The Pharmacoepigenomics Informatics Pipeline and H-GREEN Hi-C Compiler: Discovering Pharmacogenomic Variants and Pathways with the Epigenome and Spatial Genome

by

Ari Lawrence Allyn-Feuer

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2018

Doctoral Committee:

Professor Brian D. Athey, Chair
Assistant Professor Alan P. Boyle
Professor Ivo D. Dinov
Research Professor Gerald A. Higgins
Professor Wolfgang Sadee, The Ohio State University

Ari Lawrence Allyn-Feuer

ariallyn@umich.edu

ORCID iD: 0000-0002-8379-2765

Dedication

In the epilogue of Altneuland, Theodor Herzl famously wrote:

"Dreams are not so different from deeds as some may think. All the deeds of men are dreams at first, and become dreams in the end."

Medical advances undergo a similar progression, from invisible to visible and back. Before they are accomplished, advances in the physician's art are illegible: no one differentiates from the rest the suffering and death which could be alleviated with methods which do not yet exist. Unavoidable ills have none of the moral force of avoidable ones.

Then, for a brief period, beginning shortly before it is deployed in mainstream practice, and slowly concluding over the generation after it becomes widespread, an advance is visible. People see the improvements and celebrate them.

Then, subsequently, for the rest of history, if we are lucky, such an advance is more invisible than it was before it was invented. No one tallies the children who do not get polio, the firm ground that used to be a malarial swamp, or the quiet fact of sanitation. Few remark on the novelty of the un-novel.

This thesis contemplates medical advances which do not yet exist, which are invisible. It contemplates the means by which medical science may make the genetic component of variation in disease risk, drug response, adverse events, and other medical phenotypes visible to the clinician so that informed decisions on the basis of this information may improve patient outcomes. It is offered in the fervent hope that it can contribute, in a small way, to improving and hastening the subsequent work of making these informed decisions possible.

Accordingly, it is dedicated to the patients, currently largely anonymous, who suffer invisibly in ways which can be ameliorated by genetic testing and by actions undertaken on the basis of knowledge gleaned from genetic testing. May their alleviated suffering soon be equally invisible.

Acknowledgements

This work is the result of seven years of exertion, during which time I have been aided, in a variety of ways ranging from the direct to the oblique, by a very large number of people.

It is cliché for an author to announce in the acknowledgements of a work the impossibility of acknowledging everyone who deserves it. This truism has never felt more true, for me, than it does now. I therefore first acknowledge those whom I have inadvertently omitted.

I acknowledge all the scientists in this field and outside it who shared their work with me, by publication, presentation, or conversation.

I acknowledge the patients, literally millions of them, who volunteered for the clinical research studies, including GWAS, on which this work and all the cited works are based.

I acknowledge all of my teachers, professors, editors, mentors, and other instructors over the last twenty five years.

I acknowledge my colleagues and coauthors.

I acknowledge all the members of my thesis committee, particularly my advisor, Brian D. Athey, and Gerry Higgins.

I acknowledge everyone whom I consulted for advice and who reviewed drafts of any of my publications and other documents, including this thesis.

I acknowledge everyone on whom I practiced any of my various talks.

I acknowledge everyone in the support staff of the department and university who accommodated all my various administrative and technical needs.

I acknowledge everyone who endured a social slight or any form of inattention from me because I was working.

I acknowledge all of my friends and lovers.

I acknowledge most of all the support, encouragement, and assistance of my family.

Preface

This work describes my modest contribution to a scientific project which has been underway, in one form or another, for thousands of years: the analysis and parsing of genetic inheritance and the use of genetic inferences to aid humanity's unceasing labors.

Specifically, it describes a suite of bioinformatics tools I have constructed, in an attempt to translate recent advances in the use of the epigenome and spatial genome to identify and parse important phenotypic variants, to an area of medical science, pharmacogenomics, which has largely failed to benefit from them.

The bulk of this work is based on published and soon-to-be-published papers which were collaborative in nature and have other authors. I have endeavored in this narrative to tell the story of my work, but that narrative involves collaborative work which is not fully separable. Where there is any doubt, I urge you to maximize your appraisal of my colleagues.

I begin by describing, in the first chapter, the preeminent role these advances have taken in biomedical science and in genetics in other contexts, and the disjunction between this picture and the picture in pharmacogenomics. This chapter is partially based on the 2015 review **[Higgins et al 2015]**.

I then describe, in the second chapter, three variant analyses I and others undertook to explore pharmacogenomic phenotypes with this powerful new explanatory framework, in neuropsychiatric small molecule medications, lithium, and valproic acid, and the success of these analyses in discovering mechanistic and predictive variants. This chapter is largely based on the primary papers reporting these experiments **[Higgins et al 2015, Higgins et al 2015, Higgins et al 2017]**.

I then describe, in the third chapter, the Pharmacoepigenomics Informatics Pipeline, a tool I built with the object of making such analyses easier to perform, more reproducible, and amenable to advanced methods which were not tractable in the semi-manual analytic mode in which such analyses had previously been run. I further describe the success of the PIP in reproducing the glutamatergic lithium pathway previously discovered, and in uncovering a previously unknown genetic basis for warfarin response. This chapter is largely based on the PIP paper **[Allyn-Feuer et al 2018]**.

I then describe, in the fourth chapter, the H-GREEN Hi-C compiler, a tool I built to function as one portion of a target gene module for a future version of the PIP, to address the finding of distal target genes, an important gap which existing methods could not fill. I further describe its success at detecting distal chromatin contacts existing methods could not find in a head-to-head comparison, and in building spatial contact networks of phenotypically important chromatin domains. This chapter is currently being adapted for publication.

I conclude by describing, in the fifth and final chapter, my vision for a future evolved PIP featureset, for the use of such pipelines to develop genetic tests, and for the extension of such methods in parallel across thousands of phenotypes to develop a pharmacophenomic atlas. This chapter is currently being adapted for publication, and a small part was based on the 2018 machine learning review **[Kalinin et al 2018]**.

This work is not complete; no scientific work is ever truly complete. There are logical developments to be undertaken, separate elements to combine, more experiments to perform, and ultimately, widespread clinical deployments to be undertaken. It is my fond wish that this work take place, under whatever aegis, so that future students may have more to learn, and so that some of those who are sick may be well.

Ari Lawrence Allyn-Feuer

Ann Arbor, Michigan, United States of America

July 2018

Table of Contents

List of Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 5:

Abstract

Over the last decade, biomedical science has been transformed by the epigenome and spatial genome, but the discipline of pharmacogenomics, the study of the genetic underpinnings of pharmacological phenotypes like drug response and adverse events, has not. Scientists have begun to use omics atlases of increasing depth, and inferences relating to the bidirectional causal relationship between the spatial epigenome and gene expression, as a foundational underpinning for genetics research. The epigenome and spatial genome are increasingly used to discover causative regulatory variants in the significance regions of genome-wide association studies, for the discovery of the biological mechanisms underlying these phenotypes and the design of genetic tests to predict them. Such variants often have more predictive power than coding variants, but in the area of pharmacogenomics, such advances have been radically underapplied. The majority of pharmacogenomics tests are designed manually on the basis of mechanistic work with coding variants in candidate genes, and where genome wide approaches are used, they are typically not interpreted with the epigenome.

This work describes a series of analyses of pharmacogenomics association studies with the tools and datasets of the epigenome and spatial genome, undertaken with the intent of discovering causative regulatory variants to enable new genetic tests. It describes the potent regulatory variants discovered thereby to have a putative causative and predictive role in a number of medically

important phenotypes, including analgesia and the treatment of depression, bipolar disorder, and traumatic brain injury with opiates, anxiolytics, antidepressants, lithium, and valproate, and in particular the tendency for such variants to cluster into spatially interacting, conceptually unified pathways which offer mechanistic insight into these phenotypes.

It describes the Pharmacoepigenomics Informatics Pipeline (PIP), an integrative multiple omics variant discovery pipeline designed to make this kind of analysis easier and cheaper to perform, more reproducible, and amenable to the addition of advanced features. It described the successes of the PIP in rediscovering manually discovered gene networks for lithium response, as well as discovering a previously unknown genetic basis for warfarin response in anticoagulation therapy.

It describes the H-GREEN Hi-C compiler, which was designed to analyze spatial genome data and discover the distant target genes of such regulatory variants, and its success in discovering spatial contacts not detectable by preceding methods and using them to build spatial contact networks that unite disparate TADs with phenotypic relationships.

It describes a potential featureset of a future pipeline, using the latest epigenome research and the lessons of the previous pipeline. It describes my thinking about how to use the output of a multiple omics variant pipeline to design genetic tests that also incorporate clinical data. And it concludes by describing a long term vision for a comprehensive pharmacophenomic atlas, to be constructed by applying a variant pipeline and machine learning test design system, such as is described, to thousands of phenotypes in parallel.

Scientists struggled to assay genotypes for the better part of a century, and in the last twenty years, succeeded. The struggle to predict phenotypes on the basis of the genotypes we assay remains ongoing. The use of multiple omics variant pipelines and machine learning models with omics atlases, genetic association, and medical records data will be an increasingly significant part of that struggle for the foreseeable future.

Chapter 1: Pharmacoepigenomics

Pharmacogenomics Lags Other Genetics Disciplines in Benefiting from the Epigenome

Pharmacogenomics began with pharmacokinetics, and pharmacokinetics began with association. From the antique period, commenters as early as Pythagoras noted that susceptibility to some types of food poisoning ran in families **[Pirmohamed 2001]**. More formal reports of such heritability of response to pharmaceutical drugs dates from the 1950s **[Evans et al 1961]**, and formal study of such relationships from the 1960s **[Vesell et al 1968]**. In the 1960s it was discovered that cytochrome P450 (CYP) enzymes in the liver metabolize many pharmaceutical drugs and natural substances **[Danielson et al 2002, Zanger et al 2013, Zanger et al 2014]**. With the cloning of CYP genes and the emergence of crystal structures **[Poulos et al 1987]**, it was discovered that coding SNPs altering the sequences of these enzymes altered the metabolic rates of these enzyme activities, and that differences in metabolic rates altered the peak concentrations and time courses of these drugs and their active metabolites in the body **[Skoda et al 1988]**. With these developments, the stage was set for the eventual development of tests to stratify patients by drug response by assaying these SNPs.

During the 1990s, the realization dawned that this pattern, of the approximately 60 CYP enzymes metabolizing drugs with varying speeds based on genetics, underlay not only the metabolism of

most pharmaceutical drugs **[Zanger et al 2013, Zanger et al 2014]**, but also the metabolism of many food substances, and that competition for enzyme activity underlay many drug-drug **[Guengerich 1997, Ogu et al 2000]** and drug-food interactions for substances as diverse as grapefruit **[Bailey et al 2004]**, starfruit **[Zhang et al 2007]**, and watercress **[Leclercq et al 1998]**. Even the genetic basis of Pythagoras' fava bean sensitivity was parsed, with "Favism" syndrome attributed to deficiency in Glucose-6-phosphate dehydrogenase **[Laosombat et al 2006]**, which, although not a CYP, is a metabolizing enzyme.

The search for variants that predict a phenotype is also intimately connected, in pharmacogenomics as in other disciplines, with the search for mechanisms underlying that phenotype, which may be both explanatory and also amenable to medical intervention. For example, the discovery that warfarin acts by inhibiting vitamin K epoxide reductase (VKORC1) enabled both the development of warfarin pharmacogenomics tests **[Pirmohamed et al 2013, Kimmel 2013]**, and the development of novel VKORC1 antagonist drugs **[Griminger 1987]**. The discovery of pharmacogenomic loci and mechanisms in other systems may also contribute to the development of diagnostic and therapeutic methods.

Due to this historical legacy, pharmacogenomics variant discovery has prioritized the search for protein coding variants with genetic association and biochemical methods **[Black et al 2007]**. With the sequencing of the human genome, the initial hope for immediate discovery of highly penetrant coding variants for many phenotypes **[Collins et al 2001, Ganguly et al 2001]** did not bear out, both in pharmacogenomics and in many other fields **[Weinshilboum et al 2004]**. And with the advent of GWAS (genome wide association studies), pharmacogenomic phenotypes

began to be investigated with this powerful new tool **[Giacomini et al 2017]**. While most available pharmacogenomics tests were and are based on highly penetrant coding variants, GWAS studies have revealed that the bulk of genetic variation in drug response, like many other phenotypes, comes from noncoding regulatory variants which are cooperative and combinatoric **[Boyle et al 2012]**.

In the first decade of this century, the accumulation of knowledge about such variants and genes began to take on systematic scope **[Mrazek 2010]** and to result in the construction and deployment of genetic tests to predict drug response and adverse event phenotypes in a number of clinical domains **[Mrazek 2010, Hall-Flavin 2013, Trinks et al 2014, Moaddeb et al 2013]**.

Pharmacogenomic tests developed with coding variants discovered by mechanistic work have a spotty record. They have demonstrated clinical validity and utility in systems as diverse as neuropsychiatry **[Health Quality Ontario 2017]** and oncology **[Xin et al 2017]**, and in these domains, have been used widely in the clinic. However, in other domains, great effort has been expended on the development of tests for highly heritable phenotypes without yielding clinically useful tests **[Pirmohamed et al 2013, Kimmel 2013]**. Even in domains wherein pharmacogenomic tests have been deployed and added value, they often account for only a fraction of the heritability of the phenotypes they predict **[Health Quality Ontario 2017]**.

If pharmacogenomics testing and concomitant advances in diagnosis and treatment are to realize their potential, the discipline must make wider use of high throughput methods, regulatory variants, and the epigenome. This chapter will chronicle the ways in which these methods have

transformed much of biomedical science, how they have as yet failed to transform pharmacogenomics, and the stirrings of movement in pharmacogenomics toward the "Pharmacoepigenomics" vision. It will set the stage for my contributions to this transformation, which are described in subsequent chapters.

The GWAS Interpretation Challenge

Genome wide association studies (GWAS) **[Visscher et al 2017]** have become a cornerstone technique of biomedical locus discovery over the last twelve years, and all signs point to the continuing escalation of this trend. The parallel measurement of millions of SNPs throughout the genome on a microarray at tractable cost **[LaFramboise 2009]** has allowed the traditional methods of genotypic association studies to be carried out in parallel across the entire genome. The resulting explosion in locus discovery for many systems has yielded fundamental discoveries in every area of biology and medicine **[Visscher et al 2017]**. As a result of this, the GWAS catalog **[MacArthur et al 2017]** has swelled to contain over five thousand GWAS from the published literature, while industrial concerns have amassed large numbers of proprietary GWAS.

This trend shows no sign of stopping and every sign of accelerating. Over the last five years, the genetic association methods of GWAS have been applied to two additional forms of high throughput biomedical data analysis, which add additional dimensions of parallelization across many phenotypes. In the case of molecular quantitative trait locus (mQTL) screening **[Delaneau et al 2017]**, GWAS is performed in parallel for a collection of molecular QTLs spanning the entire genome. This may include, for example, gene expression QTL analysis **[Gilad et al 2008, Peters**

et al 2016], but also including an ever-expanding array of molecular phenotypes like DNA methylation **[Banovich et al 2014, Hannon et al 2016]**, DNase accessibility **[Degner et al 2012]**, histone acetylation **[McVicker et al 2013]**, etc. And in the case of Phenome-wide association study (PheWAS) **[Denny et al 2010]**, GWAS methods are parallelized across a large number of medical phenotypes (disease diagnoses, drug responses, physiological measurements, etc) extracted at scale from an electronic medical record **[Hebbring et al 2015]**. The amount of GWAS data available, the array of phenotypes tested, the amount of undiscovered insight latent in these experiments, and the interest in interpreting them will continue to grow.

At the same time, however, the results of GWAS have not lived up to some of the original excitement. At the inception of the technique after the sequencing of the human genome, it was believed by many that the GWAS technique would lead to the recovery of the bulk of the genetic heritability of studied phenotypes **[Collins et al 2001, Ganguly et al 2001]**. As the GWAS techniques matured, however, it became clear that for most complex phenotypes, discovered loci accounted for only a fraction, often a minority, of the heritability of the phenotype **[Slatkin et al 2009, Manolio et al 2009, Gusev et al 2013]**. And in some cases, well powered GWAS recovered few or no significant loci for a heritable trait **[MacArthur et al 2017]**. This "missing heritability" problem has been the topic of perennial debate, with opinion settling around a number of hypotheses:

1) That the missing heritability is accounted for largely by independent causal variants with very small effect sizes below the detection threshold of even powerful GWAS, known as the "gold dust" **[Shi et al 2011]** or "omnigenic" hypothesis **[Boyle et al 2017]**.

2) That the missing heritability is accounted for largely by rare variants not covered in individual GWAS, known as the rare variant hypothesis **[Schork 2009, Gibson 2011]**.

3) That the missing heritability is accounted for largely by cooperative epistatic interaction between multiple loci, rather than linear summation of the statistical heritability of individual loci, known as the epistatic hypothesis **[Zuk et al 2011, Sivakumaran et al 2011]**.

4) That the missing heritability is accounted for largely by direct epigenetic inheritance not mediated by DNA sequence, known as the epigenetic inheritance hypothesis **[Slatkin 2009, Franklin et al 2010]**.

5) That the missing heritability is largely illusory, the result of shared environmental factors between family members creating concordant phenotypic outcomes which inflate heritability estimates, known as the environmental hypothesis **[Gage et al 2016]**.

Despite the enduring popularity of the environmental hypothesis in the lay press **[Rossiter 1996, Feldman et al 2018]**, detailed investigation of family members of different levels of relatedness, raised separately or together, from the same and different pregnancies, have decisively refuted this hypothesis for a number of well-studied phenotypes **[Plomin et al 2015, Plomin et al 2018]**. Less clarity has emerged on the topic of the omnogenic, rare-variant, epistatic, and epigenetic hypotheses. It is likely that each of these factors contributes to the heritability problem to some

extent and that such extents vary for different types of phenotypes, but this landscape remains murky.

Interpreting GWAS results is a very difficult problem. It involves a number of major challenges, principally discerning the reality of associations, attributing causal character to them, finding the causal variants within linkage regions, and discerning their function.

In GWAS design, investigators typically attempt to find a socially and genetically homogeneous population with a good match between cases and controls, in order to minimize the potential for lifestyle and environmental factors with a causative role in the phenotype to cause erroneous association hits through their association with ethnicity-linked or other genetic variants **[Amos et al 2007]**.

The Nexus Between Genetics and Epigenomics Integrates Environmental Stimuli

To some extent these distinctions may be artificial; recent results document a significant enrichment of regulatory SNPs identified by GWAS in chromatin domains carrying epigenomic marks consistent with enhancer functions, affecting transcription factor (TF) binding sites (TFBS) **[Farh et al 2015]**. As a result, specific causal variants affect TF binding and recruitment of epigenomic mechanisms, which leads to alteration of enhancer and promoter function. This event can be detected with chromatin immunoprecipitation methods, as documented for *VKORC1* (vitamin K epoxide reductase complex subunit 1) **[Wang et al 2008].** These effects are often tissue-specific and influenced by external stimuli, such as drug exposure, generating a dynamic

nexus between genetics, epigenetics, and environmental factors such as stress, abuse, famine, and others **[Provencal et al 2014, Stankiewicz et al 2013, Yao et al 2014]**.

Because of chromatin looping between enhancers and promoters, a genetic variant at one enhancer can influence epigenomic changes over long distances. As each enhancer and promoter can reach multiple target regions, evidenced by chromatin conformation assays, a single variant can spread epigenomic marks across multiple chromatin regions and distinct gene loci. This paradigm is incompatible with a simplistic gene variant-effect relationship, but rather highlights the highly non-linear and adaptive dynamic nature of genomic-epigenomic processing.

Statistical Problems in GWAS Interpretation

Performing millions of association tests in parallel creates serious multiple comparison problems, rendering statistical significance far from a binary. Because of the linkage between adjacent loci, not all of the methods for multiple comparison control are appropriate for GWAS **[Benjamini et al 1995, Blanchard et al 2009]**. The resultant widespread use of Bonferroni comparison control (in the form of a single threshold of 1e-8 for "genome wide significance") **[Johnson et al 2010, Fadista et al 2016]** has exacerbated the statistical problems. In many cases, causative variants discovered in subsequent GWAS on a particular phenotype were trending toward significance in prior less-powered GWAS, but were not distinguishable from noise with those samples and the methods then in use **[Visscher et al 2017]**.

However, this situation also resists the thoughtless use of more sophisticated tools like the Benjamini and Hochberg FDR **[Benjamini et al 1995]**, which evade the stringency of the Bonferroni method but have the cost of assuming statistical independence between comparisons. With adjacent loci exhibiting linkage, the naïve use of FDR methods may allow a cluster of linked SNPs around a lead SNP to falsely "demote" real and significant hits. Accordingly, thinking in this area is trending toward the use of fine mapping and coimputation approaches which are informed by linkage structure and do not blithely assume either a single genome wide threshold or statistical independence **[Fadista et al 2016, Wang et al 2016]**.

Once investigators are confident of the reality of a GWAS hit, there is still no certainty which variant or variants are causatively linked to the study phenotype and responsible for the association. Significance regions typically span a number of linked variants all attaining genome wide significance of association with the study phenotype. And because over 90% of causative GWAS SNPs are noncoding regulatory SNPs **[Boyle et al 2012]**, it is often unclear what the mechanism of causation is: which genes are being regulated by the causative SNP, in what tissues, under what conditions.

The Epigenome

During the same period of time, array- and sequencing-based assays for a large number of epigenome features, and experiments and atlases conducted with such assays, have revealed that the multifaceted, tissue-specific epigenome has a relationship with gene expression, cell fate, organismal function and dysfunction, and disease, which is both intricate and powerful.

In contrast to the genome, which is relatively well defined, the epigenome comprises a large and growing number of modalities measurable with various assays, most of which take the form of "tracks" comprising numeric levels of observed signal at various positions in the genome, and often a genome-wide vector of signal. These include gene expression (RNA microarrays and RNA-seq, and proteomics) **[Lonsdale et al 2013]**, DNA methylation (assayable with MeDIP-seq **[Staunstrup et al 2016]**, RRBS **[Yong et al 2016]**, WGBS **[Olova et al 2018]**, and other methods), hydroxymethylation (hydroxylmethylation WGBS) **[Huang et al 2010, Wen at al 2016]**, chromatin accessibility (DNase-seq **[Song et al 2010]**, ATAC-seq **[Buenrostro et al 2015]**), histone post-translational modifications principally including acetylation and mono-, di-, and trimethylation of lysine residues at H3K27, H3K9, H3K36, and H3K4 **[Roadmap Epigenomics Consortium 2015]**, but also including a large and growing number of subsidiary histone marks (ChIP-seq, ChIP-Exo), transcription factor binding for hundreds of transcription factors (ChIP-seq), and others. In addition to this are epigenome assays relating to the spatial and functional genome which express themselves as contacts and relationships in squared and other higher-dimensional genome spaces, including molecular QTL (mQTL) screening, enhancer mapping, and spatial genome measurements including 3C **[Dekker et al 2002]**, 4C **[Simonis et al 2006]**, 5C

[Dostie et al 2006], Hi-C **[Lieberman-Aiden et al 2009]**, ChIA-PET **[Li et al 2014]**, Genome Architecture Mapping **[Beagrie et al 2017]**, Hi-ChIP **[Mumbach et al 2016]**, and SPRITE **[Quinodoz et al 2018]**.

Epigenome elements differ from cell type to cell type and across the cell cycle in multicellular organisms **[Roadmap Epigenomics Consortium 2015]**, and from person to person **[Flanagan et al 2006, Schneider et al 2010]**, and according to physiological, environmental, and medical conditions **[Schneider et al 2010]**. Thus, the space of possible assays dwarfs any real dataset. Nevertheless, there have been increasingly systematic attempts to produce comprehensive spanning sets of epigenome data with standardized methods, yielding a set of epigenome "atlases" under the rubrics of ENCODE **[ENCODE Project Consortium 2012]**, the Epigenome Roadmap **[Roadmap Epigenomics Consortium 2015]**, the International Human Epigenome Consortium **[Stunnenberg et al 2016]**, and the upcoming Human Cell Atlas **[Regev et al 2017]**, as well as more focused efforts from many quarters. IHEC data now includes a set of core epigenome marks for over a hundred tissues throughout the human body. As a result of this, despite the inherent sparsity of any real dataset, the epigenome is increasingly regarded like the reference genome: as a resource to be consulted for systems and loci of interest, rather than an unknown quantity to be queried experimentally in specific contexts.

The epigenome atlases have identified a set of "core" epigenome elements which determine a set of chromatin states corresponding to the various categories of regulatory states (for genes) and regulatory elements (for noncoding regions of the genome). The most influential method for calling chromatin states is ChromHMM **[Ernst et al 2012]**, which uses a hidden Markov model

on “core” epigenome tracks to call fifteen chromatin states including seven types of promoters/enhancers and eight types of activity/repression states. Chromatin states which emerge in particular genomic locations in particular tissues have become a widely used and powerful guide to the functions of the host loci and the tissues in which they operate.

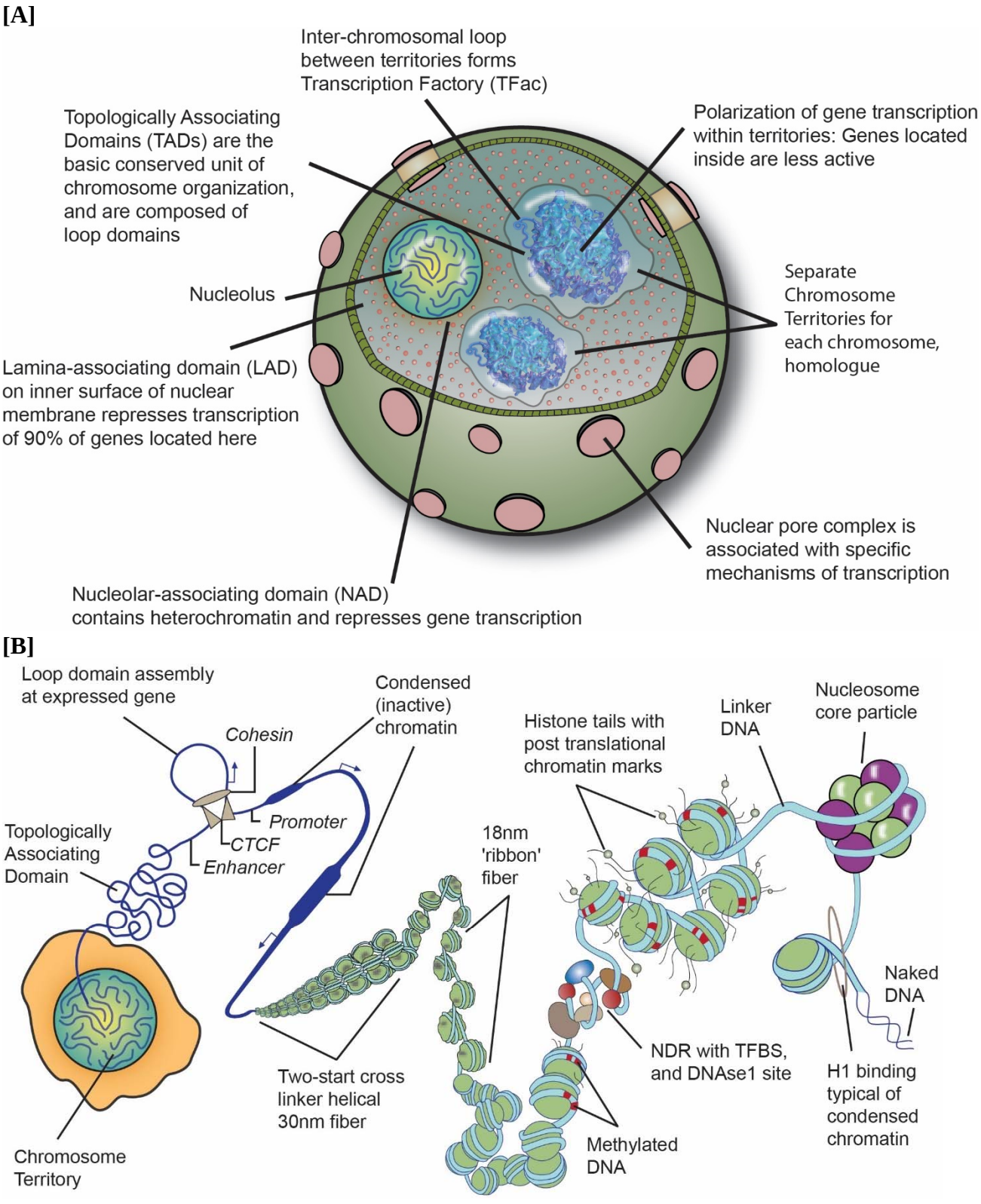


Figure 1-1. The Spatial Epigenome from the Nucleus to Atoms

Schematic of a generic interphase nucleus showing different compartments involved in determining transcriptional regulation. **[A]** Higher order components of 4D nucleome organization, including the nucleus as a whole, the nucleolus, and chromosome territories containing TADs, LADs, NADs, and loop domains in Transcription Factories. **[B]** Lower order components of 4D nucleome organization, including chromatin fibers of several types, nucleosomes and histones with associated subtypes and covalent chemical modifications, and naked DNA. In the nucleosome core particle, H3 and H4 are shown in purple while H2A and H2B are shown in green. See text for further details.

In addition to this, the validation of enhancer and promoter elements with reporter assays based on the transcription of their target genes under the influence of element excision with CRISPR genome editing **[Lopes et al 2016, Gasperini et al 2017, Klein et al 2018]**, what is referred to as "CRISPR validation" of an element, has been increasingly used. Under the influence of these methods, researchers are increasingly able to assess the function, activity, and targets of regulatory elements in a tissue-specific manner by consulting these resources, even in the absence of any purpose-specific experiments.

In addition to this, a powerful set of tools have emerged that use machine learning and explicit algorithms to make predictions about the effects of sequence changes and other perturbations on epigenome outcomes. These have included both explicit algorithms and machine learning methods for predicting TF binding, machine learning methods for predicting variant effects on epigenome tracks and chromatin states, and machine learning models for predicting enhancer targets.

Position weight matrices (PWMs) **[Stormo et al 1982]** have been used to define transcription factor motifs since the 1980s. Despite their algorithmic crudity, their performance at defining binding sites has been surprisingly strong relative to competing methods, and they have the advantage of being very simple to define and use, and very interpretable. With the sequencing of the genome and genome wide ChIP-seq experiments for many transcription factors, it became possible to define motifs for large collections of TFs **[Liefooghie et al 2006]** and assay them genome wide. With the reference and alternate alleles of SNPs, it is thus possible to gauge the comparative adherence of the flanking sequences to the TF under the influence of the various alleles of the SNP. This is performed at scale by algorithms like TFM-Scan **[Liefooghie et al**

2006] and MotifDB **[Shannon et al 2018]**. In addition to this, recent work by Nishizaki **[Nishizaki et al 2017]** and others has focused on predicting the occupancy of TF binding sites by reference to both variants and the chromatin states exhibited by the host loci in tissue-specific data.

In addition to this a large number of machine learning algorithms have been published and widely used that predict the influence of variants on epigenome tracks, especially chromatin accessibility, and on chromatin state. Support vector machines based on gapped k-mers and trained on positive and negative sequences from a specific epigenome track emerged as a potent predictor of variant effects on chromatin accessibility beginning with gkm-SVM **[Gandhi et al 2016]**, deltaSVM **[Lee et al 2015]**, and lsGKM **[Lee 2016]**. But in addition to this, deep neural networks including DeepSEA **[Zhou et al 2015]** and Basset **[Kelley et al 2016]** have been trained on collections of epigenome tracks in a large number of tissues to predict variant effects on epigenome tracks and chromatin states in specific tissues, and have exhibited performance increases over SVM-based methods, including the ability to attain increased speed of training by transfer learning from a trained multi-tissue model to a single-tissue model.

Finally, integrative computational methods for predicting enhancer target genes on the basis of tissue-specific integrative epigenomics and Hi-C data have proved extremely effective at predicting proximal enhancer targets **[Spurrell et al 2016, Hardison et al 2014]**. Such methods have prominently included TargetFinder **[Whalen et al 2016]**, which took an integrative approach, as well as methods of Lieberman-Aiden et al **[Durand et al 2016]**, which are based purely on Hi-C and sequence data.

The Spatial Genome

The period since the sequencing of the genome has also seen two epochal discoveries advancing in parallel: firstly that the spatial organization of the genome inside the 3D space of the nucleus varies from cell type to cell type **[Rao et al 2014]**, cell to cell **[Nagano et al 2013, Stevens et al 2017, Ramani et al 2017]**, condition to condition **[Chen et al 2017]**, and over time **[Seaman et al 2018]**, and has a powerful and intricate connection with gene expression, cell fate, and organismal function and dysfunction including disease, and secondly that such organization can be measured in parallel across the multidimensional vector spaces over genomic position which represent combinatorial genomic interactions **[Quinodoz et al 2018]**.

Prior to the emergence of genome sequencing based biochemical assays, the existence of a role for genomic spatial organization was already known from imaging studies and the biochemical methods of chromosome conformation capture (3C) **[Cremer et al 2000, Dekker et al 2002]**, to some extent. It was known that chromosomes in the interphase nucleus segregated into chromosome territories (CTs) resolvable with fluorescent in-situ hybridization (FISH) probes **[Cremer et al 2000]**, that the loci involved in common translocations in cancer were often spatially close to each other in normal interphase nuclei of the tissues from which the cancers arose **[Lee et al 1993]**, and that transcriptional activity demarcated by RNA polymerase II (Pol2) took on a punctate form inside the nucleus **[Hughes et al 1995, Eskiw et al 2008]**.

But such awareness took on a new height with the advent of new measurement methods. In imaging, labeling, imaging, and interpretation all took on new power. Not only did FISH probes

of the cell nucleus take on a new precision with the availability of genome-wide collections of bacterial artificial chromosome (BAC) libraries for the easy generation of FISH probes for arbitrary loci **[Baumgartner et al 2006]**, but oligomer FISH (Oligo-FISH) methods **[Schmitt et al 2010]** became easier and cheaper to perform with the availability of the reference genome, bioinformatics tools for probe design, and the rapidly plunging cost of de novo DNA synthesis. The wider availability of 3D confocal microscopes and more powerful imaging methods like 3D SIM **[Shao et al 2011]**, PALM **[Betzig et al 2006]** and STORM **[Rust et al 2006]** super resolution microscopy, labeling methods for multiple probes at the same time, spectral deconvolution **[Zimmerman et al 2003]** and white light lasers **[Chiu et al 2012]** to make channels easier to separate, and other methods made it possible to discern label positions better than ever before. CRISPR imaging **[Chen et al 2013, Chen et al 2014]** made it possible to image nuclear labels in live cells. And the proliferation of 2D and 3D image parsing packages, including the work of Rajapakse et al **[Rajapakse et al 2011, Seaman et al 2015]** and Misteli et al **[Shachar et al 2015, Jowhar et al 2018]**, as well as Kalinin et al **[Kalinin et al 2017]**, made it easier to extract high content semantic information from nuclear architecture imaging.

This imaging work confirmed the earlier explorations of the pre-genome era and added new discoveries. It was discovered that the punctate elements of Pol2 activity were in fact collections of co-regulated genes from different chromosomes being transcribed and regulated together **[Papantonis et al 2013]**, that they assemble stochastically on timescales of minutes for transcriptional bursting **[Ghamari et al 2015]**, that they are located on the periphery between chromosome territories and preferentially located in the center of the nucleus **[Cremer et al 2015]**, and that their assembly predates transcription **[Krijger et al 2017]**. It was discovered that

repressive lamin-associated domain (LAD) elements in the genome are transcriptionally repressed **[van Steensel et al 2017]** and that escape from transcriptional silencing depends on spatial escape from the lamin **[Robson et al 2017]**. It was discovered that proximal promoters and controlling enhancers unite spatially for transcription **[Rao et al 2014]**, including super enhancers controlling collections of co-regulated genes **[Thibodeau et al 2017, Gong et al 2018, Huang et al 2018]**. And experiments in serial FISH of a collection of FISH probes along the length of a chromosome showed that the spatial compaction of collections of loci within a collection of loci recapitulates their contact frequencies as measured by 3C-based methods **[Wang et al 2016]**.

And the developments in sequencing-based methods for gauging chromatin spatial information were even more profound. The sequential development of parallel 3C-based methods including 4C (one genomic location against the genome) and 5C (a collection of probes against each other) reached a disjunction with the development of Hi-C: high throughput chromatin conformation capture. It involves cross-linking fixed cells and digesting the genome with a restriction enzyme (or, for Micro-C, mainly used in small genomes, an unselective endonuclease like MNase), followed by religation, sonication, and paired end sequencing. The result is a paired end sequencing library wherein the genomic locations of the paired reads do not correspond to elements close to each other in sequence space, but in physical space. Such reads can be compiled into a chromatin contact map, potentially of the entire squared genome.

This method, which allowed parallel 3C-style measurement of the entire genome against the entire genome, producing maps of chromatin contacts in square genome space, electrified the field of chromatin structure and became the standard in genome structure research. Progressive protocol

optimization has allowed the methods to produce higher quality data with lower effort, and to address a wider collection of cell lines and tissues. Experiments in a large number of cell lines and conditions have highlighted the commonalities in genome organization and the ways that organization differs over time. Hi-C data is increasingly being used for fundamental genomics tasks like assembling reference genomes **[Korbel et al 2013]**, phasing haplotypes **[Ben-Elazar et al 2016, Selvaraj et al 2013]**, and detecting chromosome translocations **[Chakraborty et al 2017]**.

But even more than this, Hi-C data has been produced a revolution in our understanding of spatial genome organization, the largest component of which has been the discovery of topologically associating domains (TADs) **[Dixon et al 2012]** in the human genome and other genomes. These self-associating regions have been identified as potent spatial and functional elements in the human genome. The division of approximately 80% of the human genome into approximately 2500 TADs is remarkably robust, being largely conserved between cell types in the human body **[Rao et al 2014]**, between different humans **[Ruiz-Velasco et al 2017]**, and under disease states **[Rao et al 2014]**. In fact, syntenic regions of related genomes (e.g. mouse) often share the same TAD structure as the related regions of the human genome **[Krefting et al 2017, Nora et al 2013]**. TADs also function as replication domains **[Pope et al 2014]**. Moreover, TADs mediate long range spatial interactions **[Rao et al 2014]**: the contact frequency in any given portion of the squared genome will more closely correlate with a more sequence-distant portion which is in the same TAD pair than a sequence-proximal portion spanning TAD boundaries.

While the portion of a genome which composes a TAD is relatively invariant, TADs differ from one cell type and biological condition to another in their degree of transcriptional activity. This differentiation between the "A" and "B" compartments is connected with the sign of the dominant eigenvector of a genome-wide Hi-C matrix **[Dekker et al 2013]**, the degree of spatial openness as observed by both Hi-C, imaging, and biochemical experiments **[Roadmap Epigenomics Consortium 2015]**, the degree of gene expression **[Roadmap Epigenomics Consortium 2015]**, the presence of active histone marks **[Roadmap Epigenomics Consortium 2015]**, the presence of activating transcription factors **[Roadmap Epigenomics Consortium 2015]**, replication timing **[Pope et al 2014]**, and the extent of long range and interchromosomal contacts **[Rao et al 2014]**. Genes located in the same TAD tend to be co-regulated, and they are regulated partially by their TAD context.

Moreover, the sequence context governing these regulatory interactions is beginning to illuminate under sustained investigation. High resolution Hi-C experiments have discovered a hierarchy of super- and sub-TADs running all the way down to the level of "loop domains" **[Rao et al 2014]** which spatially unite the proximal enhancers of genes with their intra-TAD proximal enhancers. These contacts, and those above them in the hierarchy, are largely governed by the presence of convergent pairs of CTCF sites **[Nichols et al 2015, de Wit et al 2015]** which form a CTCF-cohesin anchor binding loci together by means of a loop extrusion mechanism which has been verified by biochemical and imaging methods **[Fudenberg et al 2016, Sanborn et al 2016]**. Perturbations of these sequence elements by CRISPR in vitro, or by disease-related mutations in vivo, as for example in enhancer hijacking in cancer, have the predictable effects on Hi-C maps, the epigenome, and gene expression **[Wutz et al 2016, Sanborn et al 2016]**.

The march of Hi-C in genomics shows every sign of continuing to escalate, with the number and variety of Hi-C experiments escalating year by year. Recent work shows that genome assembly benefits from Hi-C data **[Korbel et al 2013]**, so that in the future, Hi-C library preparation may be used for routine genome sequencing, making Hi-C data available in larger numbers of samples than any other epigenome modality (aside from gene expression). In addition, the upcoming Human Cell Atlas **[Regev et al 2017]** is widely expected to feature in-depth Hi-C data on every major cell type in the human body. Single cell Hi-C is shining a light on the variation of cellular organization between cell types and over the cell cycle. Promoter capture Hi-C **[Mifsud et al 2015]** is being explored as a medical diagnostic **[Mishra et al 2017, Baxter et al 2018]**.

And new methods of gauging chromatin organization with sequencing continue to proliferate, offering new discernment of various features. Micro-C **[Hsieh et al 2015]**, using endonucleases instead of restriction enzymes, has carried the resolution of Hi-C all the way down to the nucleosome level in small genomes (e.g. yeast and drosophila). SPRITE **[Quinodoz et al 2017]**, which uses serial dilution and combinatorial labeling to uniquely suspended label chromatin complexes before single ended sequencing, is producing deeper Hi-C style maps with less sequencing, and also allowing the exploration of multiple-locus contacts in higher order combinatorial genome spaces. Genome Architecture Mapping **[Beagrie et al 2017]**, which uses single cell sequencing of cryosectioned nuclei followed by statistical modeling of the co-occurrence of pairs of sequence regions, cannot produce high resolution intra-TAD information, but generates Hi-C style contact maps with less sequencing, and also can be used to produce calibrated physical distances.

This proliferation of data is increasingly being used to form three dimensional models of chromosome territories and even the entire nucleus. Early work in this area met with difficulty due to low resolution data, the complexity of the task and the overconstrained nature of the data, and the challenges caused by structural heterogeneity **[Nagano et al 2013]**. This may be clearly seen by the preeminence and then rejection of the "superaxis" model of chromosome territory structure, in which the sequence of a chromosome is approximately recapitulated by the spatial ordering of TADs along a "superaxis," with the string of tads threading back and forth between the A and B compartments of the territory, possibly by means of a coil. Multiple spatial modeling investigations based on different modeling methods and different Hi-C datasets independently discovered the superaxis **[Hu et al 2013, Nagano et al 2013]**, but it was subsequently demolished by the publication of serial FISH data on threading TAD positions in chromosome territories in single cells **[Wang et al 2016]**. The latest modeling methods use single cell data to refine models based on higher-resolution ensemble Hi-C **[Lando et al 2018]**, and do not produce superaxis behavior in the models. They show better accord with 3D FISH data than prior models.

Spatiotemporal Gene Regulation: From Atoms to Cells and Seconds to Decades

Multiple scales of spatial resolution demonstrate that the three dimensional (3D) organization of the nucleus, and its chromatin components, provides the basis for the regulation of gene expression on a genome-wide level and on a per gene basis. These 4D nucleome components exist on spatial scales ranging from higher order nuclear structures including chromosome territories, the nuclear lamina, nuclear pore complexes, nucleoli **[Cremer et al 2001, Wendt et al 2014]** to transcription

factories (TFac) that unite topologically-associating domains (TADs) **[Dixon et al 2012, Lieberman-Aiden et al 2009, Rao et al 2014]** from multiple portions of one or more chromosomes for coordinated transcription, to the coordinated loop domains **[Rao et al 2014]** that hold active genes open in TFac TADs, through chromatin fibers down to naked DNA. In the temporal dimension, phenomena such as the circadian rhythmicity of gene expression **[Zhang et al 2014]** and the timing of interactions between specific *CLOCK* (clock circadian regulator) gene loci **[Chena et al 2015]** operate on scales ranging from the rapid transcriptional stimulus response, to ultradian rhythm, to the circadian rhythm, to the life cycle of organisms.

The genome in the nucleus is divided into chromosome territories, discrete spatial domains physically containing the DNA of each chromosome **[Chena et al 2001, Cremer et al 2010]**. In a diploid human cell, there are a total of 46 chromosome territories; the autosomes occupy distinct territories and are differentially regulated. There is increasing recognition that allele-biased transcription, both stochastically and in a coordinated imprinted manner, is quite common in mammalian cells as determined using single cell RNA-seq and 3D FISH (Fluorescence *In Situ* Hybridization) at the cellular level **[Deng et al 2014, Palacios et al 2009].** The chromosome territories fold into Topologically Associating Domains (TADs) **[Dixon et al 2012, Rao et al 2014]**, compact, self-associating regions whose sequence extent is canonical between tissues of the body and between individuals, and whose boundaries are marked by coordinated CTCF and cohesin binding. These TADs condense in a fractal globule polymer model, whose configuration is regulated in response to the transcriptional program of the nucleus. TADs with transcribed genes migrate preferentially to the exterior of chromosome territories in the nuclear interior, where they mingle in TFacs for coordinated regulation. In contrast, lamin-associated domains (LADs) located

near the inner surface of the nuclear membrane, repress approximately 90% of gene expression as a consequence of their location in heterochromatin. A layer of heterochromatin also surrounds nucleoli, so localization of a gene in this nuclear compartment using datasets from, e.g. the human brain atlas of the Allen Brain Science Institute **[Sunkin et al 2013]** is suggestive of repression within nucleolar-organizing or associated domains (NADs). An overview of the higher level 4D nucleome components and their association with classes of organizing domains (TADs, LADs, and NADs) is shown in **Figure 1-1A.**

Within TADs, the genome is organized into chromatin loop domains. Chromatin loops form the basis of transcriptional regulation of many genes, as confirmed by the results of a recent study, which demonstrated that the fundamental functional topology of transcriptional regulation at a genome-wide level in humans occurs within spatial interaction networks **[Rao et al 2014]**. These are composed of enhancer-promoter loops that regulate transcription within and between chromosomes in *cis* and *trans* **[Rao et al 2014]**. Loop domains playing host to active transcription have a signature spatial interaction distinct from repressed loops, detectable with spatial mapping techniques **[Rao et al 2014]**. Inter-chromosomal loops have been visualized using super-resolution microscopy **[Cremer et al 2001]**, and these *trans* interactions may not only be functionally active, but the underlying loci's status as active TADs or repressive transcriptional hubs may be propagated through cell division and transgenerational inheritance. These features may be seen in **Figure 1-1A and B**.

At the fundamental level, 146 base pairs of naked DNA wrap in one and three quarters turns around a hetero-octamer of Histone proteins composed of 1 tetramer of Histones $H3_2H4_2$, and 2 histone H2A-H2B dimers, to form the nucleosome core particle, a disc 11nm in diameter and 5.5nm in

height. Nucleosome core particles are connected by variable lengths of linker DNA. A fifth histone subtype, Histone H1, known as the "linker Histone" regulates fiber condensation in heterochromatin. The subtype of histone proteins which make up the octamer, post translational modifications to these proteins, their position and frequency along the sequence of the genome (called nucleosome repeat length), and their organization into several types of euchromatin and heterochromatin chromatin fibers, are all epigenomically significant and regulated both by base sequence and programmatic context **[Bannister et al 2011, Luger et al 2012]**. Areas of active transcription are associated with nucleosome variants and post translational modifications indicative of lower binding strength, coordinated nucleosome positions around transcription start sites and generally lower density, more DNase sensitive sites, and less condensed, less regular fibers lacking in H1 binding **[Woodcock et al 2010]**. Repressed regions carry Histone variants and post translational modifications conducive to tighter binding, and regular nucleosome spacing, with condensation by H1 binding into regular, compact crossed-linker fibers with a diameter of about 30nm **[Williams et al 1986, Athey et al 1990]**. This area of lower level chromatin regulation is a rich matter for research with a number of techniques. Such features are shown in **Figure 1-1 B**.

Pharmacoepigenomics

Integrative Epigenome Models of Variant Function Applied to Association Hits

The combined power of multiple epigenome modalities to interpret variant function began being exploited to address the interpretation challenges of GWAS around the time of the publication of

the Phase 2 ENCODE maps **[ENCODE Project Consortium 2012]**, and was in wide development by the time the Epigenome Roadmap **[Roadmap Epigenomics Consortium 2014]** data landed. The most visible sign of this was in the variant annotation and interpretation methods being published, prominently including the early contributions of RegulomeDB **[Boyle et al 2012]** and HaploReg **[Ward et al 2012]**.

These pipelines work by annotating SNPs multimodally with information from multiple types of omics information. In the case of RegulomeDB, this included DNase sensitive regions, validated promoter and enhancer regions, transcription factor binding sites, and predicted regulatory elements. HaploReg expanded on this suit by adding tools to gauge the relevance of SNPs in the area of a lead SNP with linkage, and looking at PWM perturbation by SNPs, deeper epigenome data from the Roadmap, and eQTL data. Both have been widely used.

All such models rely on an overall paradigm of genomic regulation sometimes described as the "pharmacoepigenome," **[Higgins et al 2015]** in which, by virtue of the facts that 1) associations arise from causative regulatory variants influencing the underlying genomic machinery of a phenotype, 2) regulatory variants and their targets can be identified and parsed by looking for the hallmarks of regulation in epigenome datasets, and 3) the machinery underlying genetic variation in a phenotype is often joint with the machinery underlying the phenotype itself and related phenotypes, it is concluded that powerful insights into the mechanisms of a wide variety of phenotype, not necessarily constrained to pharmacogenomics, can be obtained by looking for tissue-specific regulatory variants and their targets in the regions surrounding association hits for a collection of related phenotypes.

Such models have been used to interpret GWAS, looking in significance regions for regulatory variants, by a large number of different methods mostly comprising individual ad-hoc analyses. They differ significantly but largely adhere (fully or partially) to a number of common themes:

1) Analyzing multiple GWAS on the same (or related) phenotypes together in the same analysis.

2) Analyzing many variants, not just lead SNPs. For primary GWAS this is often done with the significance region or with a p-value fold change cutoff from the lead SNP. In secondary analysis, it is often done by linkage. Some analyses have used sequence distance cutoffs, but this is not biologically informed and is inadvisable.

3) Looking for regulatory variants with tissue-specific data that is relevant to the phenotype under investigation.

4) Attempting to identify the target genes of candidate regulatory variants discovered in the analysis.

5) Filtering, ranking, and organizing the result genes together with pathways and ontologies, looking either for genes of preexisting known relationships to the phenotype, or genes that cohere with each other, e.g. by shared pathway or ontology membership, or coregulation by a known transcription factor.

Such analyses have become more common, with epigenome-based causal variant methods appearing in even primary GWAS analyses.

Early Pharmacoepigenomics Efforts in Neuropsychiatry

The intent of GWAS is to detect statistical correlation between genotypes, often SNPs, and a phenotype of interest using a large number of samples. As opposed to candidate gene association studies often used in neuropharmacology and in CNS pharmacogenomics, the objective is to enable unbiased prediction of phenotype based on allelic data. Early GWAS of psychiatric pharmacogenomic phenotypes yielded two startling results: (1) Most associations did not reach the stringent statistical threshold required to rule out false positive associations, which was based on Bonferroni correction for multiple comparisons **[Gao et al 2010]**, and (2) The majority of pharmacogenomic SNP associations detected are located within non-coding regions of the genome, including introns and intergenic domains, as is true in GWAS for other traits **[Farh et al 2015]**. It is now understood that these noncoding variants impact functional regulatory elements such as enhancers, and that differential regulation of enhancer-promoter interactions within chromatin interaction networks is largely responsible for variation in pharmacogenomic response, contributing much more than non-synonymous variation in the exome **[Kellis et al 2014]**.

The early agnostic approach to GWAS analysis required stringent adjustments for multiple hypotheses testing; however, with more information available on each variant, a new approach has emerged, taking advantage of prior knowledge. Since the first GWAS was conducted in 2006, there has been an increasingrecognition that the parameters needed to be changed, with consensus

around several modifications of the original precept: (1) Focus has been placed on the use of noncoding variants found in GWAS as components of related biological pathways **[Califano et al 2012]**; and (2) Joint GWAS analyses of related disease phenotypes are now commonly accepted as over-lapping sets of variants are becoming the norm, not the exception **[Califano et al 2012]**.

The clinical utility of any biomarker must be based on studies in humans. However, in variant discovery, explanatory and predictive annotation of human GWAS using a bioinformatics framework has the potential to provide insight into previously unrecognized pathways. There is precedence in the use of such approaches including: (1) Imputation and chromatin state annotation of GWAS SNPs in the context of the regulome **[Califano et al 2012, Schaub et al 2012, Tasan et al 2014, Boyle et al 2012]**; (2) Interpretation of GWAS SNPs using known cellular transcriptional networks with inclusion of new sequencing data **[Kellis et al 2014, Network and Pathway Analysis Subgroup 2015]**; (3) Joint analysis of multiple, seemingly unrelated disease GWAS to discover novel shared pathways **[McGeachie et al 2014, Kuhn 1962];** and (4) "Prix fixe" scoring using multiple GWAS to uncover shared regulatory genes involved in biological pathways **[Tasan et al 2014]**.

Joint analysis of GWAS in psychiatric disorders using open source pathway analysis software has implicated shared, common pathways among multiple disorders such as bipolar I disorder, major depressive disorder and schizophrenia **[Network and Pathway Analysis Subgroup 2015]**. Pathways from public databases were compiled, then significant SNPs from psychiatric GWAS were gathered as components of pathways. These pathways were then assessed statistically using different strategies, ranked and averaged. Empirical p-values were calculated by comparing

average ranks in disease data to average ranks in simulated data. Finally, a combined p-value for bipolar disorder 1, major depressive disorder and schizophrenia study showed that a significant shared effect was the methylation of histones H3K4 (histone H3 lysine), which are defining histone marks that differentiate H3K4me1, a histone mark that defines enhancers, from H3K4me3, a histone mark that defines promoters **[Ernst et al 2010]**.

The Paucity of Pharmacoepigenomics in Clinical Practice

By contrast, the discipline of pharmacogenomics has historically made little use of any of the advanced variant discovery and interpretation methods discussed above. To date, the most widely used biomarkers for clinical pharmacogenomic testing are a set of pharmacokinetic (PK) gene variants located in CYP genes, encoding the main drug metabolizing enzymes, while only a small family of pharmacodynamic (PD) genes have been utilized, and regulatory variants, even less **[Higgins et al 2015]**. While such tests often achieve clinical utility and cost-of-care reduction **[Higgins et al 2015]**, they often account for a minority of the inter-individual genotypic variation in important drug-related phenotypes **[Higgins et al 2015]**. By and large, however, these tests are designed manually using variants in a small pool of candidate genes.

Although pharmacogenomic phenotypes have been investigated with GWAS since its inception **[Giacomini et al 2017]**, most available pharmacogenomics tests continue to be based on highly penetrant coding variants revealed by gene-specific work, with GWAS findings, PD genes, and regulatory variants persistently underutilized. Deployment of GWAS in pharmacogenomics variant discovery has lagged deployment in other disciplines, epigenomic interpretation of GWAS

results has been underutilized, and the translation of GWAS results into clinical tests has been slower still in most areas.

Indeed, much pharmacogenomics variant discovery still proceeds along traditional lines involving the search for coding variants to be designated as star (*) alleles, with a particular emphasis on PK genes for absorption, distribution, metabolism, and excretion (ADME). Such genes have formed the focus for test development for response, dosing, and adverse drug events (ADEs) and adverse drug reactions (ADRs).

For example, consider neuropsychiatry. As of 2015, the Table of Pharmacogenomic Biomarkers used in drug labeling from the U.S. Food and Drug Administration (FDA) listed 145 precautions for 24 psychiatric medications, 9 neurology medications, and 2 anesthetics. None of these precautions reference PD genes expressed predominantly in the CNS; all are CYP metabolizer subtypes of genes expressed in liver, drug-drug interactions, or gene effects in peripheral tissues (Table 1). The predominance of CYP gene variants in early research and testing for risk of ADEs led to a commonly-used, CYP-based phenotypic stratification of patients into metabolizer classes such as poor, intermediate, extensive, and ultra-rapid, for genes such as CYP2D6 (cytochrome P450 family 2 subfamily D polypeptide 6).

Drug	Therapeutic Area	HUGO Symbol	Referenced Subgroup	Labeling Sections
Amitriptyline	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Precautions
Aripiprazole	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Clinical Pharmacology, Dosage and Administration
Atomoxetine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Dosage and Administration, Warnings and Precautions, Drug Interactions, Clinical Pharmacology
Carbamazepine (1)	Neurology	HLA-B	HLA-B*1502 allele carriers	Boxed Warning, Warnings and Precautions
Carbamazepine (2)	Neurology	HLA-A	HLA-A*3101 allele carriers	Boxed Warning, Warnings and Precautions
Citalopram (1)	Psychiatry	CYP2C19	CYP2C19 poor metabolizers	Drug Interactions, Warnings
Citalopram (2)	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Clobazam	Neurology	CYP2C19	CYP2C19 poor metabolizers	Clinical Pharmacology, Dosage and Administration, Use in Specific Populations
Clomipramine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Clozapine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions, Clinical Pharmacology
Codeine	Anesthesiology/ Analgesia	CYP2D6	CYP2D6 ultra-rapid metabolizers	Boxed Warnings, Warnings and Precautions, Use in Specific Populations, Clinical Pharmacology , Patient Counseling Information
Desipramine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Dextromethorphan and Quinidine	Neurology	CYP2D6	CYP2D6 poor metabolizers	Clinical Pharmacology, Warnings and Precautions, Drug Interactions
Diazepam	Psychiatry	CYP2C19	CYP2C19 poor metabolizers	Drug Interactions, Clinical Pharmacology
Doxepin	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Precautions
Fluoxetine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Warnings, Precautions, Clinical Pharmacology
Fluvoxamine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Galantamine	Neurology	CYP2D6	CYP2D6 poor metabolizers	Special Populations
Iloperidone	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Clinical Pharmacology, Dosage and Administration, Drug Interactions, Specific Populations, Warnings and Precautions
Imipramine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Modafinil	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Nefazodone	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Nortriptyline	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Paroxetine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Clinical Pharmacology, Drug Interactions

Perphenazine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Clinical Pharmacology, Drug Interactions
Phenytoin	Neurology	HLA-B	HLA-B*1502 allele carriers	Warnings
Pimozide	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Warnings, Precautions, Contraindications, Dosage and Administration
Protriptyline	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Precautions
Nefazodone	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Risperidone	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions, Clinical Pharmacology
Tetrabenazine	Neurology	CYP2D6	CYP2D6 poor metabolizers	Dosage and Administration, Warnings, Clinical Pharmacology
Thioridazine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Precautions, Warnings, Contraindications
Tramadol	Analgesic	CYP2D6	CYP2D6 poor metabolizers	Clinical Pharmacology
Trimipramine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions
Valproic Acid (1)	Neurology	POLG	POLG mutation positive	Boxed Warning, Contraindications, Warnings and Precautions
Valproic Acid (2)	Neurology	NAGS, CPS1, ASS1, OTC, ASL, ABL2	Urea cycle enzyme deficient	Contraindications, Warnings and Precautions, Adverse Reactions, Medication Guide
Venlafaxine	Psychiatry	CYP2D6	CYP2D6 poor metabolizers	Drug Interactions

Figure 1-2. Table of Neuropsychiatric Medications

Drug labels for medications used in psychiatry, neurology and anesthesiology from the Table of Pharmacogenomic Biomarkers, U.S. Food and Drug Administration. **[United States FDA 2015]**

Almost all current psychiatric drugs are members of pharmacologic classes discovered and first applied in the 1950s, including medications such as chlorpromazine, imipramine, iproniazid, and lithium **[Hyman 2014]**. There exist off-target effects of these psychotropic drugs whose pathways remain uncharacterized to this day**,** and such drugs are associated with at least half of all ADEs stemming from prescribed medications. Data extrapolated from IMS Health, Inc. as described in **[Hyman 2014]**, show that 15 of the top 30 prescribed medications in 2013 exhibiting pharmacogenomic risk are drugs used in psychiatry and pain management, including sertraline, citalopram, escitalopram, diazepam, tramadol, fluoxetine, codeine, venlafaxine, methylphenidate, paroxetine, amitriptyline, oxycodone, risperidone, aripiprazole, and mirtazapine. **Figure 1-3** lists major medications used in psychiatry in the U.S., and the incidence of serious adverse drug events as evidenced by resulting emergency room visits and hospitalizations from 2009 to 2011 **[Hampton et al 2014]**. Although more detailed information is lacking, it seems unlikely that most of these ~90,000 visits to emergency departments on an annual basis are due solely or primarily to *CYP* gene variations or the other precautionary indications shown in **Figure 1-2**.

It may be that some of this ADE burden, and our difficulty ameliorating it, results from lack of knowledge of mechanisms of lengthy, difficult-to-study emergent adverse events such as extrapyramidal syndromes, weight gain, and actinic keratosis. In addition, because benzodiazepines and opiates have significant consequences in substance abuse and addiction, they are widely considered an "undesirable and difficult" domain for study outside of addiction research.

Such cases are complex and probably involve a number of indications including dose, polypharmaceutical interactions among both psychiatric and non-psychiatric medications, and differential response due to ethnicity, age, gender, and variation in the genome. Data on inheritance of drug response phenotypes strongly suggest that other biological mechanisms, such as epigenomic and transcriptional regulation, are likely to play a significant role and could be applied to informing and improving combinatorial PK metabolizer-class models. Recent work **[Zanger et al 2013, Zanger et al 2014]** has indicated that noncoding regulatory SNPs could play a role in the regulation of PK *CYP* genes. In addition, the relative paucity of PD mechanisms in describing drug efficacy suggests a significant future opportunity for the development of increased understanding of new PD pathways for neuropsychiatric drugs.

	ED Visits, Adverse Drug Events		
Medication Category and Class	**Estimated Annual No. of Visits**	**% Hospitalization Rate**	**Examples of Approved Medications in the U.S.**
Sedatives and anxiolytics			
Short-acting benzodiazepine	14387	24.4	Alprazolam, Clobazam, Clonazepam, Estazolam, Lorazepam, Midazolam, Oxazepam, Temazepam, Triazolam
Non-benzodiazepine	10360	22.1	Buspirone, Butabarbital, Chloral Hydrate, Eszopiclone, Meprobamate, Phenobarbital, Ramelteon, Secobarbital, Zaleplon, Zolpidem
Long-acting benzodiazepine	2633	21.8	Chlordiazepoxide, Amitriptyline/Chlordiazepoxide, Clorazepate, Diazepam, Flurazepam, Quazepam
Miscellaneous	3327	25.5	
Antidepressants			
SSRIs	12019	11.2	Citalopram, Escitalopram. Fluoxetine, Fluoxetine/Olanzapine, Fluvoxamine, Paroxetine, Sertraline
SNRIs	3887	*	Desvenlafaxine, Duloxetine, Milnacipran, Venlafaxine
Triazolopyridines	3443	18.5	Trazodone
Aminoketones	2573	*	Bupropion
Tricyclics	1835	*	Amitriptyline, Amitriptyline/Chlordiazepoxide, Amitriptyline/Perphenazine, Amoxapine, Clomipramine, Doxepin, Imipramine,Nortriptyline, Protriptyline,Trimipramine
Tetracyclics	655	*	Maprotiline, Mirtazapine
MAOIs	*	*	Isocarboxazid, Phenelzine Sulfate, Selegiline, Tranylcypromine
Miscellaneous	838	*	Nefazodone
Antipsychotics			
Atypical	15272	16.9	Aripiprazole, Clozapine, Fluoxetine/Olanzapine, Olanzapine, Paliperidone, Quetiapine, Risperidone, Ziprasidone

Typical	5804	10.6	Chlorpromazine, Fluphenazine, Haloperidol, Haloperidol/Fluphenazine, Loxapine, Perphenazine Amitriptyline/Perphenazine, Thioridazine, Thiothixene, Trifluoperazine
Miscellaneous	*	*	Pimozide
Lithium salts	**3620**	**53.6**	
Stimulants			
Amphetamines	2331	*	Amphetamine, Amphetamine / Dextroamphetamine, Dexmethylphenidate, Lisdexamfetamine Dimesylate, Methamphetamine, Methylphenidate
Miscellaneous	*	*	Armodafinil, Atomoxetine, Modafinil
Two drugs from different psychiatric medication categories	5033	24.2	
Total	**88017**		

Figure 1-3. Table of Neuropsychiatric Medication Classes

Estimation of the incidence of adverse drug events involving psychiatric medications based on annual visits to emergency departments in the U.S. from 2009-2011. *Indicates unknown or statistically unreliable data. *Modified from* **[Hampton et al 2014]**.

For another example, consider the anticoagulant warfarin. Warfarin has a complicated pharmacokinetics and pharmacodynamics, with its influences on the action of a number of clotting factors each exhibiting a different half-life **[FDA 2017]**. It is widely acknowledged that warfarin dosing requirements and other phenotypes exhibit a strong patient-specific genetic element. Factor-of-ten differences in typical dosing requirements based on alleles in *CYP2C9*, the primary metabolic enzyme of warfarin, are present on the package insert label **[FDA 2017]**, and data on association between warfarin requirements and other genetic loci have proliferated. There have been multiple attempts to provide clinicians with genotype-guided dosing algorithms based typically on genotypes of *VKORC1* and *CYP2C9* **[Flockhart et al 2008]**. Despite the strength of variation in these two key genes, there is a "missing heritability" problem: identified loci in these genes account for only 30% to 50% of the predictive power implied by heritability estimates **[Flockhart et al 2008]**. Current generation tests use only variants in these two genes, and almost all use the same three SNPs: rs1799853, rs1057910, and rs9923231 (a promoter SNP for *VKORC1*)

The design of these tests has not benefited from the epigenome or GWAS. These variants were discovered and validated prior to GWAS. Variants discovered in GWAS of warfarin-related phenotypes have not been incorporated into warfarin pharmacogenomics tests, and prior to the experiments described here, regulatory variant discovery with the epigenome had not been applied to the epigenome.

Partially because of this, these tests have achieved only limited success despite intensive development and validation effort. Despite hopes that two large trials (EU-PACT **[Pirmohamed et al 2013]** and COAG **[Kimmel et al 2013]**) would report good results, their appearance in the

same issue of the NEJM did not bear out such hopes. COAG reported no difference in time within the therapeutic range (TTR), and EU-PACT reported a difference, but only compared to fixed starting doses with subsequent adjustment, not to initial dosing methods based on clinical indications that represent the real-world alternative to genetic dosing. Neither trial was powered to report a difference in bleeding and embolism events. Subsequently, a published meta-analysis of nine randomized controlled trials (RCTs) of warfarin pharmacogenomics dosing algorithms vs manual dose determination showed that current-generation tests using *VKORC1* and *CYP2C9* offer no improvement in TTR, percentage of patients with high INR, or bleeding and coagulation events **[Stergiopoulos et al 2014]**.

Since the conclusion of these trials, expert comment has indicated a consensus that such algorithms do not add clinical value relative to dosing on the basis of clinical indications, despite their predictive power **[Kimmel et al 2015, Johnson et al 2016]**. Although the recent GIFT RCT of genomic warfarin dosing found that genome-guided dosing provided a significant benefit in the composite endpoint of major bleeding, INR of 4 or greater, venous thromboembolism, or death **[Gage et al 2017]**, this generalizability of the study's results is limited. GIFT had narrow inclusion criteria: patients aged 65 years or older initiating warfarin for elective hip or knee arthroplasty. Those patients are at higher risk, and thus the generalizability of these results to larger patient populations and indications is debatable.

Despite their predictive power, such tests have failed to deploy in mainstream clinical practice. Accordingly, it is clear that the failure of the warfarin pharmacogenomics community to benefit from the GWAS studies it had conducted and interpret them with the epigenome was not benign.

There is a critical need for the pharmacogenomics community to begin using GWAS and advanced epigenomic methods in the design of pharmacogenomics tests and decision support products.

Chapter 2: Manual Experiments in Epigenome Variant Discovery

Pharmacoepigenomics in Application

Motivated by the incomplete predictive power of current psychotropic pharmacogenomics tests, and the dominance of coding variants in both existing tests and ongoing work, and the potential for noncoding variants to contribute to resolving this problem, the Athey lab in 2014 initiated a series of variant discovery experiments in neuropsychiatric drug-disease systems, initially falling under the aegis of its partnership with Assurex Health. By uniting association data with consortium omics and preexisting expert knowledge, we hoped to discover new noncoding variants for these phenotypes and their target genes, to enable development of second and subsequent generation neuroyschiatric pharmacogenomics tests.

This chapter will cover three semi-manual experiments undertaken under the auspices of this initiative. They are:

1. An exploration of a number of GWAS results in neuropsychiatric pharmacogenomics which had not yielded coding variants **[Higgins et al 2015]**
2. A comprehensive survey of association studies in lithium pharmacogenomics**[Higgins et al 2015]**

3. A mechanistic study of the action of valproic acid **[Higgins et al 2017]**.

All three experiments were undertaken under the lead authorship of Gerry Higgins, and involved a number of people, principally including myself and Brian Athey.

In the process of developing the methods for these exploratory experiments, carrying them out, and analyzing the results, a number of principles emerged, codified under the orienting framework I later began to call the Five Box Model of regulatory variant discovery. These principles and ideas formed the basis for the Pharmacoepigenomics Informatics Pipeline, which was designed to carry out such experiments in a more automated, more reproducible manner and with advanced features which would not be plausible in a semi-manual environment.

Epigenomic Mapping of Noncoding Psychotropic Variants

Motivation

Joint GWAS analyses of related disease phenotypes are now commonly accepted, as the evidence has mounted that overlapping sets of variants for related phenotypes are the norm, not the exception **[McGeachie et al 2014]**. In addition, joint analysis of GWAS in psychiatric disorders using pathway analysis software has implicated shared, common pathways among multiple disorders such as bipolar disorder, major depressive disorder, and schizophrenia **[Sullivan et al 2012]**. Pathway analysis from multiple GWAS studies has been applied to recognize regulatory networks shared among common complex diseases including coronary artery disease (CAD), rheumatoid arthritis (RA), Type 1 Diabetes (T1D), Type 2 Diabetes (T2D) and bipolar disorder **[Cross-Disorder Group 2013]**.

Fine mapping of associated loci in GWAS yield causal variants that may differ from a reported lead SNP and are typically located in noncoding regions of the genome. For example, epigenetic mapping of enhancers using the histone marks H3 lysine27 acetylation (H3K27ac) and H3 lysine 4 monomethylation (H3K4me1) in concert with mRNA-seq of neighboring genes was able to identify causal SNPs in enhancers in autoimmune disease, in which only 14% of associated protein-coding SNPs appear to be causal, while 60% of known causal SNPs map to enhancers **[Farh et al 2015]**. The relative scarcity of missense coding SNPs in GWAS of disease risk and other human traits can now be simply explained: regulatory DNA harbors causal SNPs in addition to protein coding domains **[Kellis et al 2014].**

An ongoing challenge in pharmacogenomics is that many medications used in psychiatry and pain are clinically suboptimal and exhibit high frequencies of serious adverse events. In pain management, opioids are highly addictive but very effective, so it would be better to develop analgesics without risk of abuse. However, the model compounds for almost all neuropsychiatric drugs were discovered in the 1950s or earlier, and included opiates, as well as chlorpromazine, imipramine, iproniazid, and lithium. Since many medications used to treat patients with neuropsychiatric disorders were discovered over 50 years ago and/or serve as antecedents to later refinements of these seminal mid-century formularies, it is clear that greater emphasis should be placed on pharmacodynamic targets in the human CNS **[Hyman 2014]**.

In this study, we developed a multi-level mapping method to identify novel noncoding SNPs in neuropsychiatry that exhibit epigenomic characteristics of genomic regulatory factors, and applied this method to screen SNPs derived from 26 pharmacogenomics association studies sourced from the NHGRI GWAS catalogue **[Welter et al 2014]** and 3 published candidate gene association studies. We used multiple known attributes of noncoding SNPs corresponding to genomic regulatory elements such as enhancers, promoters, and transcribed domains using extant public resources. These included public databases of chromatin interaction mapping for prediction of *cis*- and *trans*-regulation, including Hi-C and ChIA-PET data **[Dixon et al 2012, Li et al 2013, Li et al 2014]**, to predict putative spatial interactions at the allelic level. These interactions were also investigated for legitimacy based on "guilt-by-association" methods **[Lee et al 2011]**, leveraging multiple GWAS using pathway analysis **[Califano et al 2012, McGeachie et al 2014]**, and GWAS imputation and annotation methods published by the epigenome roadmap consortium **[Farh et al**

2015, Yao et al 2015, Roadmap Epigenomics Consortium 2015, Zhou et al 2015, Leung et al 2015, Ernst et al 2015]. At every phase of the experimental design, we used multiple independent analytic methods to ensure the validity of our *in silico* results using the current knowledge base.

Methods

An overview of methodology is shown in **Figure 2-1**. We used parallel approaches for SNP imputation, determination of functionally-significant variants for a given pharmacogenomic association, annotation of SNPs and assignment to a genome regulatory element such as enhancer, promoter and transcribed element, prediction of putative functional interactions in cis- and trans- using chromatin interaction mapping data and novel pathway reconstruction using multiple methods, including assignment to the most plausible pathway in terms of known functional, pharmacologic and/or morphologic data. Most results were independently gathered by 2 investigators who were often not aware of the results obtained by the other. Since this study was performed *in silico*, several different, but sometimes related, methods were used to determine whether the results were the same. The objective of this study was to provide evidence for subsequent experimental studies for scientific validation in cell lines, animal models and other biological systems.

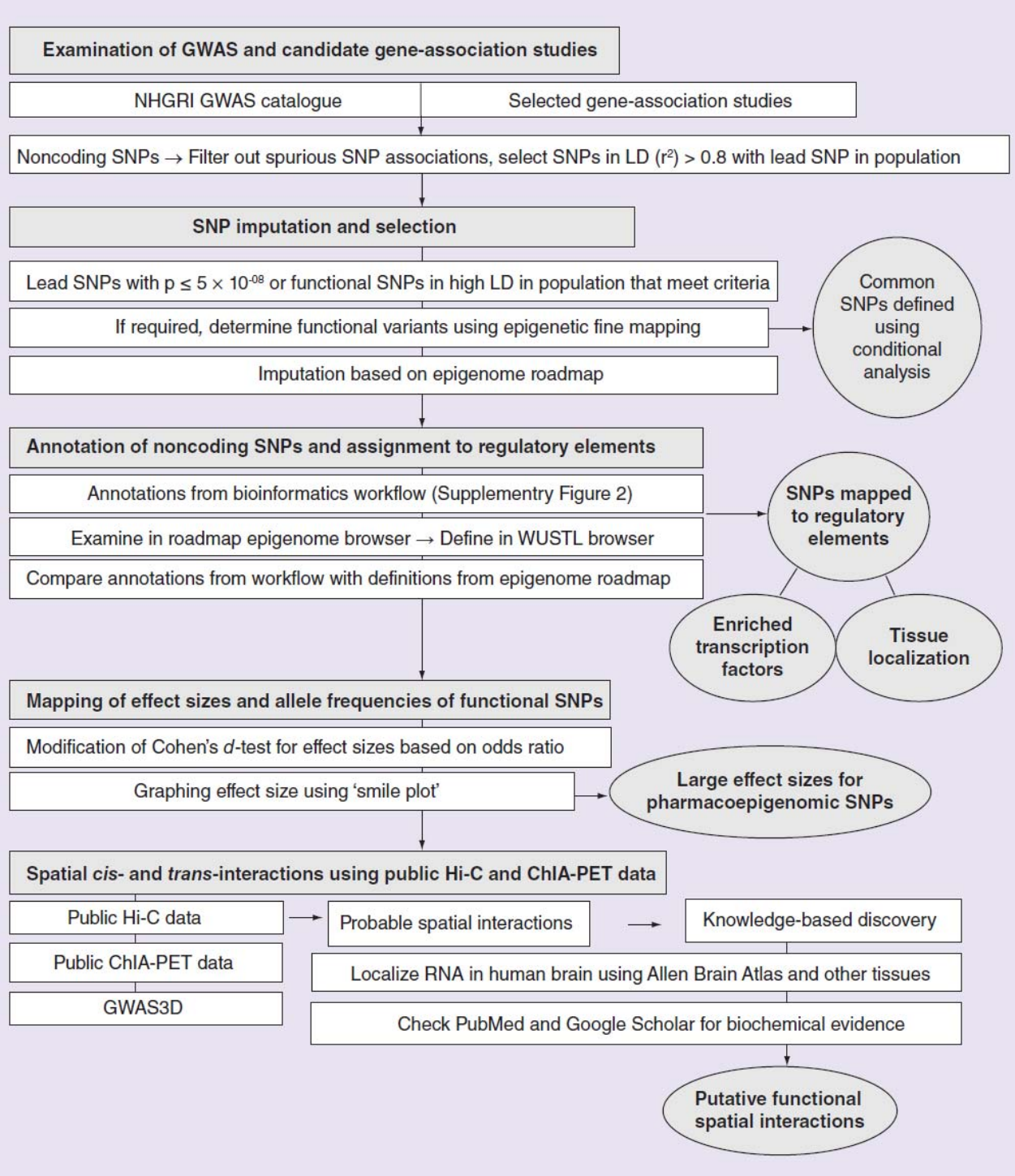


Figure 2-1. Diagram of 'Variants' Analysis Methods

Examination of GWAS and Candidate Gene Association Studies

We examined neuropsychiatric pharmacogenomics GWAS contained in the NHGRI GWAS catalogue **[Welter et al 2014]** in psychiatry, neurology, analgesia and addiction, herein referred to collectively as "neuropsychiatric." These included 26 GWAS and 3 candidate gene association studies. The candidate gene studies were chosen for analysis because they were large case-control studies performed in circumscribed human populations that had identified significant variants associated with lithium response, citalopram response, and heroin addiction, to demonstrate proof-of-concept. The gene association studies used for the analysis are shown in **Figure 2-2**. 31 variants from the 26 GWAS and 3 candidate gene association studies were initially selected that reached the criterion of a pharmacoepigenomic SNP from a total of 2024 variants in linkage disequilibrium (LD) > 0.8 (r^2) with the reported lead SNP within the population being measured. 94% (1903) of the 2024 SNPs were noncoding and 6% (121) were coding.

SNP Imputation and Selection

In certain cases, standard methods were used for dense genotyping and imputation from association loci that had not been performed in the original study, including removing spurious associations and checking population stratification **[Farh et al 2015, Wellcome Trust Case Control Consortium 2007]**. In GWAS, if lead SNPs met the stringent statistical criterion of $p \leq 5 \times 10^{-08}$, then the SNP(s) were checked for epigenomic classification using both our workflow and the reference human epigenome reference **[Roadmap Epigenomice Consortium 2015]**. In all but 1 case, the 31 SNPs analyzed using our workflow and using the human reference epigenome

produced the same results for the putative active regulatory elements that we identified. For SNPs that were associated with mutations in an enhancer that controlled a subnetwork responsible for drug efficacy and adverse events, then these variants were prioritized for further analysis. All gene definitions used in this study were acquired from the HUGO database **[Gray et al 2012]**. Genes were mapped to UCSC hg19 coordinates **[Karolchik et al 2014]**. Variation data from dbSNP **[Sherry et al 2001]** were also mapped to UCSC hg19 coordinates, and linkage-disequilibrium (LD) data for SNP pairs within population-specific haplotypes was downloaded from HaploReg **[Ward et al 2012]**. After compiling a master spreadsheet containing pharmacoepigenomic SNPs as output our workflow, we compared the results with those using the roadmap epigenome browser **[Zhou et al 2015]**. Although this study was only focused on what may be functionally active genome regulatory elements such as enhancers, promoters and transcribed domains, the epigenome roadmap classification system includes 15 chromatin states. Nonetheless, for every variant that we characterized as a promoter, enhancer, transcribed domain or quiescent, 30 out of 31 closely correspond to the chromatin state annotated from the epigenome roadmap. These data are presented in the results section.

In cases wherein dense genotyping and epigenetic fine mapping had not been performed in the original GWAS analysis, we used a simple derivation of the method used in **[Farh et al 2015]** but with several modifications. We did not restrict GWAS selection to only those studies that contained SNPs which met the rigid statistical criterion of $p < 5 \times 10^{-08}$. The rationale was that, as has been shown by others, SNPs that appear to be less significant may in fact contribute function or be casual, based on pathway analysis showing that the SNP may impact a genome regulatory elements such as an enhancer **[Onengut-Gumuscu et al 2015, Farh et al 2015, Maurano et al**

2012]. Distal H3K27ac peaks were assigned to their potential target genes if they were located within introns or within 500kb regions upstream of a transcription start site of a gene in which likely a target of an enhancer. It is presumed that within an associated locus, a linear trend should be observed such that where a causal variant has been demonstrated using functional studies, with the assumption that neutral SNPs demonstrate association signal in proportion to their LD to the causal variant. However, in GWAS, multiple causal SNPs that effect regulatory pathways may occur within a haplotype block with an LD (r^2) > 0.8 in the same population **[Farh et al 2015]**. Conversely, a given locus may harbor causal SNPs associated with more than one phenotype, in what is referred to as pleiotropy. Thus, we did not discard what may have been considered neutral SNPs in the context of a pharmacogenomics association, but instead applied both statistical and knowledge-based methods to prioritize SNPs and identify them as functional, but not causal *in lieu* of biological validation. If a reported lead SNP in an association study associated with a psychotropic drug response but was not correlated with a genome regulatory element expressed in human brain, and was not associated with pharmacokinetic response, we examined all SNPs with a LD >0.8 in the haplotype of that population to determine if there were any in tight LD that exhibited correlation with human brain. This contributed to reprioritization of the variants in the LD block based on neuroanatomical localization.

Associations	Study	Population	PMID
Habitual coffee consumption and caffeine addiction	Genome-wide association analysis of coffee drinking suggests association with CYP1A1/CYP1A2 and NRCAM	Caucasians	21876539
	Sequence variants at CYP1A1 – CYP1A2 and AHR associate with coffee consumption	European and American	21357676
	Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption	European	21490707
Number of cigarettes smoked per day and nicotine dependence	Haplotypes with copy number and single nucleotide polymorphisms in CYP2A6 locus are associated with smoking quantity in a Japanese population	Japanese	23049750
	Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior	European	20418888
	Genome-wide meta-analyses identify multiple loci associated with smoking behavior	European	20418890
	CHRNB3 is more strongly associated with Fagerström test for cigarette dependence-based nicotine dependence than cigarettes per day: phenotype definition changes genome-wide association studies results	Mixed	22524403
Opioid dependence	Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways	African-American	24143882
	NCK2 is significantly associated with opiate addiction in African-origin men	African-American males	23533358
Heroin addiction	Association of OPRD1 polymorphisms with heroin dependence in a large case-control series	Mixed American	22500942*
Cocaine dependence	Genome-wide association study of cocaine dependence and related traits	Mixed American	23958962
Alcohol consumption and dependence	Genome-wide association studies identify genetic loci related to alcohol consumption in Korean men	Korean	21270382
	Genome-wide association study identifies two loci strongly affecting transferrin glycosylation	European ancestry	21665994
Lithium response in bipolar disorder	Influence of an interaction between lithium salts and a functional polymorphism in SLC1A2 on the history of illness in bipolar disorder	European	23023733*
	A genomewide association study of response to lithium for prevention of recurrence in bipolar disorder	European ancestry	19448189
Adverse drug events and antipsychotic drugs	Genome-wide association study of antipsychotic-induced QTc interval prolongation	European	20921969
	Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs	American	20195266
	Genomewide association study of movement-related	Mixed	19875103

	adverse antipsychotic effects	American	
Carbamazepine-induced cutaneous adverse drug reactions	Genome-wide association study identifies HLA-A*3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population	Japanese	21149285
	HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans	European	21428769
Individual variability in post-surgical anesthesia	Genome-wide association study of acute post-surgical pain in humans	European	19207018
	Genome-wide association study identifies a potent locus associated with human opioid sensitivity	Japanese females	23183491
Antidepressant response and remission in depression	Citalopram and escitalopram plasma drug and metabolite concentrations: genome-wide associations	European	24528284
	Genome-wide pharmacogenetics of antidepressant response in the GENDEP project	European	20360315
	A genome-wide association study of a sustained pattern of antidepressant response	European	23726668
	FKBP5 genetic variation: association with selective serotonin reuptake inhibitor treatment outcomes in major depressive disorder	European	23324805*
	A genomewide association study of citalopram response in major depressive disorder	Mixed	19846067
Lamotrigine-induced hypersensitivity	Genome-wide mapping for clinically relevant predictors of lamotrigine- and phenytoin-induced hypersensitivity reactions	European	22379998

Figure 2-2. GWAS Studies Used in 'Variants' Analysis

Selected GWAS and candidate gene pharmacogenomic association studies in psychiatry, neurology, addiction medicine and pain medicine used for identification of pharmacoepigenomic variants using the method in this study. PMID: PubMed Identification accession number.

Annotation of Pharmacoepigenomic SNPs and Assignment to Regulatory Elements

As part of our workflow**,** we assigned SNPs to regulatory elements combining data from chromatin conformation capture, correlation with CTCF and cohesion (RAD21, SMC3) transcription factors associated with chromatin loop boundaries, chromatin state annotation, DNase I hypersensitivity, hypomethylation, anatomical localization, and biochronicity. The method focused on variants, including SNPs associated with drug response and adverse drug events, which are used to identify genomic regulatory elements that are expressed in human brain or the liver in the case of a CYP gene, or regulate gene transcription, and are located in noncoding regions of the human genome. The workflow is comprised of several components, including determination of whether an epigenome variant may be functional, followed by its expression in brain (pharmacodynamic, PK) or liver (pharmacokinetic, PD).

Input variants are screened by ethnic population, followed by mapping of functional SNPs to determine the causal nature of a given SNP using methods based on the assumptions that, in the majority of GWAS, it is necessary to discriminate SNPs that represent the most likely functional variants from those with strong but significantly less association to the drug response trait, so that the majority of SNPs at a given locus can be excluded from further analysis. As a consequence of tight LD that exists within a haplotype or haplotypes associated with a trait, the 'lead' SNP(s) as reported may or may or may not attain the threshold of significance used in GWAS and may not necessarily the best candidate to serve as the functional or 'causal' SNP.

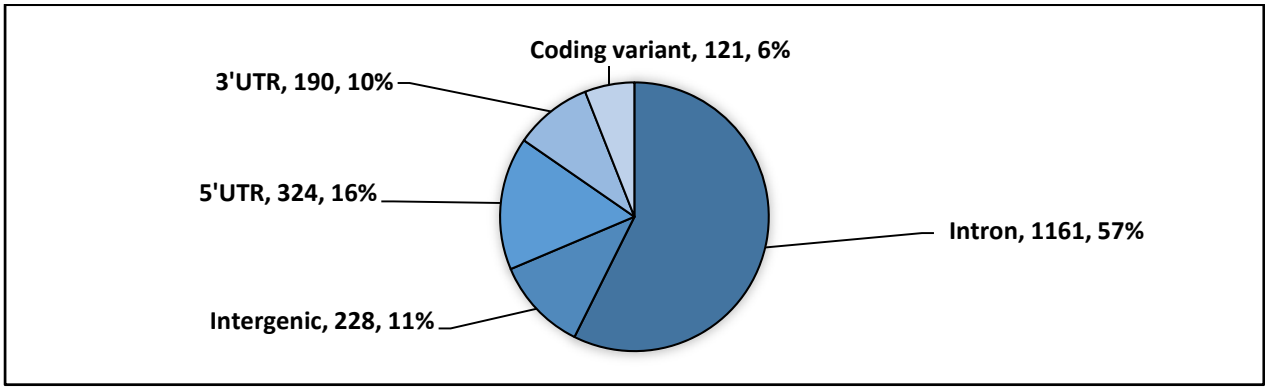
The steps of the bioinformatics pipeline workflow in were developed independently of the epigenome roadmap consortium. The method is optimized for association studies in pharmacogenomics in psychiatry, addiction medicine, pain medicine, and neurology, herein referred to collectively as "psychotropic" variants. The method focuses on variants, including SNPs associated with drug response and adverse drug events, which are used to identify genomic regulatory elements expressed in human brain or the liver in the case of a cytochrome P450 (*CYP*) gene, or to regulate gene transcription in pharmacodynamic pathways in human brain. RegulomeDB **[Boyle et al 2012]** provides an initial assessment as to what role a given variant might play in gene regulation. Variants are evaluated as to: 1) whether they were enriched as quantitative trait loci (eQTLs, meQTLs) as determined using the scoring method of GeneVar **[Yang et al 2010]** and/or GTEx **[Lonsdale et al 2013]**; 2) whether the variant was located in open chromatin in the tissue of interest (e.g., brain, neuronal cell line, liver, immune cell, HepG2, heart or other); 3) whether the variant was hypomethylated; and 4) what specific histone marks were associated with the variant that defines its assignment as a promoter, enhancer or transcribed domain to impute chromatin state **[Ernst et al 2015, Ernst et al 2012]**.

We also determined the degree to which a variant altered the strength, number and type of binding sites for transcription factors, and which transcription factors were bound to the variant to understand the function of the corresponding genomic regulatory element. These were determined using several public resources including HaploReg 3.0 **[Ward et al 2012]** and manual inspection of the variant in the UCSC **[Karolchik et al 2014]** and epigenome roadmap **[Zhou et al 2015]** browsers. Multi-level mapping in chromatin interaction networks for variants located in human brain help provide evidence for probable functional pathways. In neuropsychiatric

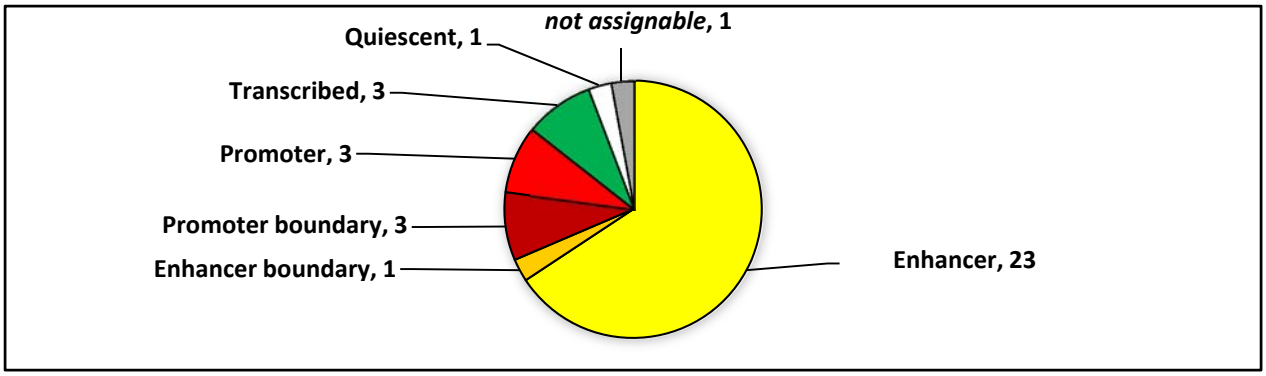
pharmacogenomics, it is also critical to evaluate whether annotated regulatory elements and their target genes exhibit biochronicity, a general feature of pharmacogenomic variation **[Zhang et al 2014]**. Results shown in **2-5**. In the workflow analysis, variants are assessed as to 1) whether they are enriched as quantitative trait loci (eQTLs, meQTLs); 2) whether the variant is located in open chromatin; 3) whether the variant is hypomethylated; 4) what specific histone marks are associated with the variant; 5) the degree to which a variant alters the strength, number and type of binding sites for transcription factors; (6) which transcription factors are bound to the variant; 7) what interactions and/or contacts the variant makes with other genomic regulatory elements, such as transcription start sties (TSS) in chromatin networks; 8) the location of the variant in brain regions or liver that are consistent with the function of the variant; and 9) whether the variant exhibits biochronicity, a general feature of pharmacoepigenomic variation. These assessments can be performed using a variety of public bioinformatics resources. Other attributes may also be important, including evolutionary conservation as determined using GERP++, or whether a variant is within a human accelerated region of the genome. Biochronicity checking using published literature resources provides additional detail to the annotation model to check whether a given variant and its corresponding genomic regulatory element exhibit circadian, ultradian or seasonal rhythmicity. These signs are indicative of dysregulated core symptoms in the clinical context of psychiatric disorders such as bipolar disorder and major depressive disorder, and can be checked using the knowledge update engine. The impact, if any, of the glucocorticoid receptor (GR) on the variant is also checked, because it mediates the stress response and is the primary driver of circadian and ultradian rhythmicity. The input variant is also checked to determine whether it is a member of the *CLOCK* gene family or a closely related gene.

We further examined the output of assignments from our workflow and compared them to those made by the epigenome roadmap consortium. This involved examination of the SNPs using the roadmap epigenome browser **[Zhou et al 2015]**, followed by definition of regulatory elements using the WUSTL genome browser **[Zhou et al 2011]**. We then compared these results with the results shown in **Figure 2-3.**

A



B



KEY – CLASSES OF REGULATORY ELEMENTS

Epigenome roadmap consortium		**Workflow**		**ENCODE**	
Enhancer		Enhancer		Enhancer	
Genic Enhancer		Enhancer boundary, including CTCF, RAD21 and SMC3			
Promoter Active Transcriptional Start Site Flanking Transcriptional Start Site		Promoter		Promoter	
		Promoter boundary, including CTCF and RAD21			
Strong Transcription		Transcribed domain		Translational Elongation	
Weak Transcription					
Quiescent		Quiescent		Quiescent	

C

Type	SNP	Association [PMID]	Workflow	Roadmap	Consensus Assignment	
ADDICTION	rs2470893	Caffeine dependence [21876539]			Enhancer	Liver
	rs2472297	Coffee consumption [21357676]			Enhancer	Liver
	rs8192725	Nicotine dependence [23049750]			Promoter boundary	Liver
	rs11878604	Nicotine dependence [23049750]			Transcription	Liver
	rs11083595	Nicotine dependence [20418890]			Enhancer	Liver
	rs1329650	Nicotine dependence [20418890]			Enhancer	Brain
	rs13280604	Nicotine dependence [22524403]			Enhancer	Brain
	rs2072134	Alcohol consumption [21270382]			Enhancer / Transcription	Brain
	rs10849915	Alcohol consumption [21270382]			Quiescent	Brain
	rs2074356	Alcohol consumption [21270382]			Transcription	Brain
	rs3811647	Alcohol dependence [21665994]			Enhancer	Brain
	rs12053259	Opiate addiction [23533358]			Enhancer	Brain
	rs60349741	Opiate addiction [24143882]*			Enhancer	Brain
	rs2236855	Heroin addiction [22500942]*			Enhancer	Brain
	rs2629540	Cocaine–induced euphoria [23958962]			Enhancer	Brain
ANALGESIA	rs2551941	Post-surgical opiate analgesia [23183491]			Promoter boundary	Brain
	rs2709394	Post-surgical opiate analgesia [23183491]			Promoter boundary	Brain
	rs2650805	Post-surgical NSAID analgesia [19207018]			Enhancer boundary	Brain
RUG RESPONSE	rs4354668	Lithium response [23023733]*			Promoter	Brain
	rs7816924	Antidepressant response [23726668]			Enhancer	Brain
	rs4986894	Antidepressant response [24528284]*			Promoter	Liver
	rs1080989	Antidepressant response [24528284]*			*not assignable*	

	rs2486416	Antidepressant response [19207018]			Enhancer	Brain
DVERSE EVENTS	rs6741819	Antipsychotic drug-induced cardiac risk [20195266]			Enhancer	Brain
					Enhancer	Heart
	rs79535779	Antipsychotic drug-triglycerides [20195266]			Enhancer	Brain
					Enhancer	Pancreas
	rs4959235	Antipsychotic drug-QT prolongation [20921969]			Enhancer	Brain
	rs77484422	Antipsychotic drug-induced dyskinesia [19875103]			Promoter	Brain
RGAN INJURY	rs2844796#	Carbamazepine-induced adverse events [21149285]			Enhancer	CD34 cells
	rs1633021#	Carbamazepine-induced adverse events [21149285]			Enhancer	CD34 cells
	rs1061235#	Carbamazepine-induced adverse events [21428769]			Enhancer	CD14 cells
	rs183266#	Lamotrigine-induced hypersensitivity [22379998]			Enhancer	CD34 cells

Figure 2-3. Epigenome Context of Reported Variants

[A] Pharmacogenomic SNPs from GWAS in neuropsychiatry map to noncoding regions of the genome. **[B, C]** Assignment of regulatory elements based on GWAS and selected gene association studies* to enhancer, promoter, transcribed and quiescent domains using chromatin state annotation based on agreement between epigenome roadmap browser scores **[Zhou et al 2015]** and our workflow assignments. In the key below, colors are based on the epigenome roadmap browser **[Roadmap Epigenomics Consortium 2015, Zhou et al 2015, Ernst et al 2015],** and those of the same shade indicate the same class of regulatory elements, and includes assignments from the ENCODE project **[ENCODE Project Consortium 2012, Ernst et al 2012]**. The

workflow further assigns a relationship to CTCF and cohesin subunits (RAD21, SMC3) for boundary definition. Epigenome roadmap scores are based on Histone H3 Lys (K) for the detected regulatory element **[Zhou et al 2015]**.For brain, the region with the highest roadmap score for H3K27ac is usually indicated, but in some cases, the regulatory element may different between brain regions. # Tagging SNPs for *HLA* alleles.

Mapping of Effect Sizes and Allele Frequencies for Functional Variants

We calculated the effect size and associated allele frequency for a number of functional pharmacogenomic SNPs (those from studies reporting odds ratios), as well as a background set of SNPs associated with neuropsychiatric diseases. Minor allele frequencies were taken directly from published GWAS studies. Effect sizes were calculated using SNP frequencies and odds ratios using a variant on Cohen's D-test **[Cohen 1992]**. These were plotted on a two dimensional "Smile Plot" graph **[Park et al 2010, Willer et al 2013, Lange et al 2015]; Figure 2-6**. The plot also includes statistical power curves for reference. Power curves demarcating the lower bound on the effect size and frequency of SNPs detectable in a GWAS can be derived using the 'pwr' package in R. **[http://cran.r-project.org/web/packages/pwr/pwr.pdf]** For reference we include the power curves of two hypothetical GWAS surveys, which would be only barely able to detect the median background SNP, for one curve, and the most powerful such SNP, for the second curve. SNPs above a given power curve are more significant, in a statistical sense, than those below and those on the curve. This plot provides a visual illustration of the comparative significance (in therapeutic terms) of a collection of SNPs based on their combined effect size and associated allele frequency.

Spatial Cis- and Trans-Interactions Using Public Hi-C and ChIA-PET Data

For prediction of cis-interactions of putative enhancer, we examined genes that fell within the purview of known enhancers identified by the ENCODE and Epigenome Roadmap consortia, and used knowledge-based prediction to identify those genes most likely to contribute to the drug

response phenotype under evaluation. These were then plotted as shown in **Figure 2-7**. A combination of known enhancer lengths are shown above each of this subset of examples, with potential TAD boundaries represented as shown based on correlation of the SNP to boundary factors such as CTCF and cohesin (i.e., RAD21 and SMC3). We used several databases of Hi-C and ChIA-PET in order to identify enhancer-target gene interactions in *cis* and *trans*. These included public Hi-C datasets and ChIA-PET data visualized in the UCSC genome browser **[Dixon et al 2012, Rosenbloom et al 2010, Karolchik et al 2014]**. In addition, we used the GWAS3D website **[Li et al 2013]** to confirm some of the spatial interactions found in the primary datasets. The 3DGD database of Hi-C interactions in certain human cell lines was downloaded and provided additional value for discrimination of spatial interactions **[Li et al 2014]**. For *cis* regulation of gene promoters by enhancers, we used the most probable targets based on a "guilt-by-association" method using known biochemical pathways and QTL relationships mined from published literature. In many cases of putative *trans* interactions, target regions of the genome contained no known genes or lincRNAs whose specific function had not yet been understood. The Ensembl genome bowser **[Flicek et al 2013]** was used to evaluate enhancer targets because they provide the most comprehensive documentation of lincRNAs, as well as extensive tracks for predicted genes.

Results

Functional SNPs Map to Predicted Pharmacoepigenomic Enhancers

In this the study, all of the predicted functional SNPs mapped to introns or intergenic regions, followed by 5'UTRs and 3'UTRs. **Figure 2-3 A** shows an overview of the distribution of SNPs by genome region. Many intronic SNPs were associated with enhancers that did not apparently regulate the gene in which they were located. Next we compared regulatory element annotation from the epigenome roadmap with the results of our workflow **(Figure 2-3 B, C)**. The assignments between the 2 independent methods was high, with concordance in regulatory element class in different adult human tissues for 34 out of 35 assignments. Comparison with ENCODE cell line annotations also showed 97% agreement with our workflow *(data not shown)*. Since we were including association with CTCF and cohesin (subunits RAD21 and SMC3) to define location within enhancers and promoters near presumptive borders between loop domains. As seen in **Figure 2-3 B**, tissue-specific assignments included 23 enhancers, 1 enhancer boundary, 3 promoters, 3 promoter boundaries, 3 transcribed domains, 1 quiescent domain and 1 which was not assignable.

Enrichment of Transcription Factors at Putative Enhancer, Promoters and Transcribed Domains

Previous research has shown that TFs and DNA binding proteins bind in a sequence specific manner. Since the majority of our predicted functional SNPs are located in regulatory sequences in the relevant tissues (pharmacodynamics genes in brain and pharmacokinetic genes in liver), we determined whether selected DNA-binding proteins such EP300 were associated with enhancers, YY1 was associated with promoters, and the splicing factor TCERG1 was associated with transcribed domains. We selected those transcription factors associated with various regulatory

elements and tissue type. **Figure 2-4** plots some of the known transcription factors and other DNA-binding proteins whose binding was predicted to be altered by the SNPs, bound using ChIP-seq, or both, and were indicative of assignment to specific classes of regulatory elements. Statistical analysis showed that these specific transcription factors and DNA-binding proteins were associated with the consensus assignment to class of regulatory element ($p<0.01$; ANOVA).

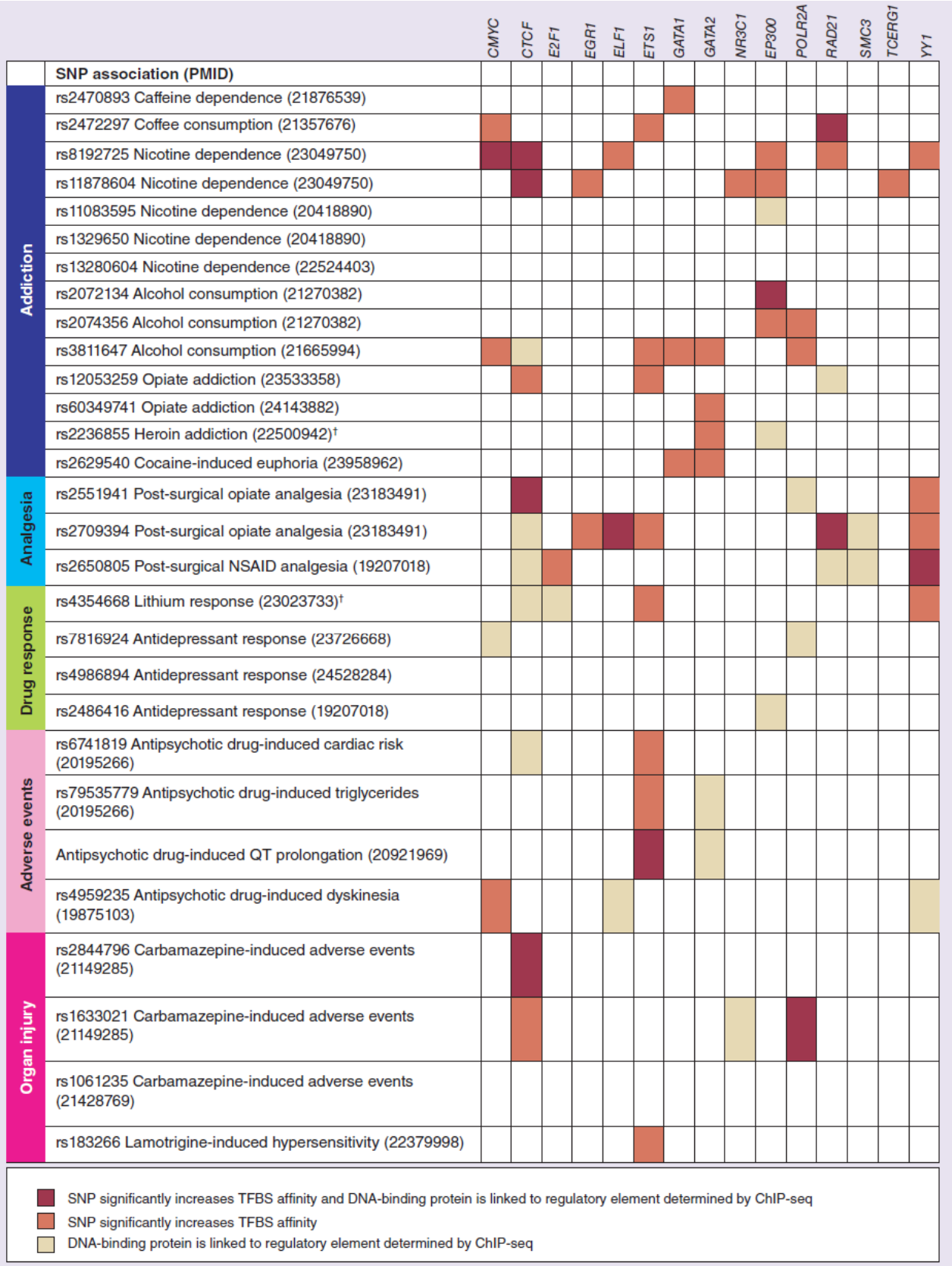

CMYC
CTCF
E2F1
EGR1
ELF1
ETS1
GATA1
GATA2
NR3C1
EP300
POLR2A
RAD21
SMC3
TCERG1
YY1
SNP association (PMID)
Addiction
rs2470893 Caffeine dependence (21876539)
rs2472297 Coffee consumption (21357676)
rs8192725 Nicotine dependence (23049750)
rs11878604 Nicotine dependence (23049750)
rs11083595 Nicotine dependence (20418890)
rs1329650 Nicotine dependence (20418890)
rs13280604 Nicotine dependence (22524403)
rs2072134 Alcohol consumption (21270382)
rs2074356 Alcohol consumption (21270382)
rs3811647 Alcohol consumption (21665994)
rs12053259 Opiate addiction (23533358)
rs60349741 Opiate addiction (24143882)
rs2236855 Heroin addiction (22500942)†
rs2629540 Cocaine-induced euphoria (23958962)
Analgesia
rs2551941 Post-surgical opiate analgesia (23183491)
rs2709394 Post-surgical opiate analgesia (23183491)
rs2650805 Post-surgical NSAID analgesia (19207018)
Drug response
rs4354668 Lithium response (23023733)†
rs7816924 Antidepressant response (23726668)
rs4986894 Antidepressant response (24528284)
rs2486416 Antidepressant response (19207018)
Adverse events
rs6741819 Antipsychotic drug-induced cardiac risk (20195266)
rs79535779 Antipsychotic drug-induced triglycerides (20195266)
Antipsychotic drug-induced QT prolongation (20921969)
rs4959235 Antipsychotic drug-induced dyskinesia (19875103)
Organ injury
rs2844796 Carbamazepine-induced adverse events (21149285)
rs1633021 Carbamazepine-induced adverse events (21149285)
rs1061235 Carbamazepine-induced adverse events (21428769)
rs183266 Lamotrigine-induced hypersensitivity (22379998)
SNP significantly increases TFBS affinity and DNA-binding protein is linked to regulatory element determined by ChIP-seq
SNP significantly increases TFBS affinity
DNA-binding protein is linked to regulatory element determined by ChIP-seq

Figure 2-4. Transcription Factor Context of Reported Variants

Association of transcriptional factors and other DNA-binding proteins help define the domain occupied by different regulatory elements that could be unequivocally assigned and are active. CTCF and cohesion subunits (RAD21, SMC3) are significantly associated with enhancer and promoter boundary domains. CYMC, E2F1, ELF1, ETS1, GATA1, GATA2 and EP300 are associated with enhancers. EGR1 is associated with estrogen regulation, and NR3C1 association is suggestive of glucocorticoid regulation. TCEGR1 is a transcriptional elongation factor associated with differential splicing. YY1 is associated with promoters.

Tissue-Specific Localization of Regulatory Elements

Tissue localization of regulatory elements were mapped based on data from the epigenome roadmap consortia **[Zhou et al 2015]**. Localization based on chromatin state annotation was in most cases consistent with organs involved in GWAS phenotypes, with HLA allele SNPs the major exception. The majority of regulatory elements were localized in liver and brain, consistent with a role in pharmacokinetic and pharmacodynamics pathways respectively. Also, in GWAS of adverse events from antipsychotic medications, the pattern of expression is consistent with affected organs. These include perphenazine-induced cardiac risk in which the enhancer is expressed in brain and heart, and clozapine-induced elevated triglycerides, wherein the impacted enhancer is expressed in brain and pancreas. In carbamazepine- and lamotrigine-induced adverse events, the predicted enhancer was localized in CD34 or CD14 cells. Although very specific liver and brain-specific transcription factor clusters have been described **[Leung et al 2015]**, we only were able to characterize 3 examples in which tissue-specific transcription clusters could be assigned.

Large Effect Sizes for SNPs Associated with Addiction, Analgesia and Drug-Induced Injury

In contrast to the prevailing understanding that complex traits are predominantly influenced by many variants with small effects, we found that causal SNPs in neuropsychiatric pharmacogenomics GWAS typically exhibit large effect sizes (**Figure 2-5**). **Figure 2-6** shows a "smile plot" comparing allele frequency from 0 to 1 versus effect size calculated using a revision of Cohen's *d*-test based on odds ratio **[Cohen 1992, Park et al 2010, Willer et al 2013, Lange et al 2015, Kato et al 2011]**. As can be seen from the plot, functional pharmacoepigenomic SNPs

associated with caffeine, nicotine, alcohol and opiate addiction are common within human populations and exhibit large effect sizes. Similarly, HLA allele tagging SNPs for carbamazepine and lamotrigine, medications used to treat bipolar disorder, exhibit very large effect sizes but exhibit minor allele frequencies. SNPs associated with lithium response and antipsychotic drug-induced cardiac risk are also common and have moderate effect sizes. Effect sizes for functional SNPs reported in psychotropic pharmacogenomic association studies were substantially larger than those reported in GWAS examining disease-risk variants for psychiatric and neurological disorders.

Effect size	Allele frequency	Phenotype	SNP	PMID
0.371992921	0.4011	Caffeine dependence in EUR	rs2470893	21876539
0.48155789	0.23	Caffeine dependence in EUR	rs2472297	21357676
0.560900928	0.2472	Nicotine dependence in JPT	rs11878604	23049750
0.27667143	0.69	Nicotine dependence in AMR	rs13280604	22524403
0.405315116	0.2706	Nicotine dependence in AMR	rs11083595	20418890
0.722835812	0.35	Alcohol dependence in EUR	rs3811647	21665994
0.582689081	0.16	Alcohol dependence in ASN	rs2074356	23183491
0.452888527	0.0902	Opioid dependence in ASW	rs12053259	23533358
0.3671001	0.3371	Post-surgical opiate analgesia in female JPT	rs2551941	23183491
0.440686308	0.3371	Post-surgical opiate analgesia in female JPT	rs2709394	23183491
0.278641	0.75	Post-surgical NSAID analgesia in EUR	rs2650805	19207018
0.270373792	0.4	Lithium response in bipolar disorder I in AMR	rs4354668	23023733
0.07841444	0.75	Response to nortryptyline in EUR	rs2486416	20360315
0.1838899	0.2	Response to citalopram in EUR	rs4986894	24528284
0.15298036	0.88	Antipsychotic drug-QT prolongation in EUR	rs4959235	20921969
0.27667143	0.28	Antipsychotic drug-induced cardiac risk in EUR	rs6741819	20195266
0.09283385	0.0059	Antipsychotic drug-triglycerides in EUR	rs79535779	20195266
0.8954187	0.101	Adverse response to carbamazepine in JPT	rs2844796	21149285
0.87222116	0.0962	Adverse response to carbamazepine in JPT	rs1633021	21149285
0.8993825	0.03	Adverse response to carbamazepine in EUR	rs1061235	21428769
Background neuropsychiatric disease risk SNPs for comparison				
0.132053536	0.2204	Bipolar disorder	rs10994415	24618891
0.0961068	0.1414	Bipolar disorder	rs12290811	24618891
0.072391305	0.1768	Bipolar disorder	rs17826816	24618891
0.062612533	0.5	Bipolar disorder	rs12202969	24618891
0.067523554	0.4394	Bipolar disorder	rs6550435	24618891
0.077216543	0.8333	Bipolar disorder	rs2011503	24618891
0.28332797	0.1667	Schizophrenia	rs114002140	23974872
0.057657467	0.3182	Schizophrenia	rs7085104	23974872
0.113328685	0.798	Schizophrenia	rs1198588	23974872
0.052657558	0.37	Schizophrenia	rs1006737	23974872
0.082000003	0.096	Schizophrenia	rs17691888	23974872
0.047611987	0.0505	Schizophrenia	rs4129585	23974872

0.118846066	0.0404	Schizophrenia	rs17504622	23974872
0.082000003	0.2273	Schizophrenia	rs6932590	19571808
0.114372469	0.0758	Schizophrenia	rs9960767	19571808
0.0938122	0.101	Schizophrenia	rs3131296	19571808
0.118846066	0.96	Schizophrenia	rs17693963	22688191
0.017528719	0.0606	Schizophrenia	rs11191580	22688191
0.063614443	0.3081	Schizophrenia	rs12666575	22688191
0.0961068	0.3889	Autism	rs4307059	19404256
0.078554294	0.5202	Bipolar disorder & schizophrenia	rs11789399	20889312
0.0795	0.3118	Major mood disorders	rs2251219	20081856
0.181538157	0.4451	Major depressive disorder	rs1545843	21521612
0.116524317	0.3485	Epilepsy	rs13026414	22949513
0.145419923	0.26	Epilepsy	rs72823592	22949513
0.213227868	0.197	Epilepsy	rs10496964	22949513
0.050644856	0.36	Migraine	rs2651899	23793025
0.074809192	0.1364	Migraine	rs12134493	23793025
0.036463696	0.65	Migraine	rs2274316	23793025
0.078560092	0.28	Migraine	rs6741751	23793025
0.043031237	0.3897	Migraine	rs9349379	23793025
0.052667603	0.1869	Migraine	rs11759769	23793025
0.058652042	0.1	Migraine	rs4379368	23793025
0.037844981	0.75	Migraine	rs827382	23793025
0.10104635	0.65	Migraine	rs3790455	22683712
0.100407764	0.38	Migraine	rs7640543	22683712
0.045353072	0.37	Migraine	rs9349379	22683712
0.08252895	0.5909	Migraine	rs6478241	22683712
0.136396365	0.17	Migraine	rs10166942	22683712
0.085162964	0.3535	Migraine	rs11172113	22683712
0.100734286	0.16	Migraine	rs11757063	22683712

Figure 2-5. Effect Sizes of Reported Variants

Effect sizes for graph for Figure 5 were calculated using a revised version of Cohen's *d*-test **[Cohen 1992]**, which was adapted to GWAS that report odds ratio. Using this method, effect sizes of 0.20 represent small effects, 0.50 represent large effects, and 0.8 represent very large effects. This method may differ from effect sizes calculated using different methods in GWAS. An effect size of unity (1.0) determines that if an individual harbors the variant, they will express the

pharmacogenomic phenotype. Since human genes (~25%) demonstrate significant allele-specific bias **[Palacios et al 2009]**, it is important to recognize that incomplete penetrance may result from skewed gene expression. Allele frequencies were determined from either 1000 Genomes Project data **[1000 Genomes Project Consortium 2012]** or from HapMap **[Gibbs et al 2013]**.

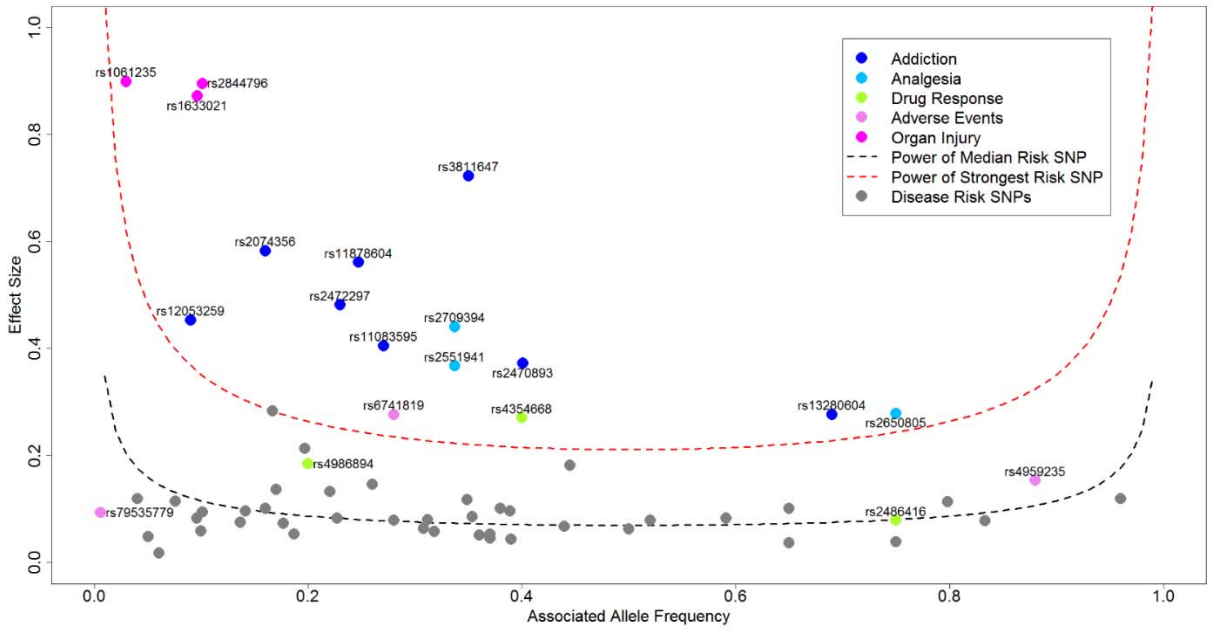


Figure 2-6. The Smile Plot

In contrast to neuropsychiatric disease risk SNPs which characteristically have low effect sizes, psychotropic drug SNPs (colors) exhibit moderate-to-large effect sizes and moderate-to-high allele frequencies. Almost all exceed the statistical power threshold of the median disease SNP (grey dotted line) and 80% exceed the power of all disease SNPs (red dotted line). Noncoding SNPs associated with addiction (dark blue) and analgesia (light blue) are notable for large effect and high frequency. Tagging SNPs at HLA loci associated with drug-induced organ injury (magenta) exhibit very high effect sizes but rarer allele frequencies. The SNP rs4354668 associated with lithium response in bipolar I patients of European ancestry (green) exhibits a moderate effect size and a common allele frequency, as does SNP rs6741819 (violet) which is associated with antipsychotic drug-induced cardiac risk. Neuropsychiatric disease risk SNPs were mined from the NHGRI GWAS catalog for bipolar disorder, schizophrenia, major depressive disorder, migraine and epilepsy, and are indicated in gray.

Knowledge-Based Discovery of Putative Spatial Interactions in Cis and Trans

Multiple public datasets were used to evaluate the spatial interactions of enhancers and promoters impacted by SNPs in gene association studies of psychotropic pharmacogenomic phenotypes, followed by knowledge-based discovery methods using published literature as described. Spatial interactions were plotted as shown in **Figure 2-7**. Several functionally-important relationships were inferred based on systematic evaluation of published literature, as well as other data sources such as www.clinicaltrials.gov. These relationships are concordant with RNA-Seq results **[Zhou et al 2015]**. In some cases, the association of enhancer RNAs (eRNAs) with this subset of known enhancers supported their role as active genome regulatory elements in the tissue of interest.

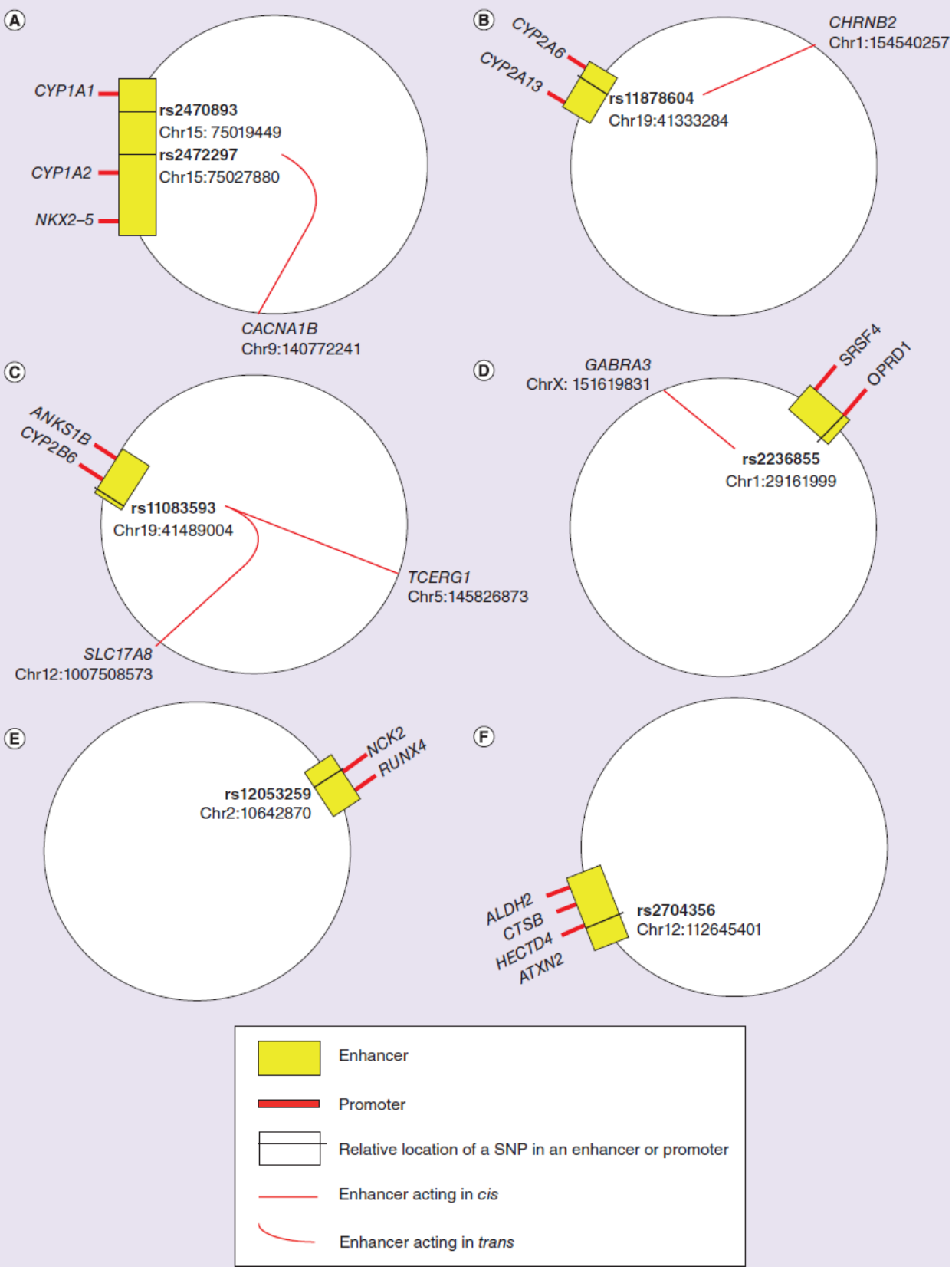
Figure 2-7-1 shows examples of SNPs that disrupt enhancers associated with addiction and analgesia. The SNPs rs2470893 and rs2472297 are well-replicated, statistically significant associations with caffeine dependence in European populations **[Sulem et al 2011, Amin et al 2012].** In *cis*, not only does the cognate enhancer control *CYP1A1* gene expression, but also the regulation of the *CYP1A2* gene, whose product is the rate-limiting enzyme for the metabolism of caffeine **[Zanger et al 2013]**. In addition, the large enhancer appears to regulate the *NKX2-5* gene, in which a SNP is significantly associated with correlated with increased caffeine metabolism in liver **[Cornelis et al 2011]**. For long distance (*trans*) interactions, the identified enhancer disrupted by these SNPs is in spatial contact with the *CACNA1B* gene as determined by Hi-C in immortalized liver cells (HepG2). Mutations in the *CACNA1B* gene have been linked to caffeine-induced anxiety in animal models **[Domschke et al 2012]**. In a study of nicotine dependence in a Japanese population **[Kumasaka et al 2012]**, the statistically significant lead SNP rs11878604 disrupts an

enhancer that regulates expression of the *CYP2A6* gene, which encodes a protein that metabolizes nicotine to cotine **[Zanger et al 2013]**, as well as *CYP2A13*, whose product CP2AD metabolizes the major nitrosamine found in tobacco, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone **[Su et al 2000]**. An enhancer involved in nicotine dependence **[Tobacco and Genetics Consortium 2010]** is linked in *trans* to the promoter of the *CHRNB2* gene, in which variants are correlated with the euphoric effects of nicotine **[Hoft et al 2011]** and the SNP rs11083593 located in the same locus, may activate differential splicing of the *CYP2B6* gene through a long-distance interaction with the *TCERG1* gene that generates a splice variant associated with increased incidence of lung cancer **[Timofeeva et al 2011]**. The *SLC17A8* gene is a glutamate transporter in brain whose expression has been significantly associated with addictive behaviors **[Enoch et al 2014].** The SNP rs2336855 is located in an intron of the *OPDR1* gene is associated with a population of individuals of European ancestry who were heroin addicts in a large case-control candidate study which has been replicated **[Nelson et al 2014]**. The same enhancer appears to control the splicing factor SRSF4 which is thought to be responsible for alternative splicing of the *OPRD1* gene **[Pandit et al 2013]**. In *trans,* it contacts the *GABRA3* gene, in which variants associated with reward deficiency and opiate addiction have been identified **[Kertes et al 2011, Blum et al 2011].** In a GWAS of opioid dependence **[Liu et al 2013]**, a significant SNP imputed from a population of African-American males shows that the enhancer regulates *NCK2*, which encodes a protein involved in synaptogenesis, and has been shown to be significantly associated with morphine tolerance in an animal model **[Huroy 2012]**. The SNP rs2074356 is significantly associated with alcohol dependence in a Korean population **[Baik 2011]** impacts an enhancer that modulates the *ALDH2* gene, in which variants are significantly associated with alcohol-related neuropathy and cancer **[Cederbaum et al 2012, Duell et al 2012]**. The enhancer also effects the *CTSB* gene, which has

been correlated with both addictive behavior in humans, alcoholism and alcohol-induced liver injury **[Lind et al 2013, Razvodovsky et al 2013, Jagannathan et al 2012]**. The enhancer is in spatial contact with the promoter of the *ATNX2* gene on the same chromosome. Variants in the *ATXN2* gene are responsible for a disorder called spinocerebellar ataxia **[Wang et al 2015]**, which is a characteristic adverse event associated with alcohol intoxication **[Modig et al 2012].**

Figure 2-7-2 shows further examples of the spatial interactions of enhancers and a promoter associated with psychotropic drug response and enhancers associated with antipsychotic drug-induced adverse events. These spatial relationships suggest the nature of the interaction that provides insight into mechanism or potential pathway components. In a GWAS of sustained antidepressant response, the significant SNP rs7816924 located in an intron of the *CSGALNACT1* gene, contacts the *CNTNAP3* gene in *trans*. *CNTNAP3* encodes a contactin-associated protein located at central synapses in which variants have been significantly associated with autism, intellectual disability syndromes and depression **[Zuko et al 2013, Verpelli et al 2014, Mitchell et al 2012, Berezin et al 2013]**. This enhancer also appears to regulate the *DLC1* gene, whose expression in human brain, methylation state and variants have been significantly associated with posttraumatic stress disorder, substance abuse, suicide attempts and suicidal ideation **[Mehta et al 2013, Mullins et al 2014, Gelernter et al 2014]**. The SNP 4354668 located in the promoter of the *SLC1A2* gene, which encodes a high-affinity glutamate transporter in human brain, is correlated with lithium response **[Dallaspezia et al 2012]**, exhibits a large effect size in patients of European ancestry with bipolar disorder, and is connected in trans to the *PLD5* gene. Variants in the *PLD5* gene have been significantly associated with bipolar disorder in bipolar patients of European ancestry **[Djurovic et al 2010]**. Mutations in the *SLC1A2* gene have been significantly linked to

essential tremor in GWAS, and tremor is a significant side effect of lithium treatment **[Their et al 2012].** In a GWAS studying the adverse event of QT prolongation in patients that take second generation atypical antipsychotic medications **[Aberg et al 2012]**, the imputed SNP rs4959235 is associated with quetiapine-induced QT prolongation, and the same enhancer impacts *NQO2* gene in *cis*, which has been associated with clozapine-induced agranulocytosis **[Ostrousky et al 2003]**, and is involved in acetaminophen hepatotoxicity **[Miettinen et al 2014]**. This enhancer is spatially connected with the *NOS1AP* gene, which has recently been shown to control QT prolongation in multiple, replicated studies **[Chang et al 2013, Jamshidi et al 2012, Kapoor et al 2014, Tomas et al 2010].** In a GWAS examining cardiac risk as a consequence of antipsychotic medication use in a European population **[Adkins et al 2011]**, the corresponding enhancer appears to regulate the *RNF144A* gene, in which variants have been significantly associated with cardiac dysfunction **[Kolder et al 2012]**.


A
CYP1A1
rs2470893
Chr15: 75019449
rs2472297
Chr15:75027880
CYP1A2
NKX2–5
CACNA1B
Chr9:140772241
B
CYP2A6
CYP2A13
rs11878604
Chr19:41333284
CHRNB2
Chr1:154540257
C
ANKS1B
CYP2B6
rs11083593
Chr19:41489004
TCERG1
Chr5:145826873
SLC17A8
Chr12:1007508573
D
GABRA3
ChrX: 151619831
SRSF4
OPRD1
rs2236855
Chr1:29161999
E
rs12053259
Chr2:10642870
NCK2
RUNX4
F
ALDH2
CTSB
HECTD4
ATXN2
rs2704356
Chr12:112645401
Enhancer
Promoter
Relative location of a SNP in an enhancer or promoter
Enhancer acting in cis
Enhancer acting in trans

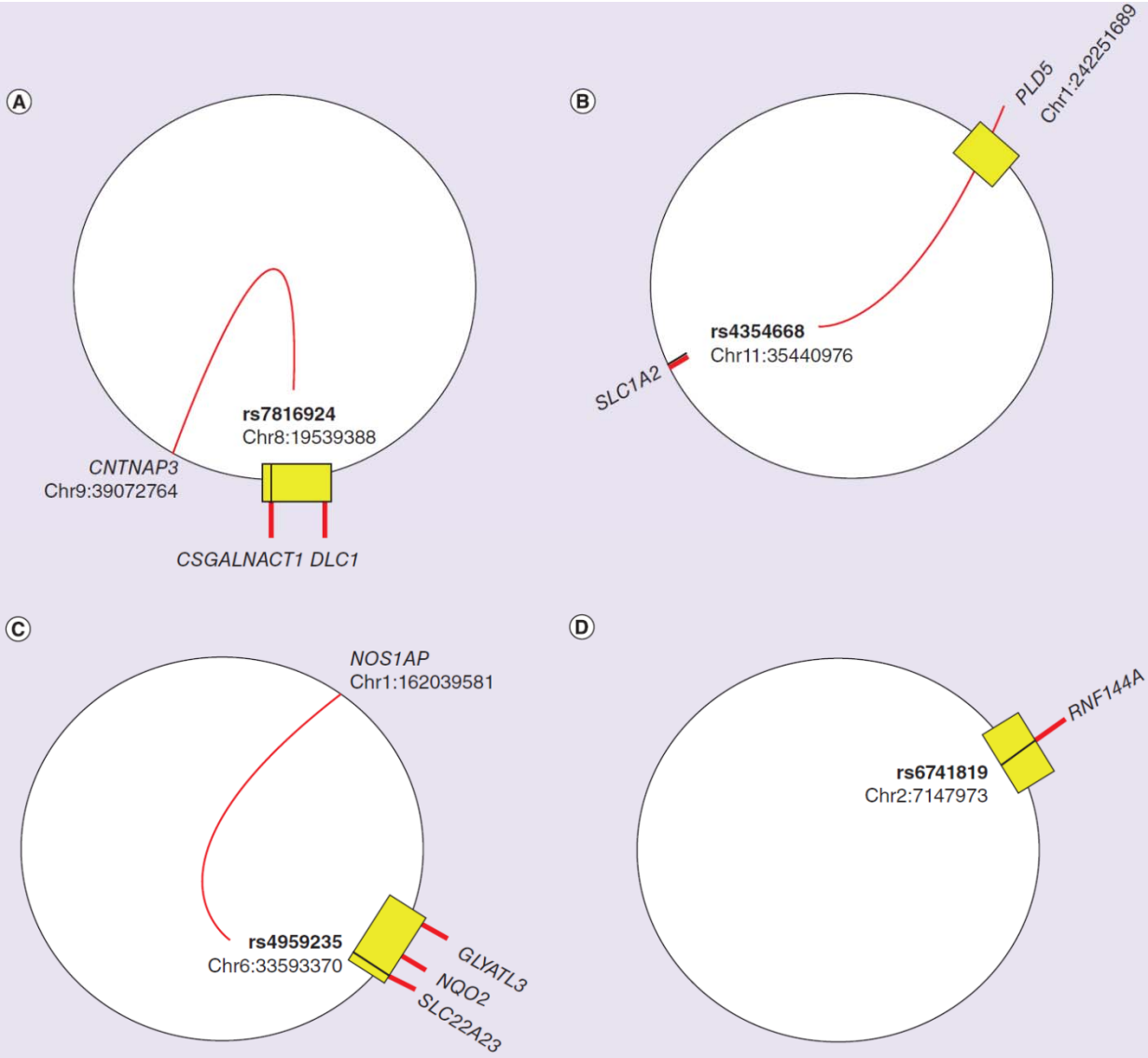


Figure 2-7: Spatial Context of Reported Variants

Figure 2-7-1: Spatial interactions of enhancers associated with addiction and analgesia plotted onto circular genome-wide maps. [A, B] SNPs significantly associated with caffeine dependence in individuals of European ancestry replicated in multiple GWAS; [B] A SNP significantly associated with nicotine dependence in a Japanese population; [C] A functional SNP associated with heroin addiction in Europeans; [D] A SNP associated with opioid addiction in

African-American males; [E] A SNP significantly associated with alcohol dependence in a Korean population.

Figure 2-7-2: Spatial interactions of enhancers associated with psychotropic drug response and antipsychotic-induced adverse events plotted onto circular genome-wide maps. [A] A SNP associated with sustained antidepressant response; [B] A SNP located in the promoter of the SLC1A2 gene associated with lithium response in bipolar patients of European ancestry; [C] A functional SNP associated with quetiapine-induced prolonged QT syndrome in an American population; [D] A SNP associated with antipsychotic drug-induced cardiac risk in a European population.

Discussion

Linear and combinatoric pharmacogenomic variant genotyping tests have already demonstrated considerable value in a number of systems. The discovery of additional variants which predict clinically important phenotypes is of great potential value. This study was undertaken to determine if candidate pharmacoepigenomic regulatory variants in humans could be identified using informatics and extant public data resources prior to further analysis in cell lines, animal models and clinical trials. Although these are bioinformatics-based predictions, they differ from genome-wide applications such as ChromHMM **[Ernst et al 2012]**, ChromImpute **[Roadmap Epigenomics Consortium 2015]**, GBR **[Libbrecht et al 2015]** and Segway **[Hoffman et al 2012]** because they make use of context-specific datasets collected in disease-relevant tissue types. For this reason, and because they are derived from gene association studies in human populations and many exhibit substantial effect sizes, our candidate variants and their predicted sites of action represent candidates for subsequent investigation and utility in biological experiments, biomedical and clinical validation research and, ultimately, in clinical practice.

The informatics-based predictions of mechanisms found in this study, as illustrated in **Figure 2-7**, can be validated by subsequent experimental studies in cells and tissues to observe spatial interactions between promoter-enhancer pairs in cell lines and tissues. These studies may use methods such as FISH microscopy, tissue-based Hi-C, and related methods for the study of dynamic genome architecture as was demonstrated in Rao et al **[Rao et al 2014]**. These variants and others found using these approaches may also be considered high priority variants for combinatoric effect size analysis. Although combinatoric effect size analysis has proven

intractable in a genome-wide context, where the number of comparisons allows statistical noise to overcome faint signals, with a reasonable number of loci such as the outputs of an experiment of this sort, the number of comparisons in combinatoric analysis is such that reasonable experiments (e.g. GWAS and clinical trial cohorts) may provide enough statistical power for such analyses to yield meaningful results. This analysis provides a straightforward and comprehensive approach to prioritizing psychotropic pharmacoepigenomic variants and corresponding regulatory pathways that will support a new generation of clinical pharmacogenomic tests and CNS drug discovery.

In addition, these results and others point to an emerging understanding about potential new ways to interpret GWAS datasets.

First, the importance of noncoding regulatory variation in the determination of pharmacogenomic response and adverse drug events is now being realized. Although non-synonymous mutations within exons contribute to disease risk and variable drug response, this biological phenomenon cannot begin to explain the human variome, with additional clarity provided by epistasis and epigenomics. Zanger et al **[Zanger et al 2014]** have discussed this phenomena concerning noncoding SNPs in *CYP* genes. Since significant evidence suggests that the majority of variation in human traits lies outside protein-coding regions of the genome **[Kellis et al 2014]**, epigenome informatics provides a step towards a better understanding of the variation in pharmacogenomic traits that exist in noncoding regions of the human genome **[Boyle et al 2012, Schaub et al 2012].**

Second, the epigenome, in addition to serving as a mechanism for non-sequence inheritance and regulation of transcription patterns, is important in both clarifying which variants are important

and predicting their mechanism of action. The brain generates the highest transcriptional complexity of any tissue by more than an order of magnitude **[Hawrylycz et al 2012, Kang et al 2011]**, involving transcription of at least 40% of all human coding genes **[Roadmap Epigenomics Consortium 2015]**, and the regulatory networks involved in neuropsychiatric diseases are likely to be similarly complex. Thus, it is probable that the spatial connections of the enhancers, promoters and transcriptional elements detected by these significant variants can provide explanatory value given the examples as shown in **Figure 2-7.** New data types including pQTLs, metQTLs, and sequence based predictions of variant effects on the epigenome are likely to provide explanatory value as well. Like any new approach in biomedical research that needs further experimental validation, epigenomic analysis of noncoding variants that explain gene regulation will demonstrate its utility through a more thorough understanding of the corresponding phenotype. We note that many of the new intragenic and intronic pharmacoepigenomic SNPs we have uncovered in this work are of the PD class, which is dramatically under-represented in pharmacogenetics today.

Third, the 4D nuclear context (3D structure plus temporal dynamics) is quite important in interpreting neuropsychiatric GWAS. In parallel with emerging results from studies of 3D nuclear architecture, phenomena such as circadian rhythmicity of gene expression **[Zhang et al 2014]** and the timing of interactions between specific CLOCK gene loci demonstrated by Hi-C **[Aguilar-Arnal et al 2014]**, emphasize the importance of temporal dynamics for understanding the functional impact of gene expression. New evidence is challenging existing dogma including recognition of the pervasive nature of mono-allelic transcription in mammalian cells as determined using single cell RNA-seq and 3D FISH at the cellular level **[Borel et al 2015, Deng et al 2014,**

Osborne et al 2004, Palacios et al 2009] and allele-specific expression at the tissue and population level, distinct from known mechanisms of genomic imprinting **[Schalkwyk et al 2010, McDaniell et al 2010]**. Non-coding variants identified in genome-wide association studies (GWAS) can be annotated as genomic regulatory elements **[Ward et al 2012, Schaub et al 2012]**, supporting the discovery of novel regulatory pathways that can be studied temporally.

Fourth, GWAS have demonstrated variable utility in different systems. GWAS research in complex human diseases has frequently shown a large number of small-effect, high frequency SNPs **[Park et al 2010, Willer et al 2013, Figure 2-6]**, and seldom shown common high-effect SNPs **[Park et al 2011]**. Our results clearly demonstrate that neuropsychiatric pharmacogenomic phenotypes do not exhibit the same constraints. Several of the pharmacogenomic systems we investigated display SNPs that are sufficiently powerful and in high frequency such that, alone or in combination **[Stringer et al 2011]**, they could potentially be used in clinical pharmacogenomic assays if validated in clinical research studies.

The reasons why this appears to be common in pharmacogenomics but rare in disease prediction remain unclear. The literature in disease GWAS is converging on the notion of an evolutionary "ceiling" on the significance (prevalence and effect size) of individual variants imposed by the deleterious fitness effect of disease **[Cohen 1992, Lange et al 2015]**. Although it could be argued that pharmacogenomic phenotypes would not be subject to the same evolutionary constraints because some of these medications only appeared recently in evolutionary terms, this explanation makes less sense in light of the relatively large effect sizes observed for SNPs for naturally-occurring addictive drugs such as alcohol and opiates which have been used by humans for at least

3500 years (opiates) to 9000 years (alcohol) **[Brownstein et al 1993, Crocq et al 2007]**. Xenobiotic substances in general often act on, or are acted on by, biological systems which are key physiological processes subject to positive of balancing selection **[Sadee et al 2014]**, and so these loci may emerge as hits in disease or other GWAS, while the particular variants responsible for response to xenobiotics may still not have been subject to direct selection. Further work will define the contours of this phenomenon, but it is clear that the utility of GWAS may vary substantially by phenotypic class, rendering global judgements in this area premature.

Fifth, the relevance of the traditional significance threshold (and in the future, any hard threshold) is declining. While the Bonferroni or "X^{2}" threshold often limits the number of significant lead SNPs reported in pharmacogenomic studies, recent studies have used imputation methods to identify variants in high LD with sub-Bonferroni lead SNPs, often greatly increasing the significance of "functional" or "causal" variants that represent bona fide associations with the trait of interest as opposed to so-called "tagging" SNPs **[Edwards et al 2013]**. Other researchers have convincingly argued that more informed systems approaches indicative of pleiotropy, including false discovery rate, are more accurate measures of significance in GWAS compared to traditional metrics **[Lee et al 2012, Benjamini et al 1995, Jia et al 2010]**. The totality of these strategies and our findings demonstrate that there is a great deal of useful data to be mined from existing GWAS that may have previously been dismissed as inconsequential **[Shi et al 2011]**. There is every indication that a strategy of permissive thresholding, followed by screening with other forms of information, is to be preferred, at least in further pharmacogenomic studies. In the future, "co-imputation" strategies may emerge which dispense with GWAS significance thresholds entirely

and screen variants with a combination of association and other methods from the very beginning of the workflow.

The results presented here are promising targets for future research that validate neuropsychiatric pharmacogenomic regulatory variants as clinical biomarkers, and the methods used to discover them, as part of an emerging new understanding of GWAS methods and significance, can offer powerful utility for discovering, interpreting, and validating functional variants in a number of systems, eventually adding to the considerable value already obtained from pharmacogenomic testing. Such variants may include not just SNPs but also indels, and these methods are extensible to probe variants called from NGS profiles of patient cohorts. Results from this study provide insight into potential regulatory mechanisms of gene expression underlying psychotropic drug response, and establish an *in silico* bioinformatics-based pipeline approach to discovering novel putative pharmacoepigenomic variants. This approach can serve to streamline the discovery of novel variants, which can then be further validated in cell lines, animal models and clinical trials using a diverse set of molecular and morphological data types.

The Lithium Glutamatergic Pathway

Motivation

Bipolar disorders are a spectrum of neuropsychiatric disorders characterized by abnormal shifts in energy, activity levels, and mood. Patients diagnosed with bipolar disorder also meet criteria for another lifetime disorder, more than half of whom had 3 or more other psychiatric disorders **[Merikangas et al 2011]**. There is a strong genetic component to susceptibility to bipolar disorder. Since bipolar spectrum disorders are heterogeneous and coincide with other comorbid disorders, estimates of heritability need to be carefully defined as to the subtype that is evaluated. Neuroimaging studies performed in family cohorts provide some of the most accurate data **[Ladouceur et al 2008]**. The lifetime risk for bipolar affective disorder is 15%–30% in individuals with 1 first-degree relative with bipolar disorder and up to 75% in those with 2 affected first-degree relatives **[Grof et al 2002]**. Recent studies have demonstrated strong overlap between gene associations and pathways shared between different psychiatric disorders **[Sullivan et al 2012, Cross-Disorder Group 2013]**. The genetic concordance rate for monozygotic twins is around 70%. The concordance rate among monozygotic twins is higher for bipolar disorder than unipolar affective disorder, suggesting a relatively greater contribution from genetic factors in the etiology of bipolar disorder compared with schizophrenia and unipolar depression **[Noga et al 2004]**. However, monozygotic twins may differ by variance in epigenome alterations produced by gene by environment interactions, so divergence is not wholly attributable to genetic variation **[Petronis et al 2003, Fraga et al 2005, Bel et al 2011]**.

Lithium is an alkali metal with no known role in human physiology, although many solute carrier proteins encoded by the SLC gene family can transport lithium across the cell membrane. Lithium salts were first used to treat mania in the 1950s and bipolar disorder in the 1960s. There appears to be a subpopulation of bipolar patients that are excellent responders to lithium salts, and this trait demonstrates familial clustering **[Grof et al 2002, Rybakowski et al 2010, Passmore et al 2003]**. Bipolar disorder patients who respond well to lithium salts exhibit higher genetic liability, which has led to a variety of studies focusing on the families of these patients, and most of these studies have confirmed an increased frequency of bipolar disorder among relatives **[Grof et al 2002, Rybakowski et al 2010]**. Studies investigating family histories of patients who respond to other drugs such as lamotrigine **[Passmore et al 2003, Turecki et al 2001]** suggest that different treatments may be most effective in patients who are clinically and biologically distinct from those patients who respond well to lithium. However, many bipolar patients take multiple psychotropic medications and may have a poorly-defined history of medication use and/or drug abuse, so definitive pharmacogenomic results have been difficult to obtain.

Two studies have examined lithium response in bipolar disorder with single nucleotide polymorphisms (SNPs) using genomewide association (GWA) in human cohorts. A genomewide association study (GWAS) examining lithium response in a Han Chinese population reported highly significant SNPs within introns of the *GADL1* (glutamate decarboxylase-like 1) gene **[Chen et al 2014]**. The human *GADL1* gene not only shares homology with *GAD1* and *GAD2*, which function in the conversion of glutamate to the inhibitory neurotransmitter GABA (g-aminobutyric acid), but also displays homology with an *archaeal* glutamate decarboxylase homolog, which functions as an aspartate decarboxylase **[Tomita et al 2014]**. L-glutamate and L-aspartate are

excitatory neurotransmitters in human brain that act as agonists at members of the NMDA and AMPA receptor families.

In another replicate GWAS examining time-to-recurrence of bipolar disorder in patients treated with lithium that was withdrawn from the National Institute of Human Genome Research Institute (NHGRI) GWAS catalogue, Perlis et al **[Perlies et al 2009]** found no associations that met the statistical threshold of GWAS. They did, however, find a SNP in a noncoding domain of the *GRIA2* (glutamate receptor ionotropic AMPA 2) gene that did not meet the rigorous "X^2" GWAS statistical threshold, but whose gene expression is significantly down-regulated following chronic lithium treatment in a human neuronal cell line and human hippocampus **[Seelan et al 2008, Du et al 2008]**. Using functional magnetic resonance imaging (fMRI), concentrations of glutamate in cortex, amygdala, striatum and hippocampus have been consistently shown to be significantly increased in bipolar patients that respond to lithium compared to controls **[Lim et al 2013]**.

Candidate gene association studies have implicated a number of almost exclusively noncoding SNPs associated with lithium in bipolar disorder. Candidate SNPs, although limited by the small sample sizes and confounding problems that are characteristic of many published gene association studies, have been linked to dozens of genes. These include genes with properties suggestive of bipolar disorder and/or lithium. Among these are genes which:

1. Convey weak genetic effects to risk of bipolar disorder and other neuropsychiatric diseases
2. Participate in inositol–based second messenger systems
3. Are involved in synaptic vesicle recycling or are structural components of post-synaptic densities

4. Regulate cell proliferation, growth and apoptosis
5. Are members of the CLOCK gene family that are dysregulated in bipolar disorder and major depressive disorder
6. Are significantly associated with central glutamatergic neurotransmission
7. Enhance neuronal cell adhesion
8. Are solute carriers or ion channels or are transcription factors involved in brain patterning, development and hormonal regulation of gene expression

Recent discoveries in human genetics, genomics and epigenomics has broadened the range of variants that can be used to classify patients by drug efficacy and drug-induced adverse events. Research over the past 5 years has upset orthodox views that emerged from the sequencing the draft human genome at the turn of the 20th century. Results of the Encyclopedia of DNA Elements (ENCODE) project **[ENCODE Project Consortium 2012]** and the Epigenome Roadmap Consortium **[Roadmap Epigenomice Consortium 2015, Leung et al 2015]** studies that have defined the three-dimensional (3D) organization of the human genome at resolution using methods such as chromatin conformation capture, including Hi-C **[Rao et al 2014]** and reinterpretation of GWAS data using pathway analysis **[Jia et al 2010, Shi et al 2011, Tasan et al 2014, Califano et al 2012, McGeachie et al 2014]**, have demonstrated the following:

1. Coding variants represent a fraction of the genetic differences between human cohorts for traits such as drug response **[Zhou et al 2013, Kellis et al 2014, Meyer et al 2013, Maurano et al 2012, Roussos et al 2014, Onengut-Gumuscu et al 2015, Weinhold et al 2014]**. The

majority of statistically-significant SNPs found by GWAS are noncoding and disrupt regulatory elements such as enhancers **[Farh et al 2015]**;

2. GWAS SNPs that may not have met the traditional, stringent statistical threshold of $p \leq 5 \times 10^{-08}$ are now being shown as integral to understanding the genetic foundation of phenotype, particularly if they provide explanatory mechanistic power using pathway analysis **[Jia et al 2010, Shi et al 2011, Tasan et al 2014, Califano et al 2012, McGeachie et al 2014]**. In addition, several psychiatric disorders share common genome regulatory pathways **[Cross-Disorder Group 2013]**, as do complex disease traits **[McGeachie et al 2014]**, and epigenome pathways contain common risk variants for seemingly unrelated diseases **[Farh et al 2015]**;

3. In diseases associated with psychological stress, epigenomic modifications of the human genome are the best indicators of disease risk and drug response, and these alterations can be transmitted from parent to child **[Crews et al 2012, Dietz et al 2011]**. These results reinforce decades of research in the social sciences, clinical psychology and psychiatry, showing that the offspring of abused children were most likely to abuse their own children **[Peled et al 2011, Kaufman et al 1987, Larrance et al 1983]**;

4. The bulk of recent studies suggest that "so-called" missing hereditability is not genetic, but rather epigenetic **[Petronis 2010, Slatkin 2009, Labrie et al 2012, Cortijo et al 2014, Badyaev et al 2014]**, with inherited disease-related variation located in genomic regulatory elements **[Kellis et al 2014, Maurano et al 2012, Farh et al 2015, Parker et al 2013, Kumar et al 2013]**;

5. The 3D chromatin environment in the nucleus provides the foundation for the regulation of gene expression, and gene transcription is organized into nuclear compartments, including

topologically-associating domains that involve both intra-chromosomal and inter-chromosomal interactions that vary in space and time **[Dekker et al 2013, Jin et al 2013, Lieberman-Aiden et al 2009, Deng et al 2012, Seitan et al 2013, Kleinjan et al 2005, Noordermeer et al 2011, Li et al 2012, Cremet et al 2001, Wendt et al 2014, Aguilar-Arnal et al 2014]**;

6. The specific milieu of any given genomic element, as determined by accessibility to transcription factors in open chromatin, and driven by genome variants, such as SNPs and transcription factor binding, dictates its function **[Ernst et al 2010, Ernst et al 2012, Hon et al 2009, He et al 2012, John et al 2012]**; *and*
7. Differences in chromatin state are a major determinant of what had previously been attributed to genetic differences between human phenotypes **[Heyn et al 2013, Hellman et al 2010, Kasowski et al 2013]**, often resulting in allele-specific methylation and allele-specific expression **[Shoemaker et al 2010, Schalkwyk et al 2010]**, largely driven by gene variants and correlative binding of transcription factors **[Kilpinen et al 2013, McVicker et al 2013, Jones et al 2013]**.

When viewed in the context of these recent findings, it is no longer tenable to maintain that the majority of complex, polygenic inherited traits such as drug response are only, or even primarily, a consequence of missense or nonsynonymous changes in exons that alter protein structure and function. The paucity of coding variants predictive of psychotropic phenotypes, and the presence of a "missing inheritance" gap, as well as the presence of a profundity of transcriptional regulation in the brain, suggest that psychiatry is no exception. Our hypothesis in the present study is that combining functional SNPs imputed from lead SNPs reported in association studies of lithium

drug response, together with those associated with bipolar disorder, will provide insight into genomic regulatory pathways that mediate psychotropic drug response **[Jones et al 2013, Flister et al 2013, Edwards et al 2013, Schaub et al 2012].** We developed and applied a novel bioinformatics pipeline method for characterization of noncoding SNPs that includes chromatin state annotation, spatial mapping of enhancers and promoters in 3D space, neuroanatomical localization using both in situ hybridization of mRNA and enhancer in postmortem human brain tissue, neuroimaging data from patients and controls, and pathway analysis. In this article, we present preliminary data on imputed SNPs that were screened using pathway analysis and RNA localization in the context of existing knowledge of human brain function and lithium response in bipolar disorder.

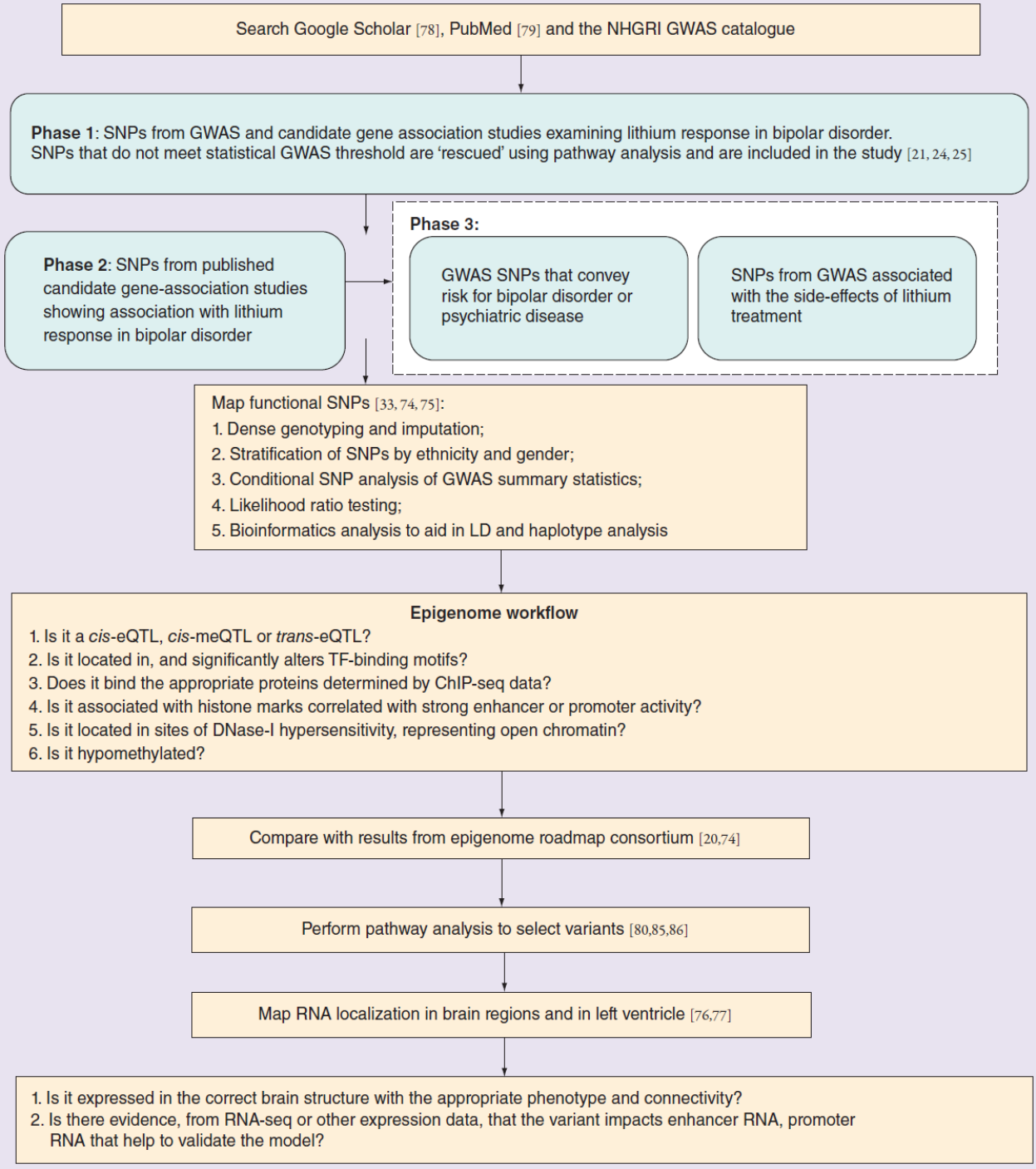


Figure 2-8. Diagram of Lithium Analysis Methods

For neuroanatomical characterization, we used a variety of resources, including (1) Human data from the Allen Brain Institute, which includes postmortem human brain tissue sections used for in situ hybridization, RNA-seq and microarray analysis **[Sunkin et al 2013]**, and (2) Data on

enhancer RNA localization from the epigenome roadmap consortium **[Roadmap Epigenome Consortium 2015]**. A federation of publically-available resources, including those of the Allen Brain Institute, the NIH epigenome roadmap project and databases in bioinformatics provide a highly detailed, comprehensive map of connectivity in human brain, including transcriptomic signatures for many neuronal subtypes and other cell types. Similarly, there are sources of detailed data on the localization of functional enhancers in major regions of human brain that were used, including validated RNA-seq measures of enhancer RNA, methylome signatures and many other epigenomic characteristics.

Methods

Figure 2-8 shows an overview of the method that was used in the study that is further described in the text. The major steps in the "pharmacoepigenomic trait workflow" included SNP imputation from GWAS and candidate gene association studies examining lithium response in bipolar disorder, genetic and epigenetic fine mapping, chromatin state annotation, pathway analysis and multiscale epigenome analysis. All are described in detail below.

Study Selection and Fine Mapping

A literature review was performed using Google Scholar **[Pomerantz 2013]** and PubMed **[Lu 2011]** for studies published prior to March 15, 2015, using different search strings including "bipolar AND lithium AND gene AND association" and "lithium response AND variants AND bipolar disorder." In addition, the National Human Genome Research Institute's (NHGRI) GWAS catalogue was searched, and spreadsheets were downloaded for evaluation. In the NHGRI catalog, many single nucleotide polymorphisms (SNPs) related to bipolar disorder, other psychiatric disorders and sleep habits were shared. For example, several GWAS had identified the same or linked 'lead' SNPs in the *ANK3* and *CACNA1C* genes.

Sets of lead SNPs, reported in genetic association studies, were examined for linkage with noncoding SNPs that could identify genomic regulatory elements such as enhancers. Several sets of variants derived from single studies and combinations of studies were evaluated. Studies were selected in 3 phases: (1) 2 GWAS of lithium response in patients with bipolar disorder, (2)

Candidate gene association studies examining association between lithium response and bipolar disorder and (3) GWAS of bipolar disorder and other psychiatric disorders in which functional SNPs could be imputed that were the same as found in the first phase of study selection.

Fine mapping to determine the causal nature of a given SNP has been described elsewhere **[Farh et al 2015, Edwards et al 2013, Mi et al 2013]**, but uses methods based on assumptions that, in the majority of association studies, it is necessary to discriminate SNPs that represent the most likely functional or causal variants from those with strong but significantly less association to the drug response trait, so that the remainder of SNPs at a given locus can be excluded from further analysis. As a consequence of tight LD that exists within a haplotype or haplotypes associated with a trait, the 'lead' SNP(s) that attain the threshold of significance used in GWAS may not serve as the best candidates for functional or 'causal' SNPs. Unfortunately, it has been difficult to obtain large sample sizes in psychiatric pharmacogenomic studies, resulting in a major limitation of the power of such association studies.

Accurate determination of a SNP's significance of association with a pharmacogenomic trait and exclusion of non-functional SNPs at a given locus included steps based on Edwards et al **[Edwards et al 2013]**. Dense genotyping and imputation methods were used to define what SNPs are contained in a haplotype block identified in a GWAS or other study. For determination of LD between SNPs, r2 values were calculated in a pairwise manner using the genotype dataset of each SNP **[Flister et al 2013, Wang et al 2012]**. LD blocks were defined among the SNPs that exhibited strength, using multiple approaches. These included re-calculation of the upper and lower 95% confidence limits on r2 for a strong LD that were set at 0.8. LD patterns between the SNPs with

minor allele frequency ≥ 0.05 in the same genomic position were similarly analyzed using HaploReg, version 2 **[Ward et al 2012]**. These methods have been independently demonstrated **[Higgins et al 2015]**.

The first phase included 2 GWAS, the first one performed by Perlis and collaborators looking at response to lithium for prevention of recurrence in bipolar disorder **[Perlis et al 2009]**. The investigators did not find any SNP associations that met the "X2" statistical threshold, but found several that showed strong sub-"X2" threshold. Using pathway analysis, including Ingenuity Pathway Analysis™ **[Kramer et al 2013]** and open source tools including pathway commons and Panther **[Cerami et al 2011, Mi et al 2013]**, we were able to identify a SNP in the *GRIA2* gene that was part of a very significant pathway. The second included GWAS found noncoding SNPs in the *GADL1* gene that were significantly associated with lithium response in a Han Chinese **[Chen et al 2014]** population.

The second phase included the following studies: (1) SNPs reported from a Sardinian population that exhibited long term responsiveness to lithium **[Squassina et al 2011]**; (2) Rare variants obtained from sequencing families that might represent lithium-responsive pedigrees **[Cruceanu et al 2013]**; (3) A SNP in the *SLC1A2* gene, which has been associated with lithium response in recurrent bipolar disorder in an Italian population **[Dallaspezia et al 2012]**; (4) SNPs in the *SLC4A10* gene that were associated with lithium response in bipolar disorder **[Rybakowski 2013, Schulze 2012, Rybakowski 2014]**; (5) SNPs in the *GSK3* gene that were associated with lithium mechanisms in brain **[Benedetti et al 2005, Lin et al 2013]**; (6) SNPs recommended for further investigation for their association with lithium response in bipolar disorder, with the exception of

brain-derived neurotrophic factor, in which data from human studies where not significant **[McCarthy et al 2011, McCarthy et al 2010]**; and (7) SNPs in *CLOCK* genes which had been significantly associated with lithium response in bipolar disorder **[Rybakowski et al 2014]**.

The third phase included the following studies, which had lead or imputed functional SNPs shared with the phase 1 and phase 2 SNPs **[Byrne et al 2013, Soronen et al 2010, Ferreira et al 2008, Psychiatric GWAS Consortium Bipolar Disorder Working Group 2011, Chen et al 2013, Liu et al 2011, Rietschel et al 2010, Green et al 2013, Schizophrenia Psychiatric GWAS Consortium 2011, Rueckert et al 2013, Schumacher et al 2013, Spellman et al 2011, Schulze et al 2009, Thier et al 2012]**. This phase also included SNPs from 2 studies that might have SNPs related to the side effects of lithium response in bipolar disorder, or with symptomatology of the disorder. These include a SNP in the *SLC1A2* gene associated with essential tremor **[Thier et al 2012],** Brugada syndrome **[Crawford et al 2015, Wright et al 2010, Darbar et al 2005],** and sleep disruption **[Parsons et al 2013]**. The third phase was important because many of the same variants associated with bipolar disorder have been associated with lithium response in the disease **[Manchia et al 2013, Philips et al 2014, Strakowski et al 2012]**. This phase also included SNPs from GWAS showing association with side effects observed with lithium response in bipolar patients, including tremor and sleep habits.

Pathway Analysis

Several different strategies were used to determine the relationship between genomic regulatory elements identified by the SNPs. Since the objective of this study was to discover novel regulatory

genes and pathways for the mediation of lithium response in bipolar disