. ISSN: 2234-943X (2012).

15. Zhou, S. *et al.* Multipoint targeting of the PI3K/mTOR pathway in mesothelioma. *British Journal of Cancer* **110,** 2479–2488. ISSN: 0007-0920 (May 2014).

16. Costa, C. *et al.* Abnormal Protein Glycosylation and Activated PI3K/Akt/mTOR Pathway: Role in Bladder Cancer Prognosis and Targeted Therapeutics. *PLOS ONE* **10** (ed Real, F. X.) e0141253. ISSN: 1932-6203 (Nov. 2015).

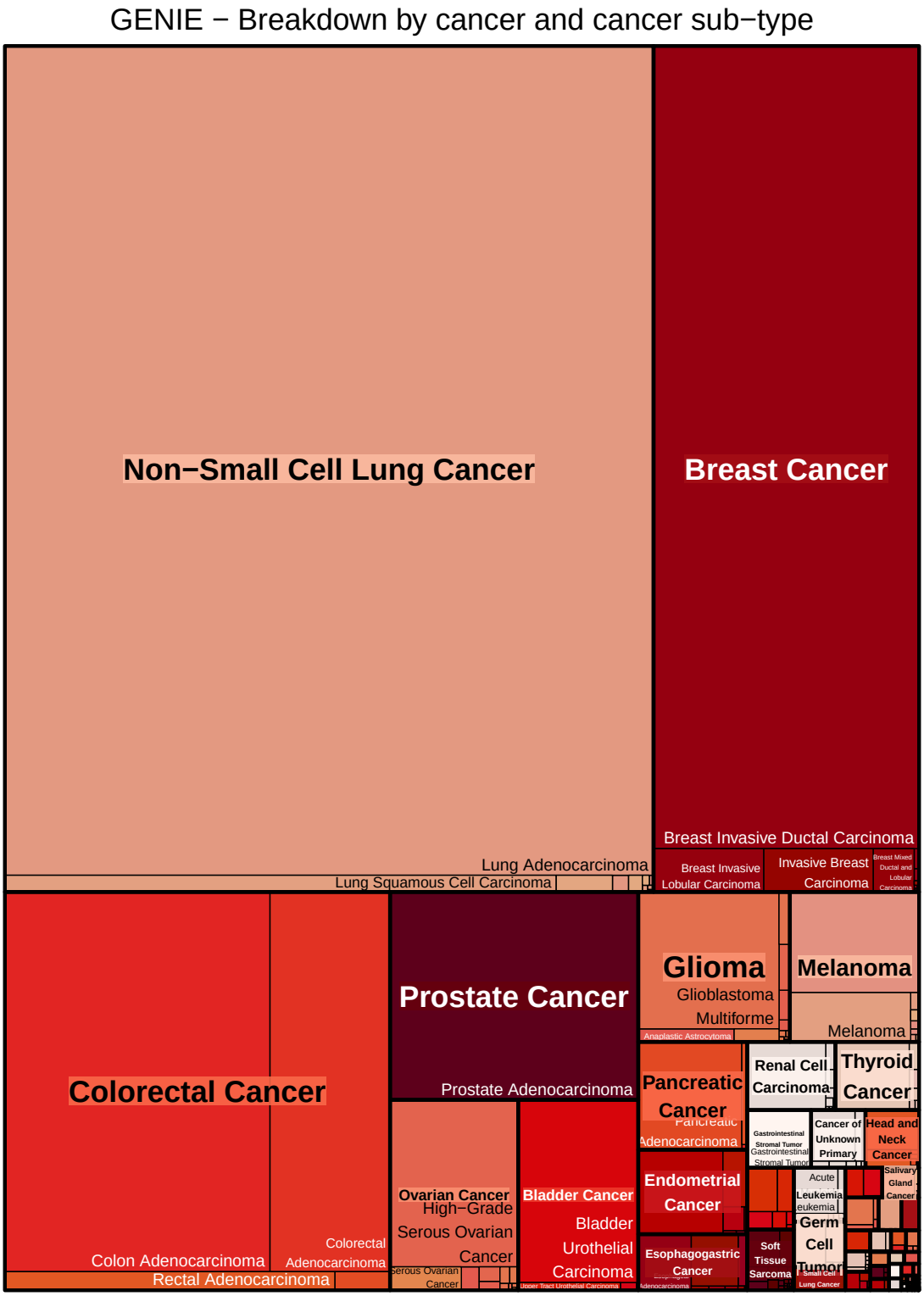

Figure 1: GENIE Cancer Types and Subtypes

This tree map shows a breakdown of the GENIE dataset by cancer type (center aligned labels) and subtype (right-bottom aligned labels). The size of each rectangle corresponds to the number of records in the dataset with that subtype. Overall we can see that non-small cell lung cancer is by far the best represented in the data. We also note that for each cancer type that there is a single subtype that dominates all others.

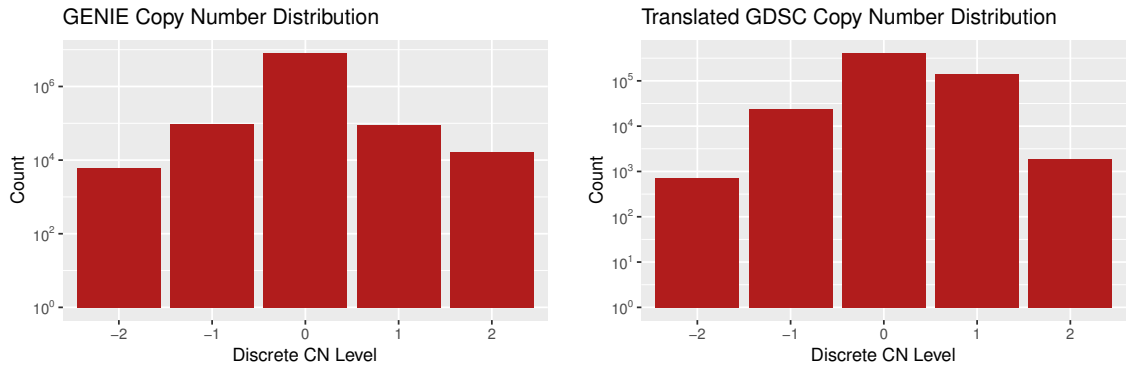

Figure 2: GENIE and Translated GDSC CN level Distributions

The GENIE CN data has five discrete values. In the above histograms the counts for each value are plotted on a log scale showing the distribution of different CN levels in both the native GENIE dataset as well as the GDSC dataset after translation (see table 1).

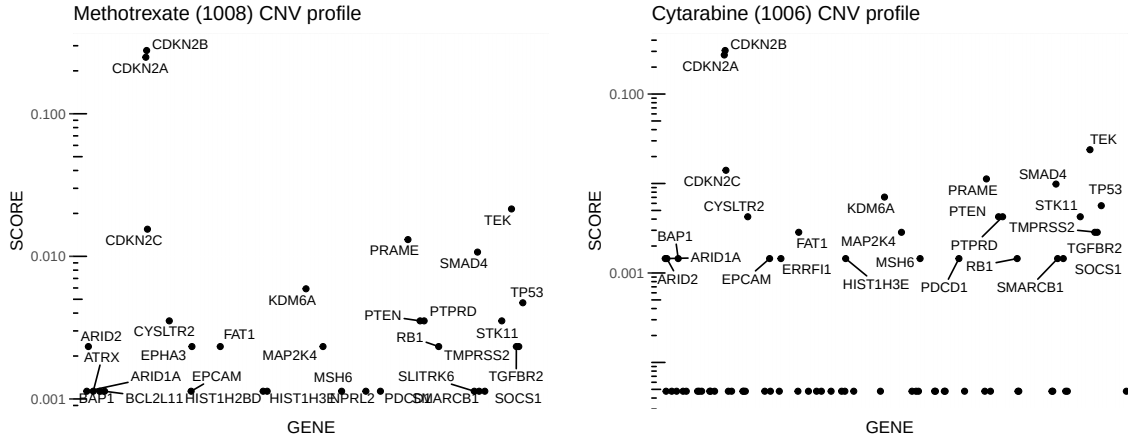

Figure 3: Example of Drug Matching Profile

These two scatter-plots show the profiles for two drugs Cytarabine and Methotrexate on a log scale. The two are both Antimetabolite drugs targeting DNA Replication. The profile is represented by showing copy proportion for all significant genes.

GENIE Cancer Subset	Drug Name	Score
Glioma, Male, White	JW-7-52-1	0.197248933698481
Bladder Cancer, Bladder Urothelial Carcinoma, Female	JW-7-52-1	0.217293693561531
Bladder Cancer, Bladder Urothelial Carcinoma, Female, White	JW-7-52-1	0.220477123893562
Glioma, Female, White	JW-7-52-1	0.22107578159198
Bladder Cancer, Female, White	JW-7-52-1	0.226530905780637
Bladder Cancer, Bladder Urothelial Carcinoma, Female	Daporinad	0.227452761479394
Mesothelioma, Pleural Mesothelioma-Biphasic Type, Male	JW-7-52-1	0.230352772468977
Mesothelioma, Pleural Mesothelioma-Biphasic Type, Male, White	JW-7-52-1	0.23193872191475
Bladder Cancer, Bladder Urothelial Carcinoma, Female, White	Daporinad	0.232090150802078
Bladder Cancer, Bladder Urothelial Carcinoma, Female	WZ3105	0.233631285126988
Mesothelioma, Pleural Mesothelioma-Biphasic Type	JW-7-52-1	0.235136309141556
Bladder Cancer, Bladder Urothelial Carcinoma, Female, White	WZ3105	0.235188230517356

Table 3: Top 12 Overall Similarity Scores

Summary table of the best scoring pairs of CSP and DRP. This is to help show empirical scores in addition to the Heat map (see figure 5).

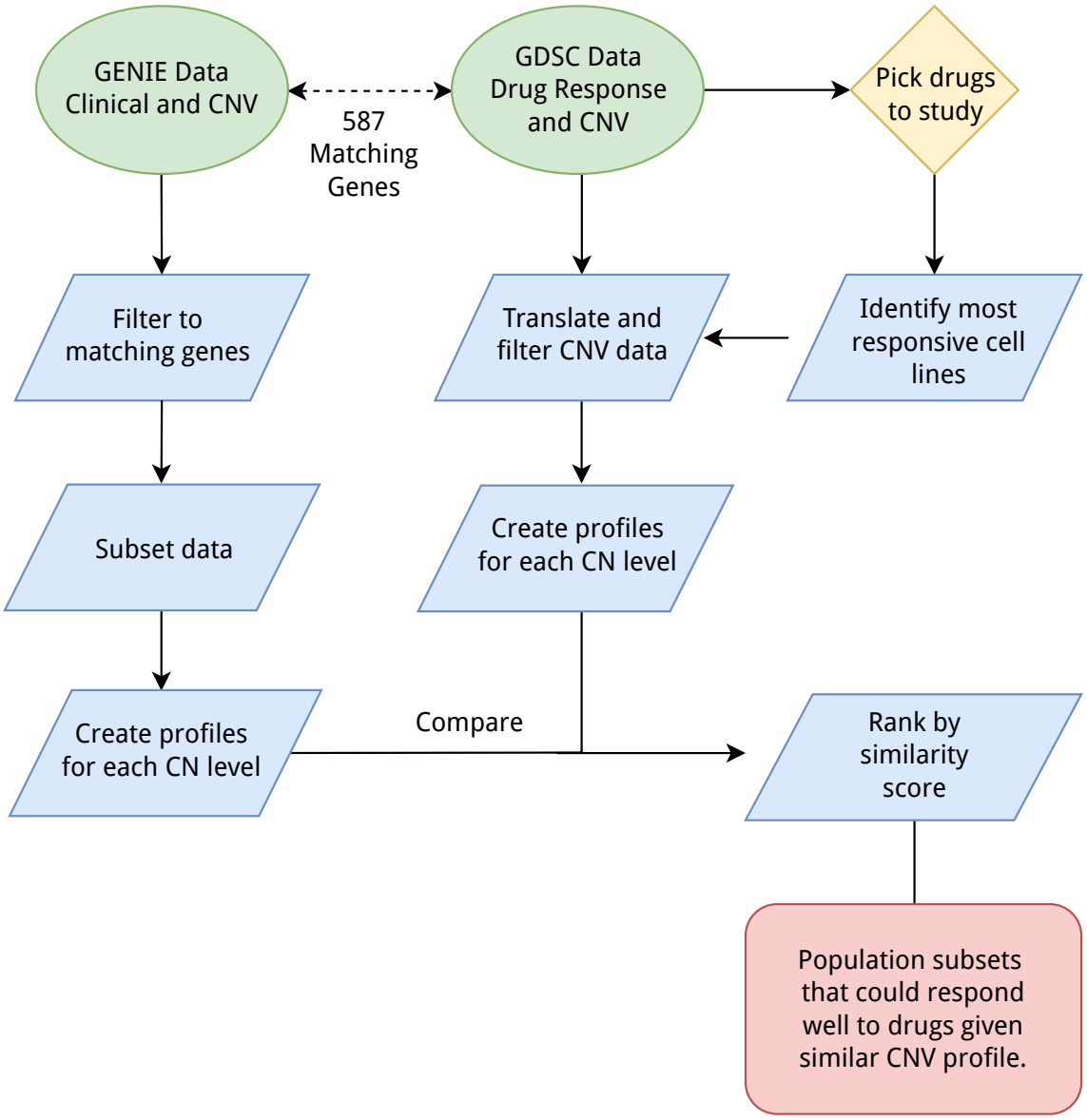

Figure 4: Project Flow Diagram

This is our experimental design. The GDSC database which correlates drug response to cancer line using copy number data. The GENIE database has over 19,000 cancer patient copy number data and we found 587 genes are present in both databases. Using the matching genes, we hope to find subtype of cancer as a potential positive responder to a particular drug due to similarities in copy number profiling.

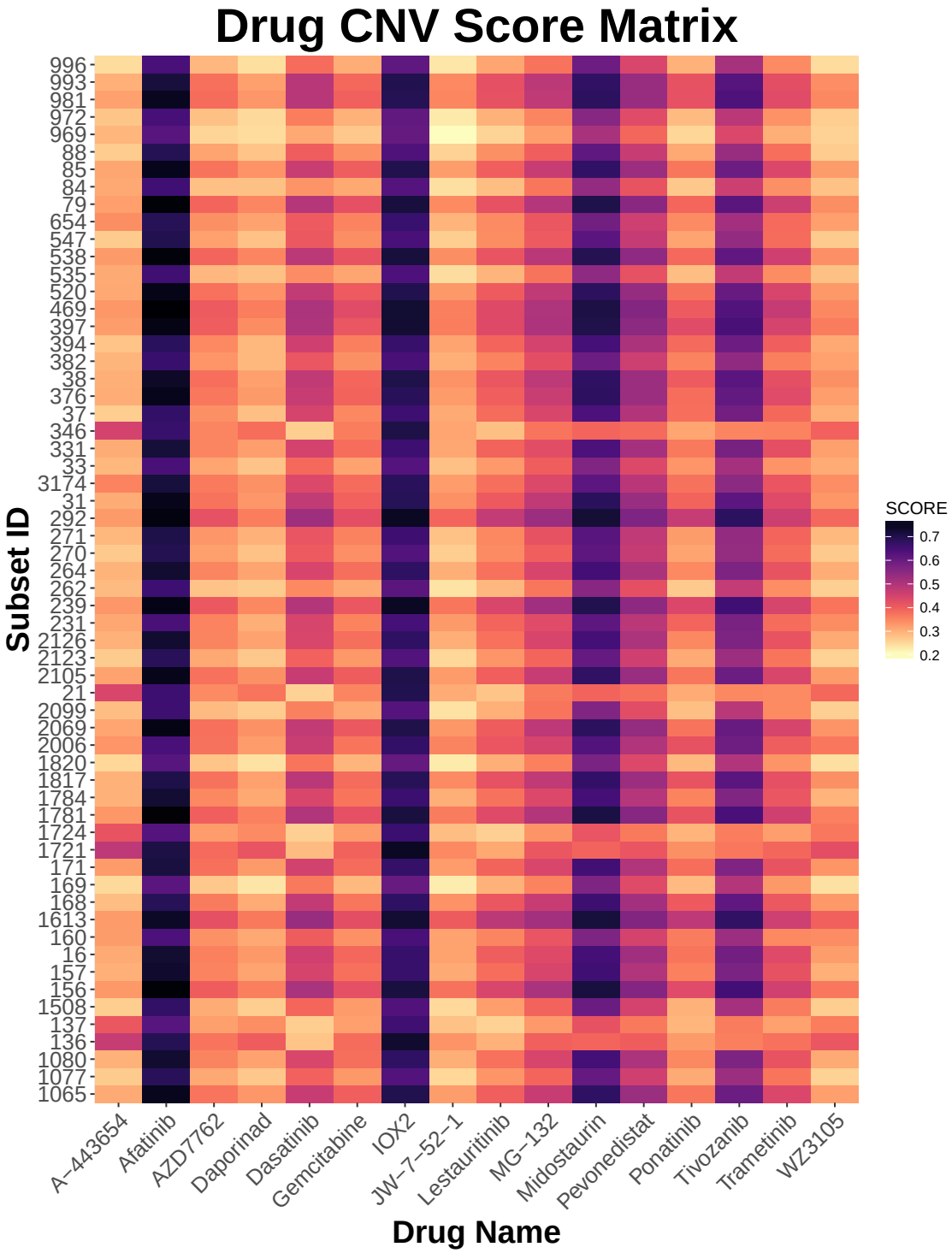

Figure 5: Scoring Matrix

Representation of the top 60 GENIE subsets across the 16 drugs we tested. Each Subset ID represents a specific filter of the GENIE database, a higher ID number generally corresponds to a more specific filter. The individual cells are a score calculating the similarity of the the two profiles by euclidean distance. The more similar the profiles, the lower the score.

Drug ID	Drug Name	Synonyms	Target	Target Pathway
9	MG-132	LLL cpd, MG 132, MG132	Proteasome, CAPN1	Protein stability and degradation
51	Dasatinib	BMS-354825-03, BMS-354825, Sprycel	ABL, SRC, Ephrins, PDGFR, KIT	Other
83	JW-7-52-1	NA	MTOR	PI3K/MTOR signaling
86	A-443654	KIN001-139	AKT1, AKT2, AKT3	PI3K/MTOR signaling
135	Gemcitabine	Gemzar, LY-188011	Pyrimidine antimetabolite	DNA replication
153	Midostaurin	PKC412, benzoylstaurosporine, CGP-41251	PKC, PPK, FLT1, c-FGR, others	Other
155	Ponatinib	AP24534, AP-24534, KIN001-192, Iclusig	ABL, PDGFRA, VEGFR2, FGFR1, SRC, TIE2, FLT3	RTK signaling
252	WZ3105	-	SRC, ROCK2, NTRK2, FLT3, IRAK1, others	Other
312	Tivozanib	AV-951, AV 951, KRN-951, KIL8951, ASP-4130	VEGFR1, VEGFR2, VEGFR3	RTK signaling
1022	AZD7762	AZD-7762, AZD 7762	CHEK1, CHEK2	Cell cycle
1024	Lestauritinib	CEP-701, SP-924, SPM-924, A-154475, KT-555	FLT3, JAK2, NTRK1, NTRK2, NTRK3	Other, kinases
1032	Afatinib	BIBW2992, Tovok, Gilotrif	ERBB2, EGFR	EGFR signaling
1230	IOX2	IOX-2, IOX 2, AK176060	EGLN1	Other
1248	Daporinad	APO866, FK866, FK866	NAMPT	Metabolism
1372	Trametinib	GSK1120212, Mekinist	MEK1, MEK2	ERK MAPK signaling
1529	Pevonedistat	MLN4924, MLN 4924, MLN-4924	NAE	Other

Table 4: Tested Drugs and Pathways

Summary table of the 16 drugs we selected for our project and for which we generated DRP's. This expounds on what pathway each drug targets.

Task	Jared Nielsen	Richard Olpin	Kaleb Olson
Figure generation	3	3	4
Pipeline Development I	3	0	2
Data exploration	6	6	5
CNV translation	2	2	4
Converting data	0	0	1
Filtering data	4	2	0.5
Pipeline Development II	4	3	4.5
Subset GENIE	0.5	5	0.5
Scoring Algorithm	3	1	4
Generate Heatmap	4	4	4
Prepare Presentation	3	3	3
Verify Results	6	2	4
Pipeline Cleanup	2	3	2
Write Report	8	13	13
Total Hours	48.5	49	51.5

Table 5: Estimation of Time Spent per Task