JOHANNES KEPLER
UNIVERSITY LINZ

Submitted by
Evren Aricanli

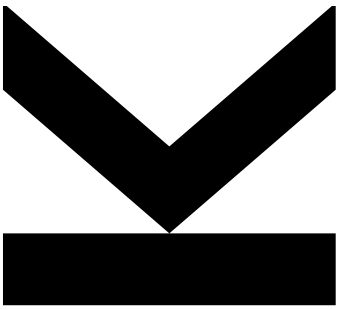
Submitted at
Institute for Machine Learning

Supervisor
Univ.-Prof. Dr. Sepp Hochreiter

Co-Supervisor
Dr. Mag. Günter Klambauer

September 2020

Reimplementation of the Connectivity Map with Accessibility Enhancements



Bachelor Thesis
to obtain the academic degree of
Bachelor of Science
in the Bachelor's Program
Bioinformatics

JOHANNES KEPLER
UNIVERSITY LINZ
Altenbergerstraße 69
4040 Linz, Österreich
www.jku.at
DVR 0093696

Bibliographical Detail

Aricanli, E., 2020: Reimplementation of the Connectivity Map with Accessibility Enhancements. Bachelor Thesis, in English. 53 p., Institute for Machine Learning, Faculty of Engineering & Natural Sciences, Johannes Kepler University, Linz, Austria

Annotation

The Connectivity Map (CMap) is a biomedical research tool which leverages modern biotechnology and bioinformatic methods to provide a platform for comparing the a subject gene expression profile against a library of drug-treatment gene expression profiles. This comparison allows for retrieval of drugs from the CMap's library which induce biological activity similar (or dissimilar) to a given subject expression profile – thereby making it a tool which allows for the finding of repositionable drugs. Herein, a number of adaptations are proposed – most focally being changes which would enhance the tool's accessibility (e.g. a simplified UI/workflow, expanded queryable gene identifiers). Using the R statistical computing environment and numerous libraries available through it, an adapted CMap tool was created from the publically-accessible CMap data, which housed the proposed adaptations. To evaluate success of the accessibility adaptations, the adapted CMap tool was focus tested on several participants – which ultimately yielded great favor for the changes made, allowing them to be conclusively suggested for future iterations of the CMap.

Declaration

I hereby declare that I have worked on my bachelors thesis independently and used only the sources listed in the bibliography.

I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my bachelor thesis, in full to be kept in the Faculty of Science archive, in electronic form in publicly accessible part of the STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages.

Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defence in accordance with aforementioned Act No. 111/1998.

I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

....

Place / Date Evren Aricanli

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Over the course of humanity's use of medicine, a virtually countless number of drugs have been encountered in pursuit of finding therapeutic treatments. These drugs are of either of natural or synthetic origin, with a large number of the synthetic compounds coming into existence under exploratory conditions. The discovery and exploration of these drugs became a possibility as the practice of medicine developed – with modern medicine bringing in an understanding of the underlying mechanisms of medical phenomena (via biomedicine), an understanding of how drug treatments work or how they can be made (via pharmaceutical sciences), and a means by which experimentation and analysis could swiftly be carried out (via advancements in the biotechnology). These developments brought on the virtually countless number of drugs encountered by medicine, of which a minority have made it though to see the light of clinical use for their proposed therapeutic intent. However, there have been numerous serendipitous cases of new life being breathed into drugs – an act known as drug repositioning where alternate therapeutic uses of drugs are found and established. Recently, tools have been developed which remove the serendipity from discovering repositionable drugs — with the most popular tool being the Connectivity Map.

1.2 Statement of Purpose

In this work, we revisit the Connectivity Map (CMap) – a biomedical research tool which leverages modern biotechnology and bioinformatic methods to provide a platform for comparing a subject gene expression profile against a library of drug-treatment gene expression profiles. This comparison allows for retrieval of drugs from the CMap's library which induce biological activity similar (or dissimilar) to a given subject expression profile – thereby making it a tool which allows for the finding of repositionable drugs. Moreover, we propose a number of adaptations which we consider to be advantageous to the platform – these primarily being adaptations to enhance the tool's accessibility.

1.3 Prerequisite Knowledge

1.3.1 Medicine

Disease, disorder, and injury have been ever-present obstacles in mankind's pursuit of good health and well being. Due to humanity's long-standing history with these health-spoiling factors, medicine – the act of managing corporeal afflictions via identification, prevention, and treatment - is a practice well ingrained in the fabric of human existence. The long history of medicine has allowed for advancements in all of its facets - herein, the focal facet being treatment development.

Early vs. Modern Medicine (Biomedicine)

Early medicine (often referred to as traditional medicine) made use of naturally occurring substances as treatments to afflictions. These early medicinal treatments were discovered, shared and developed within societies over generations as passed-on cultural traditions (hence the name traditional medicine). However, the finding of these traditional medical treatments were not rooted in science. Rather, treatments were discovered gradually via experimentation with natural (mainly, herbal) substances in certain adverse situations, where results were retained within a society's medical knowledge.

In contrast to early medicine, modern medicine is deeply rooted in science. After centuries of research, the scientific bases underlying medicine have become quite well understood. These bases, which fall under the umbrella of biological sciences, are either various direct subfields of biology (e.g. cell biology, molecular biology, genetics, physiology, etc.) or fields where other natural sciences are used to explain biological phenomena (e.g. biochemistry, biophysics). The application of contemporary biological science knowledge (with a heavy focus on physiology) to the practice of medicine is referred to as biomedicine. The field of biomedicine has proven to be an effective approach to practicing medicine, as the interplay of the various biological science subfields are able to address the different focuses of medical practice. Additionally, biomedicine has been made more effective via the utilization of biotechnology - notably, biotechnology stemming from genomics (the study of gene structure, function, and mapping within genomes). Employment of biotechnologies such as biological data-extraction tools (i.e. sequencing or gene expression analysis platforms) and bioinformatics (an applied field of informatics which focuses on analyzing and drawing conclusions from extracted biological data) have allowed for medical questions to be addressed from a genetic perspective. This shift in perspective provides a more concrete foundation upon which to base physiological observation – and therefore, medical findings. Fortunately, the state of modern genomic biotechnologies enables their use within reasonable cost, effort, and time. Ultimately, the use of these biotechnologies has substantially bolstered biomedical practice and research – an already effective approach to medicine - by enabling genetic solutions to medical problems in an effective, yet relatively efficient manner.

Pharmaceutical Sciences Next to biological sciences, another set of sciences utilized heavily in modern medicine are pharmaceutical sciences - a family of scientific fields concerned with understanding various different aspects of medical treatments. At the center of pharmaceutical sciences is pharmacology – the field concerned with studying the effects of biologically-active chemical compounds within human biological systems (i.e. the different organ / tissue systems which make up the human body). When a chemical compound is said to be biologically-active, it means that its introduction to a biological system invokes a change to the functional mechanisms within that biological system. In other words, biologically-active chemical compounds (a.k.a. drugs) are compounds which induce physiological effects. While molecular in nature, these physiological effects cascade and manifest as perceivable macro-effects (e.g. drunkenness via alcohol, pain-relief via morphine, blood-thinning via anticoagulants, or the beneficial effects of self-describing drug families like antipsychotics or antidepressants). At its most basic, pharmacology is concerned with researching what sort of effects any given drug induces within a biological system. The consequences of an induced physiological effect within a biological system can range anywhere from beneficial to detrimental, with a possible magnitude spanning from minor to significant. Among this spectrum lies the potential to find drugs which act as therapeutic treatments for adverse health conditions. However, drugs typically induce a multitude of effects - all existing anywhere on the aforementioned spectrum of consequence. All effects a drug induces result from the molecular interactions a drug partakes in – either stemming ends from the same physiological effect, or divergent effects of a drug with nonspecific properties. Pharmacology is duly involved here - researching how drugs exert their resulting effects at the molecular level. A drug's action is the result of how it biochemically interacts with a biological system – these being the drug's so-called mechanisms of action. These mechanisms of action encompass all interactions occurring between bioreactive portions of a drug and molecular targets in a biological system. When these mechanisms of action are understood, an understanding of how certain drugs work (or even further, an understanding of the grounds underlying their properties) can be gleaned. This pharmacological understanding together with other branches of pharmaceutical sciences (e.g. pharmaceutical chemistry and pharmaceutics) enable the ability to compositionally curate drugs with some intended purpose in an optimal form. It is for these reasons that pharmaceutical sciences are critical in modern medicine – as they stand as a means toward creating potential treatments to ailments with a maximal net positive impact.

Drug Discovery and Development The modern medical approach to treatment development brings together the biological-explainability of biomedicine and the treatment-centric knowledge of pharmaceutical sciences to discover effective therapeutic treatment compounds in a process known as drug discovery. Simply put, the drug discovery process consists of identifying compounds which elicit notable therapeutic properties, followed by testing to verify the feasibility of their practical use as therapeutic treatments (where in

this sense, feasibility is contingent upon a compound's efficacy and safety).

Early drug discovery centered around the identification and isolation of the active ingredients within known natural (i.e. traditional) treatments. Once isolated, these compounds could then go on to be used in new, curated treatments. While most of these early discovered treatments had origins in traditional medicines, some treatments would be discovered from novel origins via serendipitous discovery – most notably being the accidental discovery of Penicillin.

As the sciences underlying it developed and became more sophisticated, so did drug discovery itself. The most integral change came as a shift toward synthesized treatments – a shift which became possible due to advancements in pharmaceutical sciences. With an ever-expanding understanding of how compound composition affects therapeutic property (via pharmacology) and the ability to synthesize potential treatment compounds (via pharmaceutical chemistry), drug discovery became focused on screening (i.e. probing or exploring) a range of synthesized, potentially therapeutic compounds for desired activity with the goal of discovering compounds with clinical feasibility. In addition to the progress in treatment discovery, methods of compound testing improved – particularly in regards to biological assay (bioassay) methods. Bioassays are a form of quantitative testing in live systems where biological activity response is measured upon introduction of a compound. Their use is pervasive across the drug discovery process, as they are the means by which compound performance (i.e. potency in primary assays, or specificity, efficacy, toxicity in secondary) is assessed. Over time, bioassay mediums became better refined – going from in vivo testing (i.e. testing upon a whole organism), to ex vivo testing (i.e testing external to an organism, such as upon extracted limb/organ/tissue samples), eventually to high throughput in vitro testing (i.e. cell line culture testing, done quickly in parallel). The sum of these advancements in the drug discovery process have allowed pharmaceutical companies to effectively synthesize, explore, and test a seemingly countless number of exploratory compounds for feasible clinical candidacy. Compounds which are found to be feasible candidates for clinical use are subject to drug development, which consists of further refining of the candidate compound and rigorous testing – ultimately resulting in clinical trials which in turn, validate or invalidate a drug's clinical use.

The fate of the countless compounds explored in the drug discovery and development process is contingent upon their ability to meet efficacy and safety expectations – an ability which is assessed at various points of testing. This fate is often rejection (and subsequent shelving/abandonment) by testing failure in either the discovery phase (e.g. inability to meet target selectivity or potency expectations) or developmental phase (e.g. clinical trials). Sometimes (comparatively infrequently to the number of rejections), the fate is acceptance and subsequent clinical use. Occasionally, however, compounds which fail in a given therapeutic use-case can exhibit unforeseen side effects which can be potentially therapeutic in another application – thereby relaunching the drug into testing for said application. This act is known as drug repositioning.

Drug Repositioning Drug repositioning is the act of taking a drug with an intended therapeutic purpose and reapplying it as a treatment to another therapeutic end. Much like Penicillin, the discovery of these drugs is serendipitous; since the discovery of repositioned drugs occurs whilst pursuing another objective, they are the product of fortunate happenstance. Despite its coincidental nature, drug repositioning has occurred quite frequently – with Sildenafil being the most prominent example. While pursuing a treatment for angina pectoris (chest pain resulting from insufficient blood flow to the heart), Pfeizer came to discover and develop Sildenafil. The compound was aimed to treat angina pectoris via stimulation of smooth muscle relaxation in the heart by promoting the presence cyclic guanosine monophosphate (cGMP) via selective inhibition of phosphodiesterase 5 (PDE5), a cGMP degrative enzyme. While theoretically sound, the molecule did not meet therepeutic expectations in clinical trials. However, a noted side effect in some trial cases was unwanted penile erection (an occurance due to the role cGMP and PDE5 play in penile muscle action). This relaunched Sildenfil back into testing as a treatment for erectile dysfunction and in 1998, it was approved for sale and marketed under the brand name Viagra. [1]

Many other examples of serendipitous drug repositioning exist, such as:

1. Thalidomide - a treatment initially used to alleviate morning sickness, was discontinued due to negative side effects, and later repurposed as a treatment for leprosy complications and various cancers. [2]

2. Chloropromazine - an intended potentiator of general anesthesia which was repositioned after realized potential as a treatment to various psychotic disorders. [3]

3. Imipramine - an intended treatment for schizophrenia which was repositioned after realized potential as a treatment for depression. [3]

In all, these examples of drug repositioning demonstrate its undoubtedly useful role in drug discovery. However, the serendipitous nature of these repositionings leaves their occurrence up to fate.

Given the importance of serendipity in drug discovery via drug repositioning, imagine if some grounds behind this serendipitous discovery which could be realized and exploited. What if there were a way to search compounds – be them exploratory failures, failed candidates with unobserved side effects, or outright successful drugs – for potential repositionability? Surely, such a tool would be of great use - as it would allow for discovery of new treatments without broadening the already vast scape of drugs already explored. Fortunately such a tool has been created - the Connectivity Map. By now, the importance of such a tool is apparent - but before addressing the Connectivity Map itself, the functionality underlying it must be understood.

1.3.2 Genetics

The complex functioning of biological systems is characterized by the ability of its primary units (cells) to coordinate and carry out intricate actions. These actions are run mainly by the use of proteins – who have origins in genes.

Genes and Proteins In their most common interpretation, genes are defined as the inherited genetic units which dictate the characteristics of an individual organism. This definition has its roots in the work of Gregor Mendel, whose landmark research on the inheritance patterns of pea plants is considered to be the basis of classical genetics. In his work, Mendel hypothesized the presence of certain inheritance factors' despite not having any knowledge of the molecular foundation. By the 1960s, it was recognized that these 'inheritance factors' are genes – sequences of DNA found within chromosomes which hold instructions for the synthesis of different proteins. Although a major element of their existence, being the vehicle for trait inheritance is not the main purpose of genes. On a fundamental level, genes exist as a means to store templates for protein creation.

As the so-called macromolecular machinery of life, proteins are absolutely critical to a cell's (and therefore a biological system's) functioning. From conducting general intracellular actions (e.g.. enzymes required in nucleic acid replication, ATP synthesis, membrane transport proteins) to structural elements (e.g. microtubules, desmosomes, or even larger scale tissue binding compunds like collagen), to defense mechanisms (e.g. antibody generation), to cell specific necessities (e.g. digestive enzymes in stomach cells necessary for breaking down food) to a myriad of other niche uses, they pervade every facet of cellular function and are critical to any living organism.

It is important to note is that proteins are typically very complex molecules, whose shape and function are dictated by interactions arising from their primary structure – known as polypeptides. These polypeptides are simple chains of amino acids, which are actually the product encoded within genes. The process of reading genes and assembling polypeptides is referred to as gene expression.

Gene Expression Gene expression is the process by which genes are read to synthesize proteins (or other functional gene products, like RNAs). It is not a simple DNA-to-product process. Rather, it consists of two subsequent subprocesses named transcription and translation.

Transcription deals with the reading of the DNA template and the synthesis of an intermediate form of RNA known as messenger RNA (mRNA). In this step, an enzyme known as RNA polymerase recognizes the promoter region of a gene (a region of DNA recognizable by RNA polymerase as a start region of transcription) and binds around it. Once bound, RNA polymerase moves along the DNA strand, and unzips the small pocket of the DNA encased by the RNA polymerase – exposing single stranded DNA. From here, RNA polymerase uses the single stranded DNA as a template for creating

mRNA by stringing together complementary nucleotides to the template DNA, until a signal for termination has been received – causing the created product to disassociate from the DNA. After some final processing steps, a finalized strand of mRNA is made and exported to the cell's cytosol.

Translation (which occurs in the cytosol) then makes use of this mRNA to orchestrate the binding of amino acids in an encoded order to create a certain polypeptide. In translation, a cellular component called a ribosome binds to the mRNA. From here, the nucleotides of the mRNA are read in sets of three (called codons) starting from a start codon. Codons are recognized by transfer RNAs (tRNA), which detect one unique codon and carry one particular amino acid. From here, amino acids are stringed together as tRNA read the codons occuring along an mRNA until a stop codon is reached – signaling synthesis completion and polypeptide disassociation. These polypetides (i.e primary structure) undergo topological changes - taking on α-helix or β-sheet structuring (i.e secondary structure) due to hydrogen bonding between peptide group interactions, which fold and globularize (i.e tertiary structure) due to hydrophobic interactions of these structured forms. Various tertiary forms may even go on to aggregate and form ensembles (i.e quaternary structure). Whether tertiary or quaternary – these imposed topologies ultimately indicate and promote the functionality of a protein.

Gene expression is not an unregulated process - cells are constantly acting to control which genes are being expressed, and even to what extent. These actions taken by cells to control their gene expression fall within a group of actions known as gene regulation.

Gene Regulation Gene regulation refers to the series of mechanisms which cells employ to control the expression of their genes. These mechanisms act as points of control along the entire gene expression pipeline - with different regulatory mechanisms potentially acting anywhere from transcription, RNA processing, RNA transport, translation, to even RNA degradation. These regulatory actions allow for cells to express genes only when necessary. Additionally, the modular employability of these control mechanisms (along with some mechanisms allowing various levels of regulation) allows for gene expression to occur in a non-binary state; Depending on the employed control mechanisms, genes exhibit varying degrees of expression rather than binary states. This allows for genes not only to be expressed when necessary, but to a necessary extent.

The ability of a cell to regulate the expression of its genes is quite critical to its existence. This is primarily due to the inefficiency of unregulated gene expression (w.r.t. energy and synthesis resources), which makes the act an unsustainable means of existence. Luckily, simultaneous expression of all genes in a cell is completely unnecessary – in fact, only a fraction of a cell's genes are needed to be expressed at any given time. Due to this, cells employ necessity-based gene expression (via gene regulation) to optimize resource use. However, what defines "necessary" varies from cell to cell – a fact which causes cells to exhibit *characteristic expression*.

Characteristic Expression and Gene Expression Profiling In most cases, the genetic content of the different cell types comprising a given multicellular biological system is approximately the same. This applies to the human body; with some exceptions, the cells which comprise the numerous tissue/organ types of the human body's various subsystems (e.g. respiratory, digestive, circulatory) all house roughly the same DNA. This common DNA equates to ∼20,000 mutually present genes, whose encoded products are responsible for many complex actions (ranging from general housekeeping actions to more specific actions).

Despite having the same genes, cells often express most of them differently. This is because cells exhibit *characteristic expression* – meaning that their gene activity (i.e. the set of genes expressed, magnitude of expression) is characterized by the cell's overall situation. Characteristic expression occurs due to the necessity-based gene expression which cell's carry out. The genes which a cell deems necessary to express (and to what extent) is contingent upon various factors relevant to a cell's circumstances (e.g. cell type, location, cell age, disease, environment). All of these factors combine to impact a cell's gene activity; individually, they instill gene activity which is characteristic of a factor (e.g. gene activity tied to a cell type's functional purpose, aging-related gene activity) – but together, their impacts intermix and aggregate to instill gene activity which is characteristic of a cell's overall situation.

With this knowledge, the opportunity exists to elucidate the characteristic gene activity underlying individual factors by observing *differentially expressed genes*. Consider the following: Not much can be gleaned by observing two cells under the same set of circumstantial factors – genes would be expressed similarly, and there is no way to attribute certain gene activity to one particular factor. However, if the two cells differed by one factor (ex: cell type), some genes between the two would certainly exhibit differential expression. Recognizing the differential gene expression caused by a factor lends information about its characteristic gene activity. Once studied enough, the factor's characteristic expression is understood – which allows other factors (e.g. disease in a given cell type, drug exposure in a given cell type) to be studied, understood and ultimately leveraged toward advancement on a certain end (e.g. biological, medical, pharmacological).

Luckily, this opportunity can be taken, as *gene expression profiling* technology exists which makes widespread gene activity observation possible. Gene expression profiling is a measurement technique which returns the relative activity of the genes in a cell. It is typically done over an entire target genome in order to deliver a global scope of expression activity in a cell. Once retrieved, a *gene expression profile* provides a snapshot of a cell's global characteristic gene activity – exhibiting characteristic expression influenced by all of its circumstantial factors. Among the methods available to carry out gene expression profiling, the focal technique discussed herein is the DNA microarray.

1.3.3 DNA Microarrays

DNA microarrays are a technology which measure relative concentration of many different DNA sequences present in a given sample (e.g. a cell, or cell culture). This is achieved via their substrate-bound arrangement of known single-stranded DNA (ssDNA) sequences - known as *probes*. When exposed to a prepared sample, complementary ssDNA from the sample (known as *targets*) hybridizes to the probes, causing a perceivable signal to emit per probe. Observation of these probe signals enable relative concentration measurement of the different targets assayed by the microarray.

Among many established experimental uses of DNA microarrays, one of their most well-established uses is gene expression profiling. This is achieved by measuring the relative concentrations of a sample's various mRNA – a task which is carried out indirectly via reverse transcription of the sample's mRNA into complementary DNA (cDNA). This cDNA is in turn exposed to a microarray which assays the cDNA fragments as targets for hybridization, ultimately allowing concentration measurement of the sample's various mRNA (therefore, gene activity).

Herein, we are concerned with a line of commercially-available microarrays offered by the company *Affymetrix*, called the *GeneChip*. Moreover, we are explicitly interested in their use toward gene expression profiling.

Structure Structurally, the GeneChip DNA microarray is a 2D-arrangement of hundreds of thousands of microscopic regions called *probe cells*. Each probe cell contains millions of copies of a single known, short (∼25bp) ssDNA sequence (known as a *probe*). These probes are complementary sequences to sections of known genes making up the sample's genome. The purpose of probes are to hybridize their potential cDNA counterparts (known as *targets*) found in the sample, which traps them for measurement of their concentration in the sample. As these probes only represent portions of different genes, they are designated into groups called *probe sets*, which as a group represent a gene. A depiction of the GeneChip's anatomy can be found in figure 1.1.

From a scale perspective, the working surface of GeneChips are small ($<2\text{cm}^2$). However, the scope of these GeneChips are extensive; Affymetrix offers GeneChips spanning various species of microorganisms, plants, and animals – including humans. In each of these GeneChips, the probe sets span the entirety of genes found within their respective genomes.

Target Detection and Measurment While microarrays provides the structure necessary to trap (i.e. hybridize) targets, the ability to observe presence and measure concentration is due to the coupling of dyes to the cDNA fragments of the sample. GeneChip experiments utilize a biotin-based dye, in which biotin molecules are bound with other factors which render them fluorescent. Prior to exposing sample cDNA to the microarray for hybridization, they are affixed with these biotin-based compounds. Due to

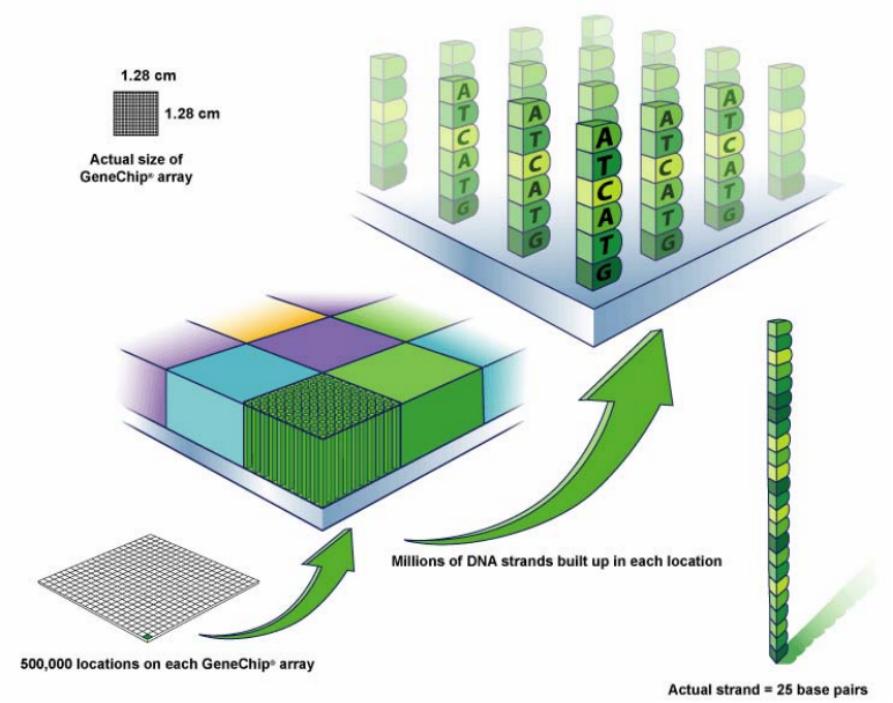


Figure 1.1: GeneChip Anatomy
The structure of a GeneChip – displaying how probes, probe copies, and probe cells all interrelate upon the miniature microarray

Source: [4]

this, probes which successfully hybridize target cDNA will emit a fluorescent signal. This signal not only allows for identification of present targets, but for measurement of their concentration – as signal intensity will vary based upon the number of targets hybridized by a probe's many copies in a probe cell.

Experimental Protocol The protocol for obtaining target concentration measurements from a given sample in a GeneChip-based gene expression profiling experiment is as follows:

1. **Isolation** of mRNA from sample
2. **Reverse Transcription** of mRNA into cDNA
3. **Fragmentation** of cDNA into probe-equivalent size
4. **Coupling** of fluorescent dye to cDNA fragments
5. **Hybridization** of cDNA fragments to microarray probes
6. **Washing** away remaining unhybridized product
7. **Scanning** of post-hybridization chip into images
8. **Processing** of images into probe-level data

The end result of this protocol is a CEL file - the digitized representation of probe signal intensities resulting from a GeneChip reading. Furthermore, as gene expression profiling is typically done toward a comparative end, this protocol is applied in tandem to multiple samples – yielding multiple CEL files.

As these CEL files only represent probe-level intensities, further processing is required in order to obtain relative gene expression values.

Data Processing The conversion of probe-level CEL files into gene expression values is typically carried out over a few computational steps in a data processing pipeline. Firstly, a set of data-adjustment procedures are regularly utilized which act to adjust probe data toward a "truer" form (as systematic variation frequently occurs in/between CEL data). These procedures are ultimately followed by summarization – the summing up of probe intensities into probe set (i.e. gene expression) values. From front to back, the steps in the processing pipeline are:

1. **Background Correction:** The removal of "background" (BG) (i.e. non-biological contribution) from each probe intensity signal. BG occurs for various reasons (e.g. nonspecific hybridization, incomplete washing, spacial heterogeneity, lab/system conditions). Its removal leads to truer intensity values.

2. **Normalization:** The adjustment of probe intensity values across individual arrays toward an equatable plane. Various systematic factors (e.g. lab/system/experimental conditions) can influence the overall magnitude of intensities measured by an individual microarray. Normalization acts to alleviate this; it brings values across CEL files to an equatable level, ensuring intensity values across CEL files are truly comparable.

3. **Perfect Match Correction:** A unique feature to Affymetrix microarrays – each probe (referred to as the perfect match (PM) probe) has an associated imperfect-match probe (known as a mismatch (MM) probe). Like BG correction, PM correction acts to remove signal due to nonspecific hybridization. It does this by adjusting PM probe signal with reference to MM probe signal – thereby leading to truer intensity values.

4. **Summarization:** The act of summing up probe intensity values within probe sets into singular intensity values. This is guided by a CEL file's associated Chip Definition File (CDF), which states the addresses of each probe set's probe cells in a CEL's origin chip type. As probe sets are equivalent to genes, this is effectively the same as to obtaining relative gene expression values.

Affymetrix offers solutions for all of these procedures, which are available under their "Microarray Analysis Suite" (MAS) toolset. However, various third-party solutions to these procedures exist and are available for employment (e.g. RMA as an entire

pipeline solution, Quantile and Cyclic Loess for Normalization, Li-Wong and FARMS for Summarization).

No matter the methods chosen, the result of summarization is gene expression data encompassing the genome of a sample in question – thereby being the sample's gene expression profile. In practice, the yielded product is typically multiple sample's gene expression profiles (as in practice, multiple CEL files are passed through processing).

Benefits of Microarrays Prior to DNA microarrays, DNA/RNA assaying were already instrumental steps in many forms biological and applied-bio analyses (e.g. biomedical, pharmacological). These steps greatly benefited from the arrival of the DNA microarray; primarily, they granted the ability to carry out mass parallel profiling of genes in a sample. Furthermore, the advent of *miniaturized* microarrays (i.e. chip-sized working surfaces) enabled higher throughput experimentation with less required resources. Since their widespread adoption and commercialization in the late 90s, microarray platforms saw further improvements; they became easier to implement, quicker to run, less resource-intensive, more robust in measurement, capable of higher throughput, and even parallelizable (e.g. parallel runs of numerous microarrays simultaneously). On top of this, their long history of employment makes their use well-established. The sum of these traits make them a powerful tool in genetic analysis. [5]

With regards to gene expression profiling, other technologies have since emerged which can also accomplish the task (namely *RNASeq*). Despite this, DNA Microarrays remain a worthy choice in this application. Firstly, their protocol from sample to data is well-established and straightforward in comparison to the freshness and complexity of profiling via RNASeq. Secondly, RNASeq in this application is excessive; RNASeq holds advantage when the identification of novel transcripts are sought. However, as the sequences of genes are already known in profiling experiments, it doesn't hold much advantage. Lastly, an RNASeq experiment is more costly than that of a microarray – a fact which is amplified as sample count grows. [6]

It is for these reasons that DNA microarrays remain a sound choice for gene expression profiling experiments, and are even the preferable for large scale gene expression profiling experiments.

Chapter 2

The Connectivity Map

The Connectivity Map (CMap) is a biomedical research tool conceptualized by Justin Lamb and scientists at the Broad Institute, whose purpose is to provide a platform for the discovery of connections between a given biological state (e.g. disease) and bioactive small molecules (i.e. drugs) via genes (or more specifically, the relative expressions of them) [7].

The main aim of the CMap is to simplify the task of finding these disease-gene-drug connections. Traditionally, pursuing the identification of these connections requires extensive domain knowledge in clinical medicine, molecular genetics, and chemistry – three distinct, yet deep disciplines. The CMap bridges these disciplines and unites them in a general analytical space via DNA microarray-led gene expression profiling. Using gene expression profiles as a universal language to encapsulate representations of disease and drug action acts to simplify the process of finding connections, thereby narrowing matters down to pattern matching between the expression values of genes.

Another goal of the CMap was to systematize and centralize a notably laborious portion of many chemical genetics based research projects – the screening process [8]. In much chemical genetics based research (e.g. drug discovery), there exists the need to screen ever-growing chemical and genetic libraries. The combination of the size of these libraries, mixed with the required scanning of these libraries with many individual readouts in a sequential and scattered manner ultimately makes this process tedious and time consuming (and therefore highly demanding). The proposed solution devised by Broad was to create a comprehensive catalog containing the global genetic effects of as many bioactive small molecule treatments as possible (stored in the aforementioned universal language of gene expression profiles). This small molecule treatment catalog is integral to the functional purpose of the CMap – it acts as the foundational dataset necessary for connections to be found.

The method by which the CMap goes about revealing connections is wholly dependent upon input data provided by a user. Connections are elucidated with respect to a user-provided query gene expression signature (a set of genes which correspond to some biological state). This query signature is then compared against each treatment instance

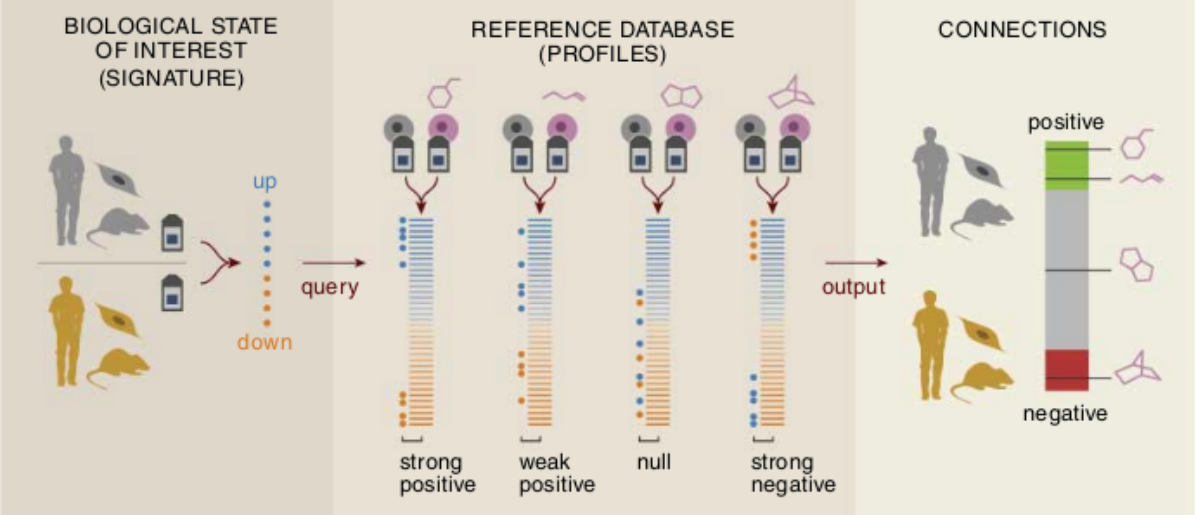


Figure 2.1: CMap Querying Concept
A query gene expression signature is checked against a database of treatment profiles, yielding the treatments sorted based upon connectivity.

Source: [8]

gene expression profile contained in the aforementioned small molecule treatment catalog. These comparisons lead to a scoring of each of the treatment instances via a similarity metric – yielding a means of sorting them from those that have high positive connectivity (i.e. best mimic the gene expression of the query – promoters of the signature's condition) to those that have high negative connectivity (i.e. those that act most oppositely to that of the query signature – acting against the the signature's condition). This process is referred to as querying the CMap (visualized in Figure 2.1).

In a practical setting, the CMap acts as an assistive tool in the field of drug discovery. Most notably, it has seen much use in drug repositioning. The value of the tool in practical use (as well as the value of the concept underlying it) have been well corroborated; the CMap publications have been cited in thousands of works – ranging from discoveries of potential repositioned theraputics [9, 10, 11], to elucidations of the molecular mechanisms underlying established medicines [12, 13], to adopters of the connectivity mapping analysis paradigm who have curated connectivity datasets and/or tools of their own [14, 15]. Overall, the project has received a remarkable amount of attention from the biomedical community and has seemingly proven its worth.

When referring to the Connectivity Map, it should be thought of as its set of parts – its dataset, its querying algorithm, and its user interface.

2.1 Dataset

The manifestation of the proposed small molecule treatment catalog comes in the form of the CMap's dataset. It is a curated collection of gene expression profile data obtained from human cell line cultures treated with bioactive small molecules (along with corresponding

control gene expression profile data of untreated cell lines in their relevant vehicle solution). The collection incorporates expression profiles over a space of experimental variables, where each unique combination captured by an expression profile is referred to as an *instance*. This space of variables includes:

- Different bioactive small molecule treatments (referred to as *perturbagens*) which range from FDA-approved drugs to bioactive research compounds.
- Varying concentrations and exposure times, which are measured in order to obtain optimal, robust signals (yet also to explore the sensitivity of expression to dose)
- Several human cell line types, which are employed in order to get a global picture of genetic effects.

Whilst pursuing the envisioned comprehensive catalog of small molecule treatments, the scope of the CMap's dataset grew considerably – jumping from 453 instances across 164 perturbagens in its initial pilot build, to 6,100 instances across 1,309 perturbagens in the CMap's first update (referred to as *build02*).

The obtention of these instances' gene expression profiles (and their related control expression profiles) from their instance samples was done using Affymetrix Human-Genome GeneChip microarrays. In build02, a combination of three GeneChip types were used in carrying out the profiling of the treatment instances (detailed in Table 2.1). The scanning of these GeneChips yielded 7,056 CEL files spanning all perturbation and control profiles comprising the treatment instance catalog of the build02 CMap. These CEL files are made available via Broad's CMap web portal.

GeneChip	Sample Capacity	# Associated Control Scans	# Probesets On Chip	# CELs in Data From Type
HGU133A	1	Single	22,283	807
HT-HGU133A	96	Multi	22,277	6,029
HT-HGU133A-EA	96	Multi	22,944	220

Table 2.1: Descriptions of the three GeneChip types from which the CMap build02 data originates.

In order to be utilized, these CEL files must be processed into instance-level gene expression data. In the case of Broad's implementation, the initial processing of CEL files into gene expression values was done using the Affymetrix Microarray Suite (MAS 5.0). This is followed by comparison of the gene expression values between each instance's perturbation scan with its control scan(s) - yielding the average difference of each probe set's expression, per instance. Finally, each instance's probe sets are ranked from their highest to lowest average difference – a nonparametric format of instance-level expression data which is necessary in order to use the CMap's nonparametric querying algorithm.

2.2 Querying Algorithm

The service of finding disease-gene-drug connections is provided by the CMap's querying algorithm; it carries out pattern matching between a query gene signature and each instance of the CMap dataset. The algorithm scores each of the CMap instances by the extent of their similarity to the query signature, acting as a means by which the most-similarly and most-oppositely expressed instances to the query signature can be retrieved. In order to obtain these similarity scores for each instance, the querying algorithm utilizes a similarity metric called Connectivity Score.

Connectivity Score (formulated in a prior Broad Institute project titled Gene Set Enrichment Analysis [16]) is a rank-based nonparametic statistic based upon the Kolmogorov-Smirnov statistic [17]. Its use requires a query signature comprised of up- and down-tag lists – two lists containing ordered up- and down- regulated genes (à la probe set IDs) which correlate to a biological state of interest. A series of calculations are done per instance to obtain each instance's up- and down- scores (as seen below).

$$a = \max_{j=1}^{t} \left[\frac{j}{t} - \frac{V(j)}{n} \right] \tag{2.1}$$

$$b = \max_{j=1}^{t} \left[\frac{V(j)}{n} - \frac{(j-1)}{t} \right] \tag{2.2}$$

$$ks^i = \begin{cases} a & a > b \\ b & b > a \end{cases} \tag{2.3}$$

For a given an up- or down- tag list of length t being checked against instance i, Equations 2.1 and 2.2 are solved – where j is a subjective tag index out of t tags, *V(j)* is that tag's ranking in the ranked instance-level data, and n the number of features (probe sets) in the instance data. Once solved, the tag list's score is decided in Equation 2.3. This is carried out for both tag lists in the query signature – resulting in up- and down-scores (ks^i_{up} and ks^i_{down}, respectively) for an instance. These scores, which represent the enrichment of the up- and down- tag list in a given instance, are used further to calculate an instance's connectivity score (as seen below).

$$S^i = \begin{cases} 0 & sign(ks^i_{up}) = sign(ks^i_{down}) \\ ks^i_{up} + ks^i_{down} & sign(ks^i_{up}) \neq sign(ks^i_{down}) \end{cases} \tag{2.4}$$

Connectivity score (as seen in Equation 2.4) is the sum of the up- and down- scores (unless up- and down- score signs are the same, in which case an instance has null connectivity). After calculation for all instances, the connectivity scores are normalized on a [-1,+1] range. The connectivity scores represent the relative strength of a signature in a given instance – where the highest ranked instance (score = +1) is the most positively correlated CMap instance to the query signature, and the lowest ranked instance (score

= -1) is the most negatively correlated CMap instance to the query signature. These scores act as a means by which the list of CMap instances can be returned sorted by the strength and direction of their relationship to a user-provided query signature. The entire process of a user utilizing the querying algorithm (i.e. providing an input query, initiating the query and receiving sorted results) is carried out via the CMap's user interface.

2.3 User Interface

A lesser discussed – albeit critical – aspect of the CMap is its user interface (UI). The CMap's UI is the means by which a user can utilize the formulated querying algorithm and dataset to their own scientific benefit. Given that the intention behind the CMap was to ease the biomedical research process, the UI to its services should reflect this goal. To this end, it is crucial that the UI's design be made with clear-use and accessibility in mind - that if any bench researcher wanted to carry out CMap analysis on an in-house microarray experiement, they should be able to do so without confusion or extensive training.

With these necessities in mind, the build02 UI for the CMap is serviceable. The tool offers a sizable help section consisting of a sample query-guided tutorial and a comprehensive glossary of topics which span the entirety of the CMap – the combination of which yields an extensively detailed guide to the tool's use. The querying functionality (loading queries, performing queries, viewing query results) is present via different menu options, with query results being saved to the server. The tool offers a multitude of additional options - from generating a CMap instance-based query, to revisiting and exporting past query results, to browsing CMap instance details.

2.4 Proposed Adaptations and Inclusions

We conceived a number of potential avenues toward improving the CMap. Focally, these aim to heighten the tool's accessibility. Additional avenues have been formed which when leveraged, could improve the tool's performance. In all, these avenues consist of:

1. **Redesigned UI:** Implementing a UI with a more intuitive, fluid workflow.
2. **Adapted Microarray Processing:** Modernization of GeneChip probe definitions (via **custom CDFs**) and utilization of **FARMS** (*Factor Analysis for Robust Microarray Summarization*) in processing.
3. **Query Signature Expansion:** Allowing various gene identifier types other than probe set IDs to be used in querying.
4. **Adapted Querying Algorithm:** Implementing a querying algorithm and scoring method which serves a more direct comparison of the fold changes between query and instance.

2.4.1 Redesigned UI

While the UI provided in the Broad Institute's build02 of CMap does serve its purpose - it is in some ways suboptimal; the functionality offered is scattered behind different menu choices, making navigation quite cumbersome. Also, instructions behind its use are considerably lengthy. We decided to provide a simplified UI to enhance user-friendliness. The solution decided upon was a single page application which contains the querying pipeline in its entirety – from importation of the query signature, to returning and saving query results in a user-defined form – along the way providing clear control options, relevant status updates and concise instructions to guide the user through the querying process. It is felt that this redesign would make the tool considerably more accessible without losing any functionality.

There exists an R package named *shiny*, which allows for the creation of interactive web applications from within the R statistical computing environment. The package is essentially a catalog of functions for generating HTML, CSS, and JavaScript code - the use of which requires minimal knowledge of these languages. This stands as a major convenience for any data scientist with the desire to present their work in an interactive medium, but lacks front-end development experience. Considering that we had decided to utilize R in this project due to its well-established connection with biological data analysis, it came as an obvious choice to leverage shiny as the means by which the UI for the tool will be implemented.

2.4.2 Custom CDFs

One well-established way of improving the performance of Affymetrix GeneChip analysis is through the use of custom CDFs. Custom CDFs are a means of addressing a glaring issue of the microarray platform - the gradual obsolescence of any chip's probe set-gene mappings. When any given microarray chip is designed, the decisions behind their probe cell / probe set assignments are based upon the latest genomic annotation information available at the time. However, due to the rapid progression of genomic and transcriptomic research, a probe set's definition may become outdated due to updates in the gene annotation relevant to that probe set. In the case of Affymetrix GeneChips, many have fallen victim to changes in annotation and are left with outdated probe set identities. For example, the HGU133A GeneChip was created when the human genome was only 25% sequenced [18] - demonstrating the large potential gap between probe set design and the present state of genome annotation. Unfortunately, Affymetrix seldom provides official updates to their chips' probe set definitions.

However, there exists a means by which probes can be reorganized into specified probe sets – via the CEL file's CDF. By default, stock CDFs are used when CEL data is loaded in. But as CDFs are open for redefinition, they have potential to be updated to current standards, thereby improving the ability of Affymetrix based gene expression analyses to

GeneChip Type	# Probe Sets in Stock CDF	# Probe Sets in Brainarray CDF
HGU133A	22283	12322
HT-HGU133A	22277	12316
HT-HGU133A-EA	22944	12792

Table 2.2: The GeneChip types used in build02 and their probe set counts between both their stock CDF and corresponding Brainarray CDF

provide relevant, up-to-date insights.

Brainarray CDFs [19] are one such attempt at creating custom CDFs which reorganize probes into probe sets based on modern genomic and transcriptomic knowledge. The process behind creating these CDFs starts with the mapping of all probes on a chip to entries of a relevant sequence database (e.g. Entrez Gene), dbSNP, and the sequenced genome of the corresponding species – where only perfect matches are retained. These hits in the related sequence database provide a target gene, acting as the basis for the new probe sets. Various filtration steps are applied to these mapped probes, which enforce a number of standards: (i) Probes must have only one perfect match in the corresponding genome sequence; (ii) Probes must have a perfect match to an mRNA/EST sequence in the relevant sequence database; (iii) Probes must not have matches with cDNA/EST sequences of different genes; (iv) Probes of a probe set must be mappable to the target sequence in a unidirectional fashion; (iv) Probe sets must consist of ¡3 probes. Updated CDFs are created on a yearly basis, over a range of different focal sequence databases. Ultimately, these CDFs are recommended as an effective way of improving Affymetrix-based gene expression analyses, as they have been shown to yield more accurate expression levels [20, 21].

With regard to the CMap data, the three chip types used in amassing the build02 dataset all have available custom CDFs which offer newly defined probe sets, thereby affecting their probe set counts (Table 2.2). These custom CDFs are to be utilized into the CEL data processing pipeline.

2.4.3 Query Expansion

In the Broad Institute's implementation, querying is limited to use of probe set IDs. This could prove to be a hindrance in the accessibility of the tool, as it forces any user interested in utilizing it to have the gene signature provided in the form of Affymetrix probe set IDs (implying that the tool is only designed for use with gene expression analyses carried out with Affymetrix GeneChips). Due to this, one conceived improvement of the CMap is to open the query up to allow other gene identifiers – namely, enabling the use of Gene Symbols and Entrez Gene IDs. Luckily, the use of Brainarray CDFs makes implementing this improvement uncomplicated – not only are probe sets unique to a given gene, but corresponding packages to the CDFs are supplied which enable the quick

linking of Brainarray CDF probe set names to various database identifiers.

2.4.4 FARMS

Another often-explored path toward improving the performance of Affymetrix GeneChip analysis is the adaptation of its processing steps. One such adaptation to be applied here is a summarization technique called Factor Analysis for Robust Microarray Summarization (FARMS) [22].

Consider the following: If examining the activity of probes of a single probe set across multiple arrays, it is expected that probe intensities would vary synchronously across arrays. However, this is almost certainly never the case due to the presence of measurement noise. FARMS is an approach to probe set summarization which utilizes factor analysis models to estimate true probe set expression values from observed, noise-influenced probe-level intensities. Factor analysis is a statistical method by which the variance of observed variables is explained by latent (i.e. underlying) variables, known as factors. Factor analysis models of observed variables are composed of the linear combination of factors for a given set of observed variables, plus some error.

FARMS proposes that the sole factor underlying a probe set's observed probe intensities is the true target concentration in the hybridization mixture, where any deviation from this is attributed to the error of measurement noise. As a factor analysis model, this linear relationship between observed probe set intensities x and true underlying concentration of target content z is represented as:

$$x = \lambda z + \varepsilon \tag{2.5}$$

where λ represents the factor analysis loading matrix and ε represents the added measurement noise. The factor z is said to be $N(0,1)$-distributed, where the noise ε is said to be $N(0,\psi)$-distributed (and is independent per probe, per array). As a result, the model stated in Equation 2.5 is $N(0,\lambda\lambda^{\mathrm{T}}+\psi)$-distributed. These model parameters (λ and ψ) are optimized per probe of a probe set (per array) via Bayesian maximum a posteriori estimation as:

$$p\left(\lambda,\psi \mid \left\{x\right\}\right) \propto p\left(\left\{x\right\} \mid \lambda,\psi\right) p\left(\lambda,\psi\right) \tag{2.6}$$

where the posterior probability of the model parameters λ and ψ given observed data $\{x\}$ (being the set $\{x_1...x_N\}$, where N is the number of arrays) is proportional to the product of the likelihood of observed data x given model parameters λ and ψ and the prior probability of the model parameters. Once optimal parameters are found, the factor z of an array can be estimated, from which true summarized signal of a probe set of an array is inferred.

Due to inconsistent noise effects when applying PM correction to low signal probes, FARMS does not utilize PM correction. Furthermore, FARMS was not created with

background correction in mind – therefore, it is not carried out when FARMS is used. The only processing step required prior to application of FARMS is probe normalization – which the FARMS package promotes the use of either quantile or cyclic loess normalization [23].

FARMS is noted for its performance advantage over other methods of Affymetrix GeneChip summarization in terms of sensitivity and specificity [24] – which is to say that it is capable of acquiring high, true signal measurements while being robust against measurement noise. It is for these reasons that it was decided to be utilized in the CMap data processing pipeline.

2.4.5 Adapted Querying Algorithm

The querying algorithm of the CMap also stands as a potential outlet for improvement; A systematic evaluation of the CMap in 2014 showed via implementing and assessing alternative metrics that the tool could benefit from the use of a more effective querying algorithm [25]. Due to this, we devised an alternative to the standard connectivity score based querying algorithm – one which utilizes gene fold-change values directly rather than assigning global relative expression rankings, and measures query/instance similarity through the use of more universal correlation methods. This idea was inspired by research done in 2011 at the University of Montreal, where the evaluation of 800 different analytical pipelines on 100 different microarray experiments lead to the suggestion that simple fold-change comparisons are preferable to the conventional statistical methods commonly employed in microarray analytics [26].

Implementation of this idea requires changes to the structure of both the provided query signature and processed instance-level expression data. Rather than being given as a tag list, the query signature is now required as a listing of genes with their corresponding fold-change values. Instance-level data is to also be processed from the dataset's CEL files into gene fold-change values per instance, rather than their relative expression rankings. As is standard in microarray analysis, the fold-change values in both the query signature and instance-level expression data are log2-transformed [27].

The newly devised querying algorithm is to use a correlation coefficient as the similarity metric between query and instance, where two common forms of correlation are made available for application - Pearson correlation and Spearman's rank correlation.

$$r_{xy} = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.7}$$

$$\rho_{xy} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2.8}$$

Pearson correlation is a parametric test which evaluates the linear relationship between two samples. It results in the correlation coefficient known as Pearsons's r, and is shown in Equation 2.7. To explain it in terms relevant to its use here, it compares a query

signature x consisting of n genes against a CMap instance y's same n genes via those n genes' fold-change values. Simply put, it is a measure of the covariance of the query signature x and CMap instance y in regard to some n genes - normalized by the product of their standard deviations. Ultimately, it delivers a measure of relationship between a query signature and a CMap instance based upon fold-change values directly.

Spearman's rank correlation is a nonparametric version of Pearson correlation which evaluates the monotonic relationship between two variables. It results in the correlation coefficient known as Spearman's *rho* (ρ), and shares Equation 2.7 with Pearson correlation - but also has a simplified no-tie form shown in Equation 2.8. Spearman's correlation differs from Pearson correlation primarily in its observation of variable ranks in a sample, rather than base values. Its explanation relevant to its use here is nearly the same as for Pearson – except for that the n genes in query signature x and CMap instance y are ranked, and these ranks are used in place of fold-change values in calculation.

The rationale behind providing these two correlation methods as options in querying is that both carry their own pros and cons which stem from the way each interprets provided sample data. The nature of Pearson correlation puts a focus on signal strength – a trait which is preferable if a user perceives that observing a query signature's explicit gene fold-change values directly is critical to assessing similarity to CMap instances. Unfortunately, the linear relationship observance of Pearson correlation makes it error-prone in the presence of outliers. Spearman's correlation is by comparison more flexible due to its use of relative ranks over explicit fold-change values. This not only makes Spearman's more robust to outliers, but allows it to evaluate monotonic relationships – a more relaxed relationship evaluation than Pearson's linear relationships. The downside of Spearman's is exactly what differentiates it from Pearson – it experiences a loss in information (magnitude) due to its use of ranks over true value.

Regardless of which is chosen, both correlation coefficients result in a value existing on a [-1,+1] range, denoting the strength and direction of mutual change between a query and a given instance. Like connectivity score, one of these two correlation coefficients is calculated between a user-provided query signature and each instance of the CMap, yielding a means by which the list of CMap instances can be returned sorted by the strength and direction of their relationships to the query signature.

Chapter 3

Methodology

The sum of adaptations and inclusions suggested in 2.4 were implemented in an adapted CMap tool – ultimately titled the *CMap FC-Signature Querying Tool*. Implementation of this was carried out using the R statistical computing environment (v. 3.5.1) [28] along with various tools accessed via Bioconductor (v. 3.7) [29] – a platform which provides means for the analysis of high-throughput genomic data (typically in the form of R packages). The relevant data for carrying out this project was obtained via the Broad Institute's web interface for the build02 Connectivity Map. This relevant data includes the 7,056 CEL files which make up the build02 dataset, and a table (referred to herein as *instances info*) which details the 6,100 drug treatment instances contained in the dataset (i.e. instance treatment parameters, mappings of CEL files to instances).

3.1 Data Processing

Prior to implementation of the querying algorithm and UI, it was first necessary to assemble queryable datasets. The entailed the processing of raw CEL files into instance-level fold-change data, which took place over a series of steps:

1. Processing raw CEL files into a single, unified set of expression data.
2. Processing the unified expression data into instance-level fold-change data.
3. Filtering out dead probe sets in the instance-level fold-change data, and translating it into additional annotation forms for flexible querying.

Step One Depicted in Figure 3.1, step one dealt with the processing of the CMap's raw CEL files into a single unified set of data. In order to achieve this, the Bioconductor package affy(v. 1.58) [30] was utilized - a package which allows for analysis of probe-level Affymetrix microarray data within the R environment. The functionality of this package which makes it practical here is that it allows for the importing of CEL data into the R environment en masse (as a so-called *AffyBatch* objects), as well as the means for processing them into expression values. This means of processing (a function called

expresso) allows for the passing of an AffyBatch through a defined processing pipeline, yielding summarized probe set expression values from the raw probe intensities for each CEL in the AffyBatch (as an *ExpressionSet* object).

The step required the initial splitting of the CELs into three groups based upon their origin chip type. A prerequisite of reading CELs into AffyBatches is that all CELs provided originate from the same chip type. Since the 7,056 CELs of CMap's build02 dataset originate from three different GeneChip types (as seen in Table 2.1), the CELs had be split into three groups and processed separately.

Each group of CELs was loaded into three different AffyBatches. At this point, the corresponding BrainArray Custom CDF(v. 22) [19] to each AffyBatch's chip type was provided. From there, the package FARMS (v 1.32) [22] was utilized per Affybatch. FARMS contains wrapper functions for affy's expresso which allow for the easy deployment of the default FARMS-defined processing pipelines. The wrapper used herein (qFARMS) employed a processing pipeline which forwent background and PM correction, carried out quantile normalization, and finally performed FARMS summarization to transform probe-level intensity data into probe set expression values. Each of the AffyBatches were processed using qFARMS, which yielded three ExpressionSets.

Each ExpressionSet housed a data frame which contained expression values for all probe sets measured by the AffyBatch's underlying chip type, for each CEL within the AffyBatch. The final task was to unite the three data frames from the three ExpressionSets into a single data frame. This was done by limiting each of these data frames to the common set of probe sets found in all three and then concatenating them. The resulting data frame consisted of the expression values of 12,307 common probe sets among all 7,056 CELs.

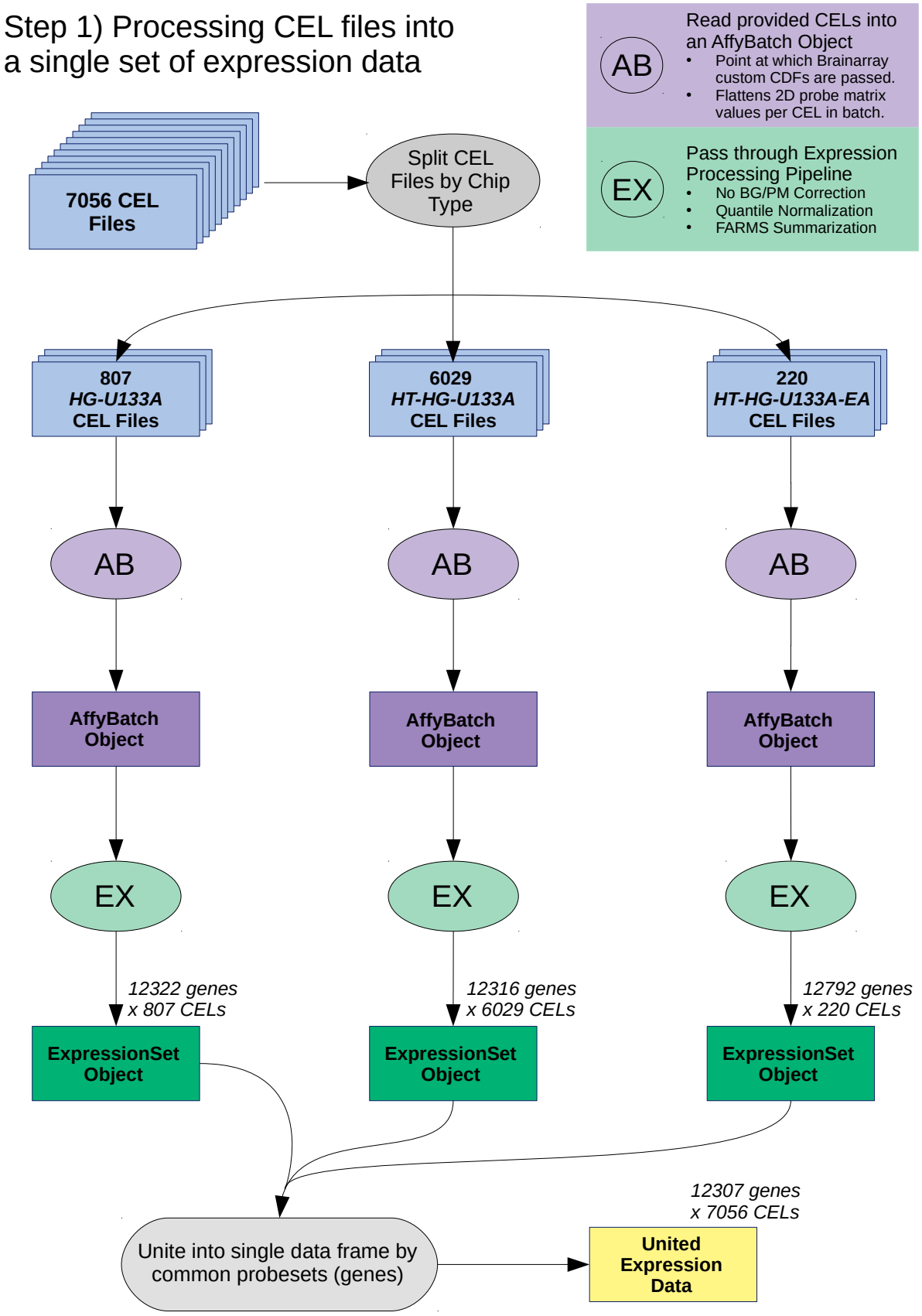


Figure 3.1: Flowchart depicting actions taken in Step 1 of data processing – the processing of CEL files into a single set of expression data.

Step Two Depicted in Figure 3.2, step two transformed the data frame of united expression data from step one into fold-change (FC) values of probe sets, per treatment instance of the CMap. This task required using of the instances info table as a guide for carrying out FC calculations. Each row of the instances info table corresponded to one of the 6,100 treatment instances of the CMap, and contained details as to which CELs act as the control and perturbation scans of each instance. Use of this information in conjunction with the united expression data allowed for the completion of the task at hand.

In practice, the rows of the instances info table were iterated over. Per row, the CEL names for the perturbation and control scans were retained. Due to the various GeneChip types utilized in obtaining the build02 dataset, an instance could either have a single control scan or multiple control scans. The retained CEL names were used to pull the relevant control and perturbation columns from the united expression data (as vectors). If the instance had multiple control CELs, the mean of each probe set across all control vectors was taken – resulting in one single control vector.

FC was calculated as the ratio of perturbation expression to control expression. Calculation of this ratio as the element-wise division of the perturbation vector by the control vector yielded the FC for a given instance, across all probe sets. To conclude, this yielded FC vector was log2-transformed and stored as a column in a new table. Once all rows of the instances info table had been processed, the resulting table was the log2-transformed FC values of 12,307 probe sets across all 6,100 instances of the CMap.

Step Three Step three handled the filtration and translation of the instance-level FC table from step two into various versions under different gene identifiers. This step required the use of the package AnnotationDbi (v. 1.42.1) [31] (in unison with the BrainArray DB package (v. 22) [19] for compatibility with the CDF), which enabled the ability to pull gene identifiers for provided probe set IDs. With this functionality, we were able to use the probe sets contained in the instance FC table to obtain corresponding gene symbols and Entrez IDs. Of the 12,307 passed, only 12,245 were able to retrieve alternate identifiers. The 62 missed probe sets (consisting of 53 control probe sets and 9 dead probe sets) were filtered out, while the remaining 12,245 were used in creating three versions of the filtered instance FC table (each under probe set, gene symbol, or Entrez ID annotation). These resulting table versions were then put forward for user-end querying.

3.2 Querying Procedure and UI

Unlike data processing, the querying procedure (depicted in Figure 3.3) is an active process which depends on supplied user input. Upon startup, the user is prompted to choose which annotated form of the CMap instance FC data to read into memory – a choice

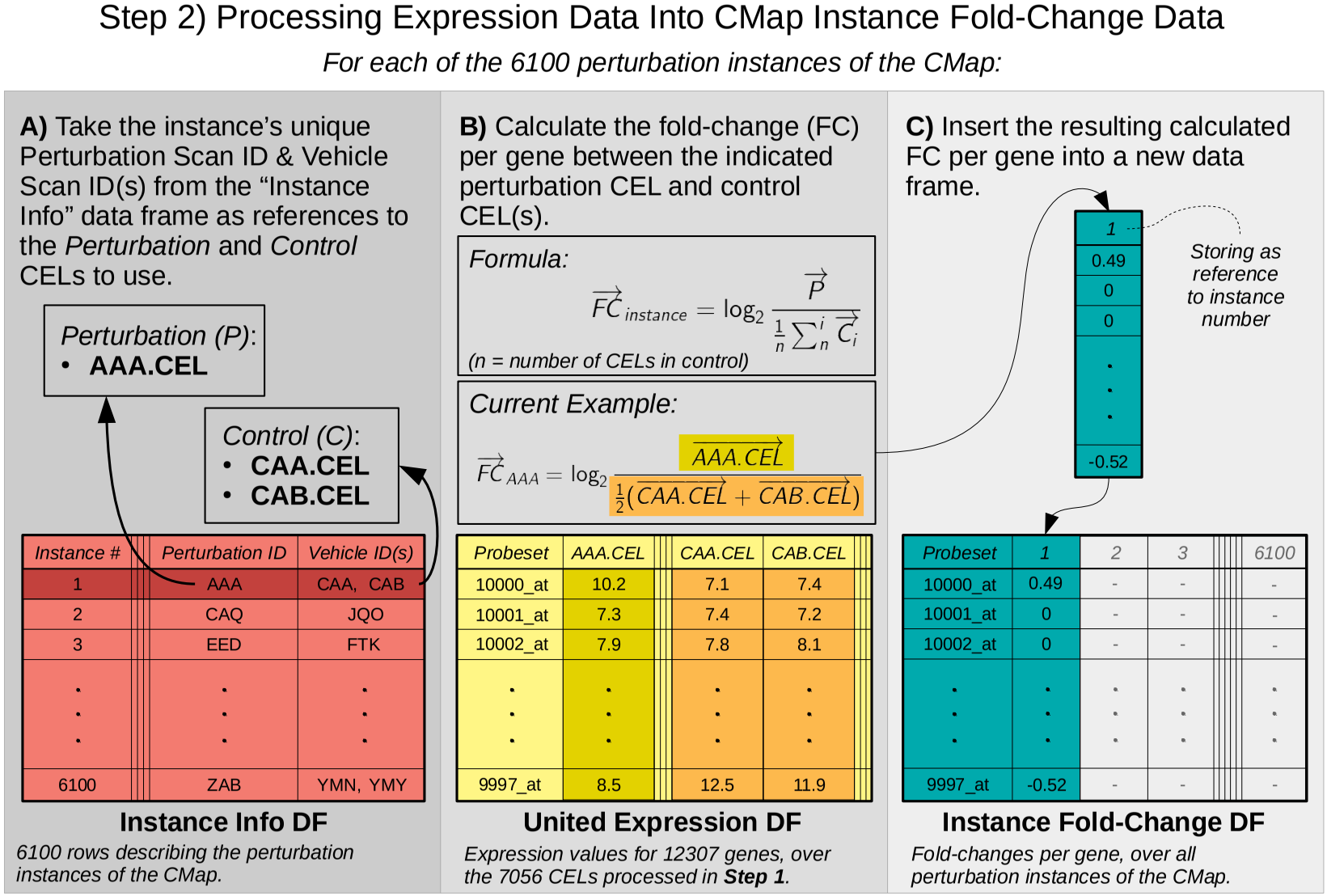


Figure 3.2: Flowchart depicting actions taken in step 2 of data processing – the processing of the united expression data frame from step 1 (Figure 2.1) into instance level FC data.

which depends upon the annotation of their query signature. Once chosen, the user is prompted to provide a query signature in the form of a CSV file – mandatorily containing the headers gene_id and log2_fold_values. Upon providing a query signature, Gene IDs within are checked against (and limited to) those found in the selected annotation form of CMap instance FC data.

Next, correlation testing (either Pearson or Spearman correlation – depending upon the user's choice) is performed between the query signature and each instance of the CMap – yielding a correlation coefficient per instance of the CMap. The resulting correlation coefficients are ultimately used to return a sorted form of the instances info table, listing the instances of the CMap in order from most correlated to the query siganture, to most anticorrelated.

The entire querying procedure was made into an interactive web-application using R's shiny package (v.1.2) [32]. Displayed data tables were made utilizing the package DT (v.0.5) [33] – an interface for the javascript DataTables library. In addition to the necessary controls to carry out the querying procedure, the application contains controls which manipulate the query output – including the ability to reverse the sorting of the query results, the ability to specify which instances info column groups to display, as well as the ability to save the output (in its manipulated form) as a CSV file.

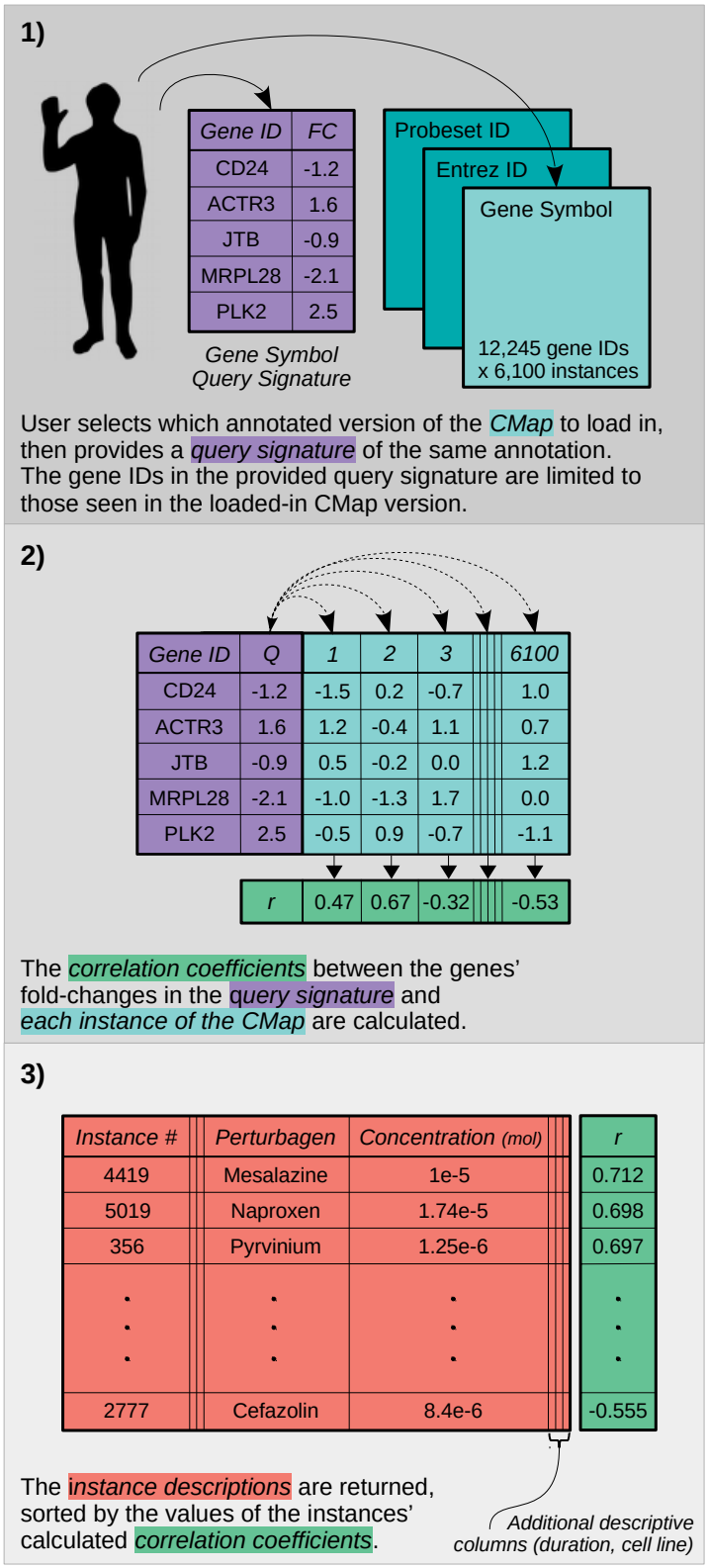


Figure 3.3: Depiction of the decided actions to be taken in CMap querying process, which utilizes the output instance tables from data processing in unison with an FC query signature provided by the user to return a listing of the CMap instances sorted from most correlated to the query signature to most anticorrelated.

3.3 Test Query Signature Gathering/Preparation

In order to functionally test the adapted CMap tool, it was necessary to procure a number of query signatures. To do so, sets of gene expression profiling data were obtained from NCBI's Gene Expression Omnibus (GEO) – a database repository of various experiment-driven gene expression data. The main criteria used when finding microarray-based profiling data in the GEO were:

- Data must originate from an Affymetrix GeneChip with reasonable probeset overlap with the CMap dataset (i.e. Human Genome GeneChips).
- Data must be a set of profiles over conditional/control samples in order to conclude FC of differentially expressed genes (DEGs) of some state.

Under these criteria, three sets of gene expression profile data were selected and downloaded from the GEO:

1. **Human breast tumors:** Expression data from tumorous human breast tissue and their paired healthy tissues. The data stems from a study which aimed to identify transcriptomic changes in breast tissue as carcinoma develops. *(43 diseased profiles / 43 healthy profiles)* [34]
2. **Type 2 diabetic human pancreatic islets:** Expression data from type 2 diabetic (T2D) and non-diabetic isolated human pancreatic islets. The data stems from a study which aimed to validate claims of PI3K's role in insulin secretion. *(7 diseased profiles / 6 healthy profiles)* [35]
3. **F05 treated MCF7 cell lines:** Expression data from MCF7 cell lines treated with *F05* (a novel compound which, under experimental application, promoted regenerative action in damaged CNS neurons) and untreated MCF7 cell lines. Data stems from a study which aimed to use the CMap for identifying regeneration-promoting compounds akin to F05. *(3 treatment profiles / 3 control profiles)* [10]

To create a query gene signature from each expression data set, it was necessary to process each into gene expression values. This mostly followed suit with Step One of Section 3.1; for each data set, all CELs were read into AffyBatches with their corresponding BrainArray CDF, and were subsequently processed into ExpressionSets using qFARMS. However, an additional end-step was performed where each of these ExpressionSets were passed through the FARMS-provided gene filtration method Informative/Non-Informative (INI) calls [36]. INI calls is an expansion of the FARMS algorithm which observes probe behavior underlying features (i.e. genes, probe sets) to obtain feature-wise signal-to-noise ratios, which ultimately allows for identification of significantly noisy (i.e. *non-informative*) features. Application of INI calls upon each ExpressionSet allowed

for filtration of genes deemed non-informative, leaving only informative genes in each ExpressionSet.

From there, queries were formed by collecting each data set's DEGs between their conditional/control groups along with their corresponding FC values. This was achieved via the package Limma (v. 3.36.5) [37]. Limma (Linear Models for Microarray Data) identifies FC and differential expression of genes by way of per-gene linear model fitting followed by empirical Bayes moderated statistical testing (via t-statistics, F-statistic, log-odds) of differential expression. Per ExpressionSet, its use toward query signature obtention was as follows:

1. CELs of the ExpressionSet were labeled via a *design matrix* - a $n \times m$ labeled binary matrix (n corresponding to CELs, m to groups) which states per CEL the group which they belong to.

2. The ExpressionSet object and design matrix were fed into Limma's *lmFit* function – a function which fits per-gene linear models to the gene expression values across all CELs between the two groups, yielding generalized log FC values per gene. Its use resulted in an *MArrayLM* object (held within, the found log FC values per gene).

3. The MArrayLM object resulting from lmFit was then fed into Limma's *eBayes* function – a function which runs the aforementioned statistical testing and test statistics and p-values per gene of their possible differential expression between the declared groups. Its use resulted again in an MArrayLM object (housing the same as previously with the inclusion of these calculated test statistics and p-values).

4. The MArrayLM object resulting from eBayes was then fed into Limma's *topTable* function – a function which reveals gene-wise FC values and p-values held within MArrayLM objects. Additionally, topTable is capable of p-value adjustment (a measure to account for the typically high false discovery rate of microarray studies [38]). This was leveraged, as topTable was called with Benjamini-Hochberg (BH) p-value adjustment. From there, DEGs were identified via their adjusted p-values (p-value ≤ 0.05). The names and FC values of identified DEGs were pulled and written to a CSV under the format specified for the querying procedure – ultimately yielding a query signature.

Chapter 4

Results

4.1 Resulting UI and Workflow

The implementation methodologies described in Sections 3.1 and 3.2 resulted an operational version of the proposed adapted CMap tool (i.e. the Connectivity Map FC-Signature Querying Tool). More specifically, the querying process methodology detailed in 3.2 manifested as the tool's UI, while the data processing methodology in 3.1 yielded three differently-annotated CMap instance FC data sets which are utilized within the tool.

The tool operates as a single window web application consisting of several elements. Its UI is depicted in Figure 4.1, where elements of the UI are highlighted, labeled A-D, and are interconnected by arrows to exhibit the tool's workflow, where:

- Element A prompts the user to select and confirm the annotation type of their query signature. The button "Confirm Selected Annotation" loads the correspondingly-annotated CMap instance FC data set into the environment.
- Element B prompts the user to select and load a query signature. The button "Load File" adds all valid gene entries within a selected file into an environmental query signature, while also displaying the success status of the load (being success, warning, or failure).
- Element C displays the loaded environmental query signature, and prompts the user to start querying the CMap under a certain correlation method. The button "Start Query" initiates correlation testing between the environmental query signature and the each instance within the environmental CMap data set.
- Element D displays the query results, as well as offers options for manipulation of the results (column restriction, result sorting). It also permits the ability to save the results (in their manipulated form) to a CSV file.

Element visibility coincides with progress within the workflow; element A is visible upon tool initialization, while subsequent elements become visible as the main interaction point of their preceding element is triggered.

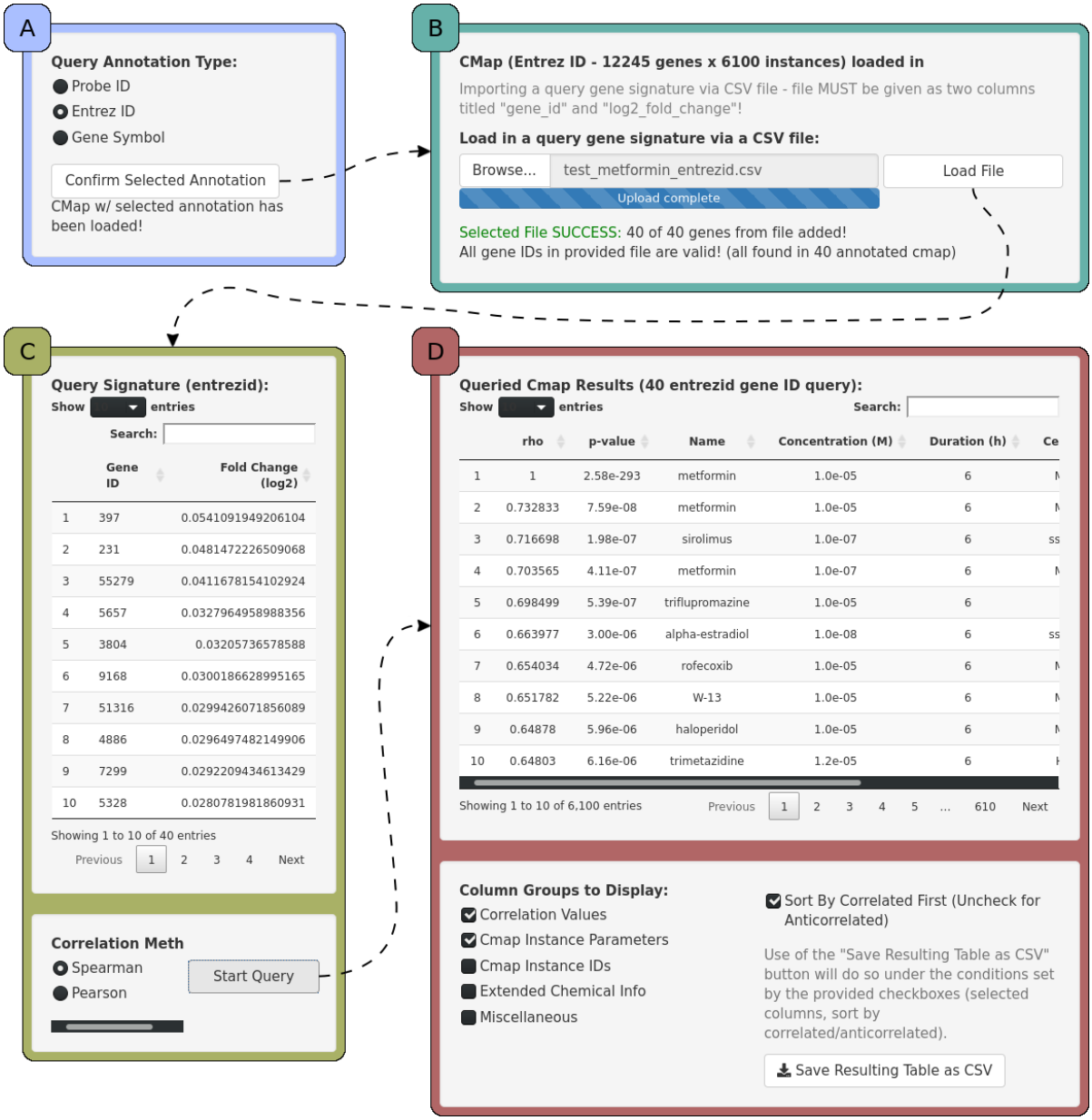


Figure 4.1: The UI of the CMap FC-Signature Querying Tool querying tool, with highlighted/labeled UI elements interconnected by arrows to showcase the workflow.

4.2 Breast Cancer Signature Query Results

Application of the defined query processing methodology to the breast cancer gene expression profile data set resulted in a set of differently-annotated query signatures, each consisting of 199 genes (visualized in Figure 4.3). Step-wise effects of the processing methodology were as follows:

- Whilst processing the data set's 86 CELs (originating from HGU133A GeneChips) into an ExpressionSet object, use of the corresponding BrainArray CDF remapped the CELs's 22,283 probe sets into 12,322 probe sets.

- Application of INI calls to the processed ExpressionSet narrowed down the 12,322 probe sets to 300 informative probe sets.

- Utilization of the Limma pipeline upon these 300 probe sets identified 199 which were differentially expressed between control and conditional.

The gene symbol annotated version of this 199 gene query signature was then successfully loaded into the adapted CMap tool and used to query the CMap (via Spearman correlation). The 5 most anticorrelated/correlated instances from the query results can be seen in Table 4.1.

4.3 T2D Islet Signature Query Results

Application of the defined query processing methodology to the T2D islet gene expression profile data set resulted in a set of differently-annotated query signatures, each consisting of 1,440 genes (visualized in Figure 4.4). Step-wise effects of the processing methodology were as follows:

- Whilst processing the data set's 13 CELs into an ExpressionSet object, use of the corresponding BrainArray CDF remapped the CELs's 22,283 probe sets into 12,322 probe sets.
- Application of INI calls to the processed ExpressionSet narrowed down the 12,322 probe sets to 3,349 informative probe sets.
- Utilization of the Limma pipeline upon these 3,349 probe sets identified 1,440 which were differentially expressed between control and conditional.

The gene symbol annotated version of this 1,440 gene query signature was then successfully loaded into the adapted CMap tool and used to query the CMap (via Spearman correlation). The 5 most anticorrelated/correlated instances from the query results can be seen in Table 4.2.

4.4 F05 Treatment Signature Query Results

Application of the defined query processing methodology to the F05 treatment gene expression profile data set resulted in a set of differently-annotated query signatures, each consisting of 5 genes (visualized in Figure 4.5). Step-wise effects of the processing methodology were as follows:

- Whilst processing the data set's 6 CELs into an ExpressionSet object, use of the corresponding BrainArray CDF remapped the CELs's 54,675 probe sets into 20,481 probe sets.
- Application of INI calls to the processed ExpressionSet narrowed down the 20,481 probe sets to 1,552 informative probe sets.

- Utilization of the Limma pipeline upon these 1,552 probe sets identified 5 which were differentially expressed between control and conditional.

Both due to the low number of DE genes and the suspicious volcano plot (seen in Figure 4.2), the resulting signature required investigation – which revealed that the resulting oddity was due to the applied BH p-value adjustment. Therefore, the query signature was made using the unadjusted p-value – which identified 47 DE genes, rather than 5. The gene symbol annotated version of this 47 gene query signature was then successfully loaded into the adapted CMap tool and used to query the CMap (via Spearman correlation). The 5 most anticorrelated/correlated instances from the query results can be seen in Table 4.3.

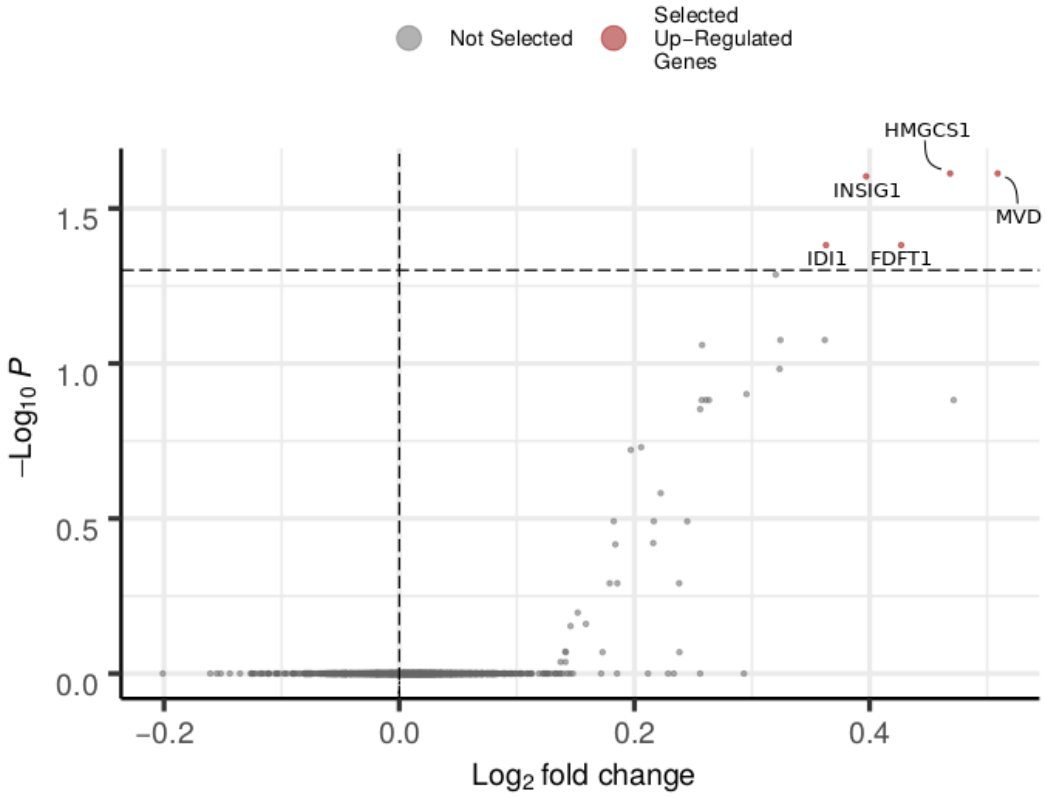


Figure 4.2: Volcano plot of the rejected F05 treatment query signature.

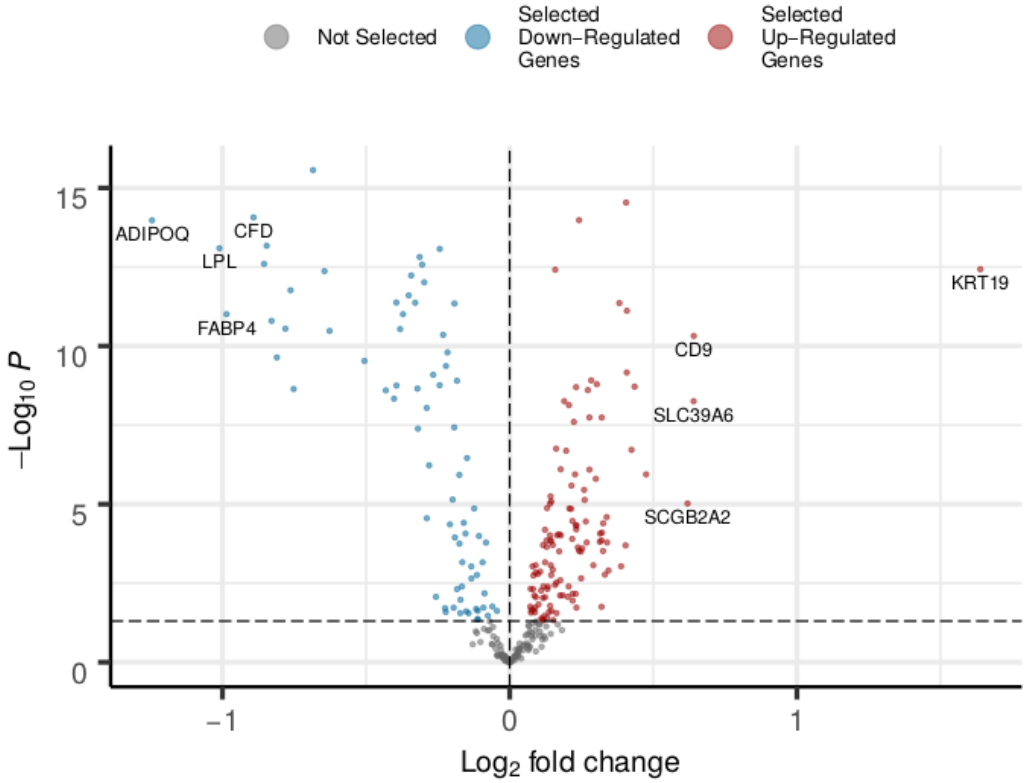


Figure 4.3: Volcano plot showing the $\log_2 FC$ and -$\log_{10} p\text{-}value$ of the 300 INI-deemed informative genes in the breast cancer data set. Selected genes for the signature are shown as blue/red, while non-selected genes are show as grey – these failing to meet the 0.05 significance level represented by the horizontal dotted line.

	ρ	p-value	Name	Concentration(M)	Duration(h)
1	-0.380065	3.08E-08	fluorometholone	1.1E-05	6
2	-0.354858	2.71E-07	chlorpropamide	0.0001	6
3	-0.341876	7.73E-07	mesalazine	0.0001	6
4	-0.328051	2.24E-06	metformin	1E-05	6
5	-0.327573	2.33E-06	wortmannin	1E-08	6
. . .	. . .	. . .	. . .	. . .	. . .
6096	0.280975	5.82E-05	quipazine	9E-06	6
6097	0.291524	2.95E-05	H-89	5E-07	6
6098	0.29263	2.74E-05	isoxsuprine	1.2E-05	6
6099	0.299922	1.68E-05	piroxicam	1.2E-05	6
6100	0.365159	1.14E-07	tetraethylenepentamine	1E-05	6

Table 4.1: The resulting five most anticorrelated/correlated instances from querying within the adapted CMap tool using the formed breast cancer query signature detailed in Figure 4.3.

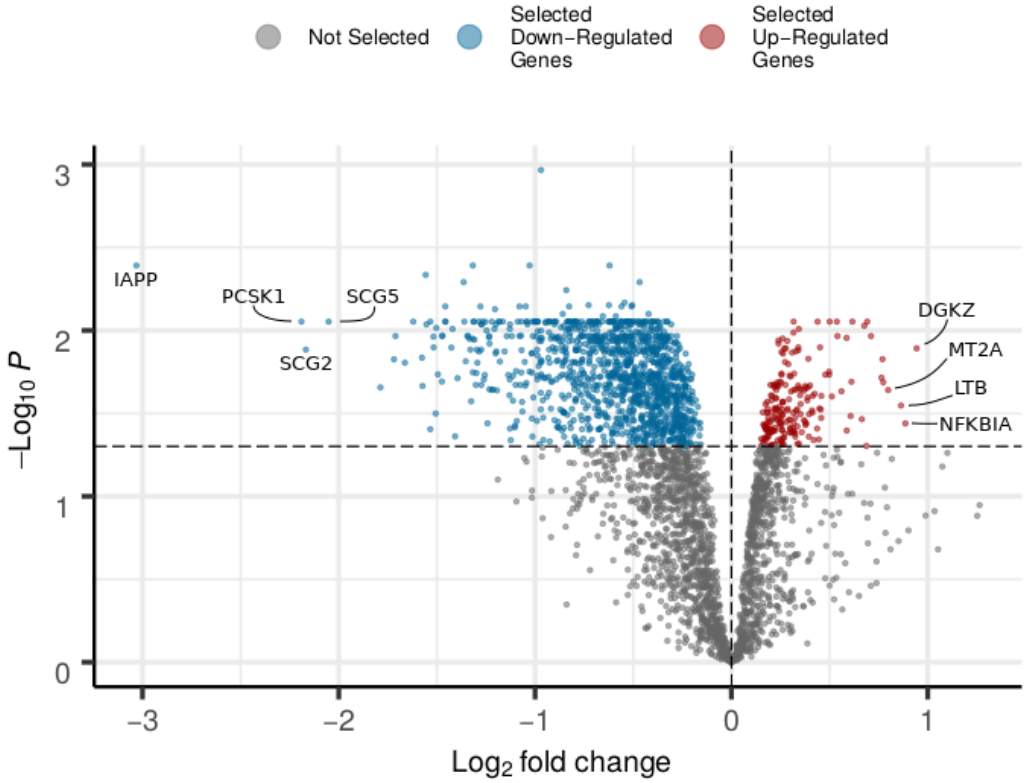


Figure 4.4: Volcano plot showing the $\log_2 FC$ and -$\log_{10} p\text{-}value$ of the 3,349 INI-deemed informative genes in the T2D pancreatic islet data set. Selected genes for the signature are shown as blue/red, while non-selected genes are show as grey – these failing to meet the 0.05 significance level represented by the horizontal dotted line.

	ρ	p-value	Name	Concentration(M)	Duration(h)
1	-0.2880521	6.5E-29	chlorpropamide	0.0001	6
2	-0.2852026	2.36E-28	genistein	1E-05	6
3	-0.2782384	5.18E-27	pirinixic acid	0.0001	6
4	-0.2736216	3.82E-26	iloprost	1E-06	6
5	-0.2709525	1.19E-25	Gly-His-Lys	1E-06	6
. . .	. . .	. . .	. . .	. . .	. . .
6096	0.2879394	6.84E-29	cimetidine	1.6E-05	6
6097	0.2979831	6.46E-31	mesalazine	0.0001	6
6098	0.2980587	6.23E-31	benperidol	1E-05	6
6099	0.3072941	7.24E-33	16-phenyltetranorprostaglandin E2	1E-05	6
6100	0.3095002	2.44E-33	CP-863187	1E-05	6

Table 4.2: The resulting five most anticorrelated/correlated instances from querying within the adapted CMap tool using the formed T2D islet query signature detailed in Figure 4.4.

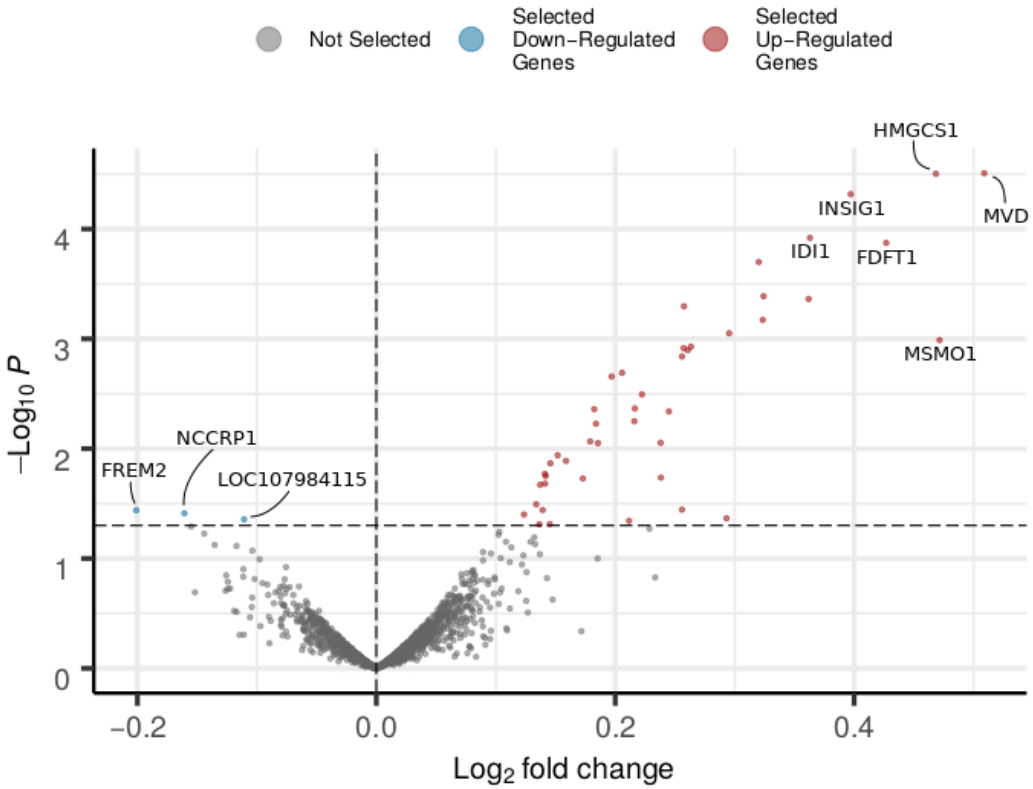


Figure 4.5: Volcano plot showing the $\log_2 FC$ and -$\log_{10} p\text{-}value$ of the 47 INI-deemed informative genes in the F05 treatment data set. Selected genes for the signature are shown as red, while non-selected genes are show as grey – these failing to meet the 0.05 significance level represented by the horizontal dotted line.

	ρ	p-value	Name	Concentration(M)	Duration(h)
1	-0.630972	1.66E-05	kinetin	1.9E-05	6
2	-0.610729	3.64E-05	kinetin	1.9E-05	6
3	-0.567206	0.000166	lansoprazole	1.1E-05	6
4	-0.562348	0.000195	nordihydroguaiaretic acid	1E-06	6
5	-0.555263	0.000243	papaverine	1.1E-05	6
. . .	. . .	. . .	. . .	. . .	. . .
6096	0.627733	1.89E-05	prenylamine	9.6E-06	6
6097	0.63664	1.32E-05	thioproperazine	6.2E-06	6
6098	0.648785	7.93E-06	ciclosporin	3.4E-06	6
6099	0.654049	6.31E-06	thapsigargin	1E-07	6
6100	0.655061	6.04E-06	terconazole	7.6E-06	6

Table 4.3: The resulting five most anticorrelated/correlated instances from querying within the adapted CMap tool using the formed F05 treatment query signature detailed in Figure 4.5.

Chapter 5

Discussion

5.1 Accessibility Adaptations

With regards to the accessibility enhancements brought forward in the adapted CMap tool, the implementation has proved to be successful. While the effects of enhanced accessibility (via the redesigned UI/workflow and query signature expansion) were presumably beneficial, it remained necessary to validate their efficacy. In attempting to assess the efficacy of these adaptations, both the original build02 and adapted CMap tool were demonstrated to several people of varying backgrounds – from bioinformatics, biology, computer science, to even those without relevant background. The feedback from these demonstrations is as follows:

- **Workflow:** In comparison to the multi-page workflow of build02's implementation, the seamless single-page workflow of the adapted CMap tool was unanimously preferred across all participants.

- **Functionality:** Participants generally responded favorably to the core querying functionality provided in the adapted tool. However, it was noted by some participants that additional functionalities from the build02 implementation was missing – these being in-line result options (e.g. perturbagen ChemBank search, perturbagen highlighting, barview search, per-probe set connectivity graphs) and the alternate results mode titled *permuted results* (these being aggregated per-perturbagen results, rather than per-instance results). As the goal of this project was to provide enhanced accessibility to the CMap tool's core functionality, implementation of these additional functions were not deemed necessary. However, it would certainly be advantageous to incorporate these functionalities into an updated version of the adapted tool – respectively as per-instance buttons allowing different actions to be taken, and as a button which toggles the results panel between instance and permuted results (with permuted results coming from some to-be-determined method of calculation upon toggling).

- **Instructions:** Despite trying to make clear yet concise instructions, some participants thought that the instructions provided in the adapted CMap tool were insufficient when compared to build02's hefty help document. Unfortunately, this is a product of circumstance; treading the line between making an approachable, lightweight UI while also providing comprehensive instructions proves to be difficult. However, a viable solution to this would be to provide all instructions as hover-enabled tooltips rather than ever-present lines of text. In this setup, the look of the UI would be even less cluttered, and the dedicated tooltip boxes would be able to house even more text information. This would definitely be a goal in a future version of the adapted CMap tool, as it would enhance its accessibility even further.

- **Expanded Gene IDs:** All bioinformatician and biologist participants responded very positively to the inclusion of gene identifiers beyond probe set IDs. This positive response makes clear sense, as gene profiling data may originate from sources outside of Affymetrix GeneChips (e.g. other chip manufacturers, in-house chips, non-chip methods such as RNASeq).

Overall, the insight gleaned from the participants' feedback shows that the accessibility changes applied were effective – albeit with some room for improvement. Given this, the changes would be nice additions to a future official build of the CMap – with the UI/workflow as the main UI for querying the CMap, or even as an alternate "simplified" UI.

5.2 Querying Performance

While the applied accessibility adaptations have clearly proven to be successful, it is difficult to say the same for adaptations which affect the querying performance itself (i.e. adapted microarray processing measures and querying algorithm). To elaborate, it is necessary to examine each query signature detailed in Section 3.3 and their subsequent query results.

Breast Cancer Query Signature Upon brief inspection, the query signature yielded from the breast cancer data set seems OK; the genes which make up the extreme ends of up/down regulated genes in the query signature (marked in Figure 4.3) are known for their directional regulation in breast cancer.

The results of querying the CMap with the signature (seen in Table 4.1) provide a potentially positive view of the tool's performance. Looking at rows 1-5 – the top most anticorrelated (i.e. potentially therapeutic) hits – there are some noteworthy observations:

- *Flurometholone* (row 1) is a drug used as a treatment for inflammatory eye diseases. Past studies have shown that an esterified verion of this compound (Fluorometholone

acetate) is effective against breast cancer [39] – however isn't used due to negative side effects.

- *Mesalazine* (row 3) is a drug used to treat inflammatory bowel disease (IBD). While not found to be chemotherapeutic, its use has been identified to be chemopreventative in colorectal cancer. [40]

- *Metformin* (row 4) is a well-established T2D medication. Despite its typical use, it has gotten some recent traction as a repositional treatment for breast cancer [41, 42].

- *Wortmannin* (row 5) is fungal steroid metabolite with use in biological research, but has no official clinical use. It is a known PI3K inhibitor, and is thereby researched as a potential chemotherapeutic treatment. There have been multiple studies stating its potential efficacy as a breast cancer treatment [43, 44].

The remaining, unmentioned compound among these top five – *chlorpropamide* (row 2) – is a T2D drug which has no prior studies suggesting its therapeuticity in breast cancer. Without investigating too deeply, it may be worth to investigate it for potential therapeuticity – as it is alongside another T2D drug in the top anticorrelated hits (Metformin) and also that its use counteracts high blood sugar (which has been found to be a contributive factor in breast cancer [45]). However, without corroborating the performance of the tool, investigation of chlorpropamide can not be confidently suggested.

One potential point of criticism to be brought up is the absence of any traditional anticancer medications within the top anticorrelated hits. Within the CMap build02 data exist many instances containing several different anticancer drugs. The highest ranked hit among these is *irinotecan* (rank 8, ρ: -0.29), a chemotherapy drug typically used to treat colorectal cancer. With regards to typical breast cancer treatments, *doxorubicin* (rank 773, ρ: -0.12) is the highest ranked. The reason behind Doxorubicin's lesser correlative strength can potentially come from the signature – which likely contains genes irrelevant to these chemotherapeutic drugs' pathways, ultimately weakening their correlative strength. However, the possibility exists that the fault lies elsewhere (e.g. within the applied adaptations, querying algorithm, or query signature preparation).

Also, it must be noted that the correlative values of the results are quite weak (failing to break +/- 0.4). This is likely due to the number of genes in the query signature. As more genes are included in a query signature, the likelihood of finding genes which do not have correlating activity rises – thereby lowering correlation values. FC thresholding (the limiting of genes in the signature which fail to exceed a certain FC value) was attempted. This resulted in a smaller query signature which (when used for querying) yielded results with higher correlation values. However, the perturbagens within the top hits of these results held no relevance toward the query's condition. Therefore, use of the non-thresholded query was kept. Ultimately, this is not necessarily negative – as within

Broad's CMap, connectivity scores are normalized on a [-1,+1] range. Therefore, perhaps some means of addressing this normalization would be an advisable adaptation here.

T2D Islet Query Signature Upon brief inspection, the query signature yielded from the T2D islet data set also seems OK; the genes which make up the extreme ends of up/down regulated genes in the query signature (marked in Figure 4.4) are known for their directional regulation in T2D pancreatic islets.

The results of querying the CMap with the signature (seen in Table 4.2) again provide a potentially positive view of the tool's performance. Starting with rows 1-5 – the top most anticorrelated (i.e. potentially therapeutic) hits – some notable occurrences are:

- *Chlorpropamide* (row 1) is drug used to treat T2D – a matching, clinically used treatment to the condition represented by the query signature.
- *Genistein* (row 2) is a soy isoflavone compound with many researched uses, but no official clinical use. Among these reseached uses is as a treatment for T2D [46, 47]
- *Iloprost* (row 4) is a drug used to treat hypertension/vasoconstriction. Research has shown that the drug's use helps in treating peripheral vascular disease (a notorious complication of T2D) – which may have to do moreso with the drug's known therapeutic effect than its relation to the gene signature, but is interesting nonetheless.

The remaining two compounds from the top anticorrelated hits (*prinixic acid* and *Gly-His-Lys*) don't have any prior studies suggesting its therapeuticity in T2D. Furthermore, surface-level investigation does not lend any logical connection to these drugs as potential therapies. Regardless, it can not be definitely be said if these are potentially correct suggestions of T2D therapies or false hits due to poor performance of the tool without in-depth analysis.

While the match of chlorpropamide is certainly a positive look for the tool's performance, the absence of other T2D medications could possibly raise some questions. For example, the T2D medication Metformin (mentioned above in the breast cancer query results) does not correlate well – with its most anticorrelated hit reaching rank 1117 ($\rho : -0.08$). Like with the absence of doxorubicin from the breast cancer data set's results, this also could be the fault of genes in the signature or even incorrect tool performance.

As was with the breast cancer query signature, correlative values of the results are noticeably weak (maxing out around +/- 0.3). FC thresholding was tried as well here. Unlike before, thresholding did not yield higher correlative values. Again – this may potentially not be bad, as Broad's CMap normalizes their connectivity scores on a [-1,+1] range.

F05 Treatment Query Signature As previously mentioned in Section 4.4, use of the query preparation pipeline from Section 3.3 upon the F05 treatment data set yielded

a suspicious query signature consisting of 5 genes. Use of this signature for querying would have been senseless, as the small sample size of genes present in the signature would provide uninformative results. However, as the data set holds relevancy here (due to its origin in a CMap-driven experiment), it was attempted to salvage it by identifying the source of its questionability. The source identified was the use of BH p-value adjustment. This raises concern over the validity of the findings from within the data set's associated paper – wherein the authors propose piperazine phenothiazine compounds as repositionable treatments for injured CNS neural cells, yet do not disclose the exact computational methodology behind procuring their 21 gene query signature. However, this suspicious query signature may also be a product of the other steps taken in creating the query signature (e.g. FARMS, custom CDF). Nevertheless, it was decided to proceed with the 47 gene signature made without BH p-value adjustment.

Identifying the validity of the genes making up the extreme ends of the up/down regulated genes of the query signature (marked in Figure 4.5) was less clear than the previous two data sets; Finding resources stating these genes' roles in CNS neural cell damage/regeneration proved to be difficult. Regardless of this, the query was still used to query the CMap.

The results of querying the CMap with the signature (seen in Table 4.3) provide a mixed view of tools performance. Of the top most anticorrelated hits, only the antispasmodic drug Papaverine (row 5) has literature suggesting potential therapeuticity [48]. However, much more notable is that the top two anticorrelated hits are both instances of kinetin (a plant hormone which stimulates cell division and plant growth). While there is no literature stating kinetin's therapeuticity in neural cell regeneration/repair, this would be grounds for investigation of kinetin's potential repositioning if the correctness of both the tool's performance and the query signature were corroborated. However, given the uncertainty behind both, this may just be a coincidental (albeit, strong) finding.

Things get convoluted when looking at rows 6096-6100 – the top most correlated (i.e. potentially condition-causing) hits. On one hand, the 2nd most correlated instance is of *thapsigargin* (row 6099) – a plant-originating inhibitor of endoplasmic reticulum Ca^{+2} ATPase. It has research stating its promotion of neural cell oxidative stress and apoptosis [49], making its appearance here fitting. However, the 3rd most correlated instance is of the immunosuppresant drug *ciclosporin* (row 6098). It has relevant research suggesting the opposite – that it positively affects proliferation of neural cells [50]. Like with Doxorubicin and Metformin in the previous data sets, these conflicting findings could be the result of genes in the gene signature (i.e. that relevant genes in ciclosporin's therapeutic pathway are absent). It is also possible that these results are incorrect due to the now-questionable data set. However, as with all preceeding findings, these could potentially be attributed to faulty tool performance.

Most interesting of all though is that *thioproperazine* (row 6097) is present in the most correlated hits. Thioproperazine is a piperazine phenothiazine – the compound

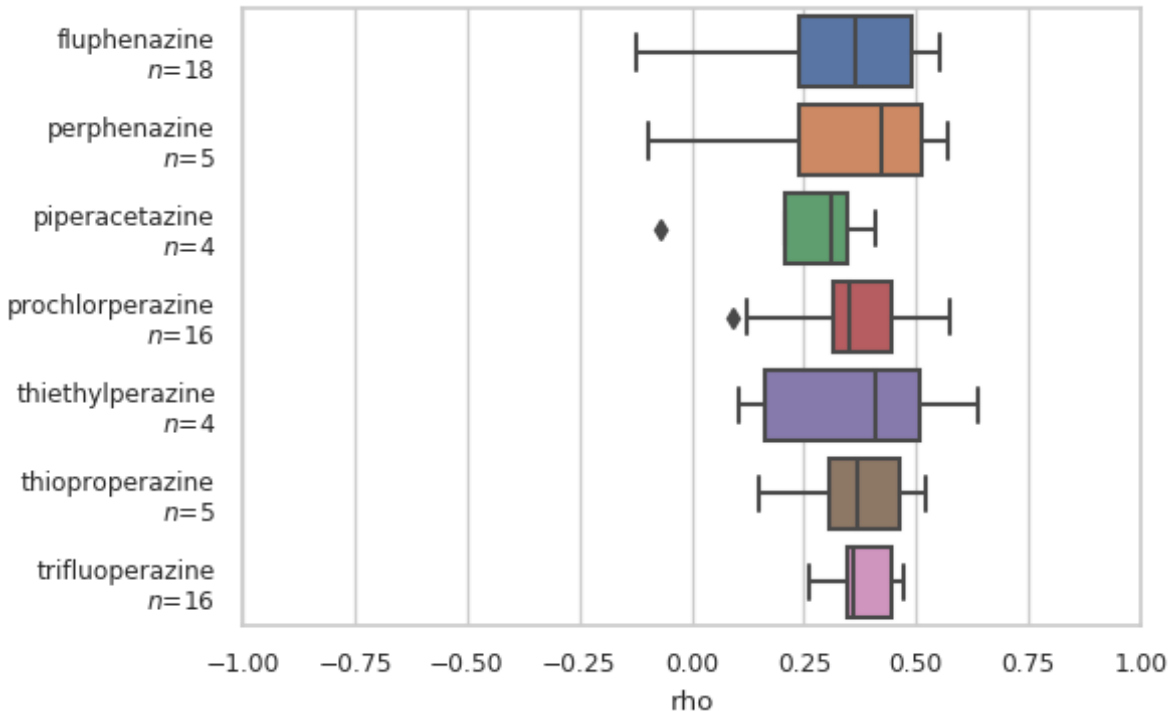


Figure 5.1: Distributions of rho values across all instances of the seven Piperazine Phenothiazine compounds in the CMap.

family which the data set's origin paper suggested to be therapeutic. Its presence here would suggest exactly the opposite of the origin paper's thesis statement. In fact, when looking at the ρ distributions for all seven piperazine phenothiazine compounds in the query results (visualized in Figure 5.1), they are primarily positively correlated. If there were no reservations over the data set, query signature, or tool performance – this would be a strong indication that the origin paper's findings were false. Again however, as all parts garner uncertainty, this can not definitively be said.

As an additional remark – the task of creating Figure 5.1 reinforced the potential positivity of some requested accessibility features by the UI demonstration participants. Identification of piperazine phenothiazine compounds in the query results was a manual process here, which would have been greatly simplified by the requested perturbagen ChemBank search (or even in-line, toggleable perturbagen information). Moreover – aggregation of each perturbagen's correlative performance over their results instances here was also a manual process, which would have been instantly available with the requested permuted results option. Furthermore, the inclusion of extended result search tools (e.g. limiting results view to chemical family) or even plotting of results (similar to that of 5.1) would further enhance accessibility in presenting findings. Unlike the UI changes to increase accessibility to the tool's intermediate use, all of these changes would increase accessibility for the user to locate and present their findings in the data – which ultimately would streamline the user's ability to achieve success with the tool's main goal of finding repositionable drugs.

After looking at the three data sets' yielded query results, two things are clear: that

further accessibility changes could help result exploration and presentation; and that querying is functional (albeit with some ambiguous level of efficacy). A few of the top returned perturbagens show promise behind the querying's effective performance. However, numerous questionable results keep this observation of efficacy from being definite; with occurances of expected perturbagens not ranking as expected (e.g. doxorubicin in breast cancer, metformin in T2D, piperazine phenothiazines in F05) and uncorroborated perturbagens occuring as top hits (e.g. chlorpropamide in breast cancer, prinixic acid and Gly-His-Lis in T2D, kinetin and others in F05) its hard to say whether these are true findings or the result of incorrect tool performance / gene signature curation without more focused evaluation. Ultimately though, evaluating the querying performance or pinpointing potential problems has lesser relevance here – as the focus of this work was to deliver further accessibility to the CMap as a tool (which has been achieved).

Chapter 6

Conclusion

Within this work, we explored the idea of including adaptations into the Broad Institute's Connectivity Map which could have positively impacted it. The main focus of these adaptations was to enhance the tool's accessibility, while others were included which could have potentially bolstered querying performance. The sum of these adaptations resulted in the implementation of an adapted CMap tool.

Focus testing of the adapted CMap tool on several participants yielded great favor for the accessibility adaptations. The adapted UI's seamless single-page workflow and stepwise-presented instructions were seen by participants as far more accessible than Broad's implementation. The expansion of queryable gene identifiers was also greatly appreciated by all participants with biology background – which makes sense, as there is no guarantee that the user's query signature is composed of Affymetrix probe set IDs. The only complaint brought forward with the adapted UI was the instructions included – which were deemed by some as too brief. This was the result of trying strike a balance between providing verbosity and an uncluttered UI. To this end, hover-enabled tooltips have been established as a remedy – which would allow for more verbose instructions without cluttering the UI. Overall, this feedback indicates that the accessibility adaptations implemented were a success – albeit, in need of some reworking. Despite this, it can be concluded that these adaptations are valuable and should be included in any future CMap tool.

Aside from the acceptance of the accessibility enhancements brought forward, additional potential accessibility-enhancing features came to light as well. It was suggested by some participants that implementing preexisting features from Broad's CMap (e.g. permuted results, in-line perturbagen details) would bring more accessibility to the already accessible adapted UI. The helpfulness of these features were seen later when analyzing results of queries made to the adapted CMap – where both permuted results and in-line perturbagen details would have greatly streamlined the result-parsing process. Furthermore, it was noticed that some result plotting functionality (either global or selective) would have been greatly beneficial. If included, these features would enhance the tool's accessibility with regards to result exploration and presentation of findings –

ultimately fast-tracking the user's ability to achieve success with the tool's main goal of finding repositionable drugs. Overall, it can be seen that features like these are quite valuable in the "end-game" of the tool's use. Therefore, it can be concluded that they too be included in any future CMap tool.

Efficacy of the remaining performance-affecting adaptations (i.e FARMS, custom CDFs, the new querying algorithm) was evaluated by querying the adapted CMap with three test query gene signatures made from real-world gene expression profiling data sets. The query results of these three showed some promise; of the three queries, there was one occurance of a well-known therapy to the relevant condition being the top most anticorrelated (i.e. potentially therapeutic) hit. Furthermore, many perturbagens among the top most anticorrelated hits had prior research exploring their therapeuticity with respect to their subject query signature's condition. While these act as positive indications of the querying performance, there were also occurances which raised uncertainty – these being occurances of expected hits not ranking highly for their relative query condition, perturbagens falling within the top hits which have no prior research of their relevant therapeuticity, or even opposing findings to the base research from which the query originated (as was the case with the F05 query results). The true nature of these perceivably negative occurances is not trivial to pin down. It is possible that these occurances are valid; anticipated perturbagens not ranking highly could be due to the query signature not aligning with the genes underlying their pathways, uncorroborated top-hitting perturbagens could potentially provide a degree of therapeuticity, and the opposing findings to the F05 data's origin paper could be true. However, it also remains possible that these occurances are due to malperformance of the tool via factors underlying the querying – whether it be one or either of the adaptations affecting the CMap data processing (e.g. FARMS, brainarray CDF), the adapted querying algorithm, a fault in the assembly of test query signatures, or even the data underlying the test query signatures. Ultimately, identifying whether these findings are wrong or right would require further in-depth analysis – and if wrong, deeper analysis would be required to identify which of the many factors contributed to the tool's malperformance. Therefore, despite indications that some degree of successful querying was provided, it can not be confidently concluded whether these adaptation were effective – and consequently, they can not currently be suggested for inclusion in a future CMap tool.

Lastly, it must be mentioned that there has since been a new official release of the Connectivity Map [51]. This third iteration of the CMap most notably brought changes to the data set – focally in both the profiling platform used to assay perturbation instances and the breadth of perturbation instances covered. Rather than the commercially-available Affymetrix GeneChips used in build02, a gene expression profiling platform was devised in-house titled the *L1000*. In principle, the L1000 assays a reduced representation of the human transcriptome – focusing on 1,000 "landmark" transcripts (i.e. an optimal set of transcripts which balances overall informativeness and cost efficiency) rather than the

entirety of the human transcriptome as offered by Affymetrix's HG series of GeneChips. This cost-effective yet performant assaying platform allowed for a substantially larger number of perturbation instances to be covered – with 1.3 million instances having had been profiled until now, covering 42,080 perturbagens. This shift away from GeneChips also eliminated querying via probe set ID – which has been replaced with the ability to query the CMap using a variety of gene identifiers. Furthermore – while not a focal change, the third iteration features an updated UI. The new UI is considerably more modern than build02's – offering pages with interactive dynamic elements (e.g. options for querying, data viewing) intuitively grouped and displayed by their purpose.

While all of these changes included in the third CMap are a large step forward, their inclusion either impacts or overlaps to a degree with adaptations brought forward in the adapted CMap tool. Firstly – FARMS and custom CDFs are no longer applicable to the data set, as both are geared toward use with Affymetrix GeneChip derived profiling data. This nullifies any potential suggestion of their inclusion in the CMap (unless adapted for use with the L1000). Furthermore – the new tool has expanded the queryable gene identifies, which was an idea brought forward in the adapted tool. While its appearance in the third CMap confirms the validity of the idea, it means the suggestion of its inclusion via the adapted tool is too late. Lastly – the new tool takes steps to make the UI more accessible via its presentation. While the changes made are a step in the direction of those brought forward in the adapted UI (thereby validating efforts to enhance the accessibility of the UI), they still do not provide any substantial assistance in simplifying or guiding the tool's workflow. More specifically, the third CMap still uses lengthy instructions (which is now broken into selectable sections) and a single-page UI which exposes all possible interactive elements at the user immediately without any on-page instruction. As both the guiding of workflow in the adapted tool (via shifting user focus in workflow elements) and the attempted embedding of stepwise instructions were seen as positive approaches to the UI, these still stand as applicable adaptations to include in the CMap.

In all – despite updates brought forward in third iteration of the CMap which either invalidate proposed adaptations (e.g. FARMS, custom CDFs) or already implement them (e.g. expaded queryable gene identifiers) – there are still valid, applicable adaptations proposed here which could provide further accessibility in the tool's overall use (via the adapted UI/workflow) or in the finding of results in queried data (via the proposed result-plotting functionality). Overall, their inclusion would be greatly benefit the accessibility of the Connectivity Map.

Bibliography

[1] I. Goldstein, A. L. Burnett, R. C. Rosen, P. W. Park, and V. J. Stecher, "The Serendipitous Story of Sildenafil: An Unexpected Oral Therapy for Erectile Dysfunction," *Sexual Medicine Review*, vol. 7, no. 1, pp. 115–128, 2019.

[2] E. Hargrave-Thomas, B. Yu, and J. Reynisson, "World Journal of Clinical Oncology," vol. 3, no. 1, pp. 1–6, 2012.

[3] T. A. Ban, "The role of serendipity in drug discovery," *Dialogues in clinical neuroscience*, vol. 8, no. 3, pp. 335–344, 2006.

[4] Affymetrix, "The Structure, Function, and Applications of GeneChip® Microarrays."

[5] A. A. Ewis, Z. Zhelev, R. Bakalova, S. Fukuoka, Y. Shinohara, M. Ishikawa, and Y. Baba, "A history of microarrays in biomedicine," *Expert Review of Molecular Diagnostics*, vol. 5, no. 3, pp. 315–328, 2005.

[6] S. Zhao, W. P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, "Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells," *PLoS ONE*, vol. 9, no. 1, 2014.

[7] J. Lamb, "The Connectivity Map: A new tool for biomedical research," *Nature Reviews Cancer*, vol. 7, pp. 54–60, jan 2007.

[8] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, "The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[9] A. P. Chiang, J. T. Dudley, M. Shenoy, S. Roedder, A. J. Butte, A. A. Morgan, R. K. Pai, M. M. Sarwal, P. J. Pasricha, and M. Sirota, "Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease," *Science Translational Medicine*, vol. 3, no. 96, pp. 96ra76–96ra76, 2011.

[10] A. L. Johnstone, G. W. Reierson, R. P. Smith, J. L. Goldberg, V. P. Lemmon, and J. L. Bixby, "A chemical genetic approach identifies piperazine antipsychotics as promoters of CNS neurite growth on inhibitory substrates," *Molecular and Cellular Neuroscience*, vol. 50, no. 2, pp. 125–135, 2012.

[11] P. Nygren, M. Fryknäs, B. Ågerup, and R. Larsson, "Repositioning of the anthelmintic drug mebendazole for the treatment for colon cancer," *Journal of Cancer Research and Clinical Oncology*, vol. 139, no. 12, pp. 2133–2140, 2013.

[12] Z. Wen, Z. Wang, S. Wang, R. Ravula, L. Yang, J. Xu, C. Wang, Z. Zuo, M. S. Chow, L. Shi, and Y. Huang, "Discovery of molecular mechanisms of traditional Chinese medicinal formula Si-Wu-Tang using gene expression Microarray and Connectivity Map," *PLoS ONE*, vol. 6, no. 3, 2011.

[13] A. M. Brum, J. van de Peppel, C. S. van der Leije, M. Schreuders-Koedam, M. Eijken, B. C. J. van der Eerden, and J. P. T. M. van Leeuwen, "Connectivity Map-based discovery of parbendazole reveals targetable human osteogenic pathway," *Proceedings of the National Academy of Sciences*, vol. 112, no. 41, pp. 12711–12716, 2015.

[14] H. Huang, T. Nguyen, S. Ibrahim, S. Shantharam, Z. Yue, and J. Y. Chen, "DMAP: A connectivity map database to enable identification of novel drug repositioning candidates," *BMC Bioinformatics*, vol. 16, no. 13, p. S4, 2015.

[15] A. S. Brown, S. W. Kong, I. S. Kohane, and C. J. Patel, "ksRepo: A generalized platform for computational drug repositioning," *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–5, 2016.

[16] A. Subramanian, A. Subramanian, P. Tamayo, P. Tamayo, V. K. Mootha, V. K. Mootha, S. Mukherjee, S. Mukherjee, B. L. Ebert, B. L. Ebert, M. a. Gillette, M. a. Gillette, A. Paulovich, A. Paulovich, S. L. Pomeroy, S. L. Pomeroy, T. R. Golub, T. R. Golub, E. S. Lander, E. S. Lander, J. P. Mesirov, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–50, 2005.

[17] M. Hollander and D. A. Wolfe, "Kolmogorov-Smirnov statistic," in *Nonparametric Statistical Methods*, pp. 178 – 185, John Wiley & Sons, 1999.

[18] E. S. Lander, L. M. Linton, and B. Birren, "Initial sequencing and analysis of the human genome [published correction appears in Nature. 2001; 411 (6838): 720]," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[19] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng, "Evolving gene/transcript

definitions significantly alter the interpretation of GeneChip data," *Nucleic Acids Research*, vol. 33, no. 20, pp. 1–9, 2005.

[20] X. Lu and X. Zhang, "The effect of GeneChip gene definitions on the microarray study of cancers," *BioEssays*, vol. 28, pp. 739–746, jul 2006.

[21] R. Sandberg and O. Larsson, "Improved precision and accuracy for microarrays using updated probe set definitions," *BMC Bioinformatics*, vol. 8, pp. 1–8, 2007.

[22] S. Hochreiter, D. A. Clevert, and K. Obermayer, "A new summarization method for affymetrix probe level data," *Bioinformatics*, vol. 22, no. 8, pp. 943–949, 2006.

[23] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[24] H. Gohlmann and W. Talloen, *Gene Expression Studies Using Affymetrix Microarrays*. Chapman & Hall/CRC, 1st ed., 2017.

[25] J. Cheng, L. Yang, V. Kumar, and P. Agarwal, "Systematic evaluation of connectivity map for disease indications," *Genome Medicine*, vol. 6, no. 12, pp. 1–8, 2014.

[26] F. Lefebvre, *Comparaison des méthodes d'analyse de l'expression différentielle basée sur la dépendance des niveaux d'expression*. Master's, Université de Montréal, 2011.

[27] J. Quackenbush, "Microarray data normalization and transformation," *Nature Genetics*, vol. 32, no. 4S, pp. 496–501, 2002.

[28] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[29] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole's, H. Pag'es, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan, "Orchestrating high-throughput genomic analysis with Bioconductor," *Nature Methods*, vol. 12, no. 2, pp. 115–121, 2015.

[30] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "affy: Analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.

[31] H. Pagès, M. Carlson, S. Falcon, and N. Li, *AnnotationDbi: Annotation Database Interface*, 2018.

[32] W. Chang, J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson, *shiny: Web Application Framework for R*, 2018.

[33] Y. Xie, J. Cheng, and X. Tan, *DT: A Wrapper of the JavaScript Library 'DataTables'*, 2018.

[34] I. B. Pau Ni, Z. Zakaria, R. Muhammad, N. Abdullah, N. Ibrahim, N. Aina Emran, N. Hisham Abdullah, and S. N. A. Syed Hussain, "Gene expression patterns distinguish breast carcinomas from normal breast tissues: The Malaysian context," *Pathology Research and Practice*, vol. 206, no. 4, pp. 223–228, 2010.

[35] V. Dominguez, C. Raimondi, S. Somanath, M. Bugliani, M. K. Loder, C. E. Edling, N. Divecha, G. D. Silva-Xavier, L. Marselli, S. J. Persaud, M. D. Turner, G. A. Rutter, P. Marchetti, M. Falasca, and T. Maffucci, "Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic β cells," *Journal of Biological Chemistry*, vol. 286, no. 6, pp. 4216–4225, 2011.

[36] W. Talloen, D. A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass, and H. W. Göhlmann, "I/NI-calls for the exclusion of non-informative genes: A highly effective filtering tool for microarray data," *Bioinformatics*, vol. 23, no. 21, pp. 2897–2902, 2007.

[37] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "{limma} powers differential expression analyses for {RNA}-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, p. e47, 2015.

[38] Y. Pawitan, S. Michiels, S. Koscielny, A. Gusnanto, and A. Ploner, "False discovery rate, sensitivity and sample size for microarray studies," *Bioinformatics*, vol. 21, no. 13, pp. 3017–3024, 2005.

[39] T. L. Dao, "Pharmacology and Clinical Utility of Hormones in Hormone Related Neoplasms," in *Antineoplastic and Immunosuppressive Agents*, pp. 170–192, Berlin, Heidelberg: Springer Berlin Heidelberg, 1975.

[40] C. Stolfi, R. Pellegrini, E. Franzè, F. Pallone, and G. Monteleone, "Molecular basis of the potential of mesalazine to prevent colorectal cancer," *World Journal of Gastroenterology*, vol. 14, no. 28, p. 4434, 2008.

[41] J. C. Wang, G. Y. Li, B. Wang, S. X. Han, X. Sun, Y. N. Jiang, Y. W. Shen, C. Zhou, J. Feng, S. Y. Lu, J. L. Liu, M. D. Wang, and P. J. Liu, "Metformin inhibits metastatic breast cancer progression and improves chemosensitivity by inducing vessel normalization via PDGF-B downregulation," *Journal of Experimental and Clinical Cancer Research*, vol. 38, no. 1, pp. 1–17, 2019.

[42] A. De and G. Kuppusamy, "Metformin in breast cancer: preclinical and clinical evidence," *Current Problems in Cancer*, vol. 44, no. 1, p. 100488, 2020.

[43] J. Yun, Y. G. Lv, Q. Yao, L. Wang, Y. P. Li, and J. Yi, "Wortmannin inhibits proliferation and induces apoptosis of MCF-7 breast cancer cells.," *European journal of gynaecological oncology*, vol. 33, no. 4, pp. 367–369, 2012.

[44] M. Z. Hossain, Akter, Kleve, and Gealt, "Wortmannin induces MCF-7 breast cancer cell death via the apoptotic pathway, involving chromatin condensation, generation of reactive oxygen species, and membrane blebbing," *Breast Cancer: Targets and Therapy*, p. 103, jul 2012.

[45] S. Sun, Y. Sun, X. Rong, and L. Bai, "High glucose promotes breast cancer proliferation and metastasis by impairing angiotensinogen expression," *Bioscience Reports*, vol. 39, jun 2019.

[46] D. Liu, W. Zhen, Z. Yang, J. D. Carter, H. Si, and K. A. Reynolds, "Genistein Acutely Stimulates Insulin Secretion in Pancreatic -Cells Through a cAMP-Dependent Protein Kinase Pathway," *Diabetes*, vol. 55, pp. 1043–1050, apr 2006.

[47] E. R. Gilbert and D. Liu, "Anti-diabetic functions of soy isoflavone genistein: mechanisms underlying its effects on pancreatic β-cell function," *Food Funct.*, vol. 4, no. 2, pp. 200–212, 2013.

[48] K. Itoh, T. Ishima, J. Kehler, and K. Hashimoto, "Potentiation of NGF-induced neurite outgrowth in PC12 cells by papaverine: Role played by PLC-γ, IP3 receptors," *Brain Research*, vol. 1377, pp. 32–40, mar 2011.

[49] L. Li and G.-k. Hu, "Pink1 protects cortical neurons from thapsigargin-induced oxidative stress and neuronal apoptosis," *Bioscience Reports*, vol. 35, feb 2015.

[50] J. Guo, Y. Zeng, Y. Liang, L. Wang, H. Su, and W. Wu, "Cyclosporine affects the proliferation and differentiation of neural stem cells in culture," *NeuroReport*, vol. 18, pp. 863–868, jun 2007.

[51] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I. C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O. M. Enache, F. Piccioni, S. A. Johnson, N. J. Lyons, A. H. Berger, A. F. Shamji, A. N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D. Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N. S. Gray, P. A. Clemons, S. Silver, X. Wu, W.-N. Zhao, W. Read-Button, X. Wu, S. J. Haggarty, L. V. Ronco, J. S. Boehm, S. L. Schreiber, J. G. Doench, J. A. Bittker, D. E. Root, B. Wong, and T. R. Golub, "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles," *Cell*, vol. 171, pp. 1437–1452.e17, nov 2017.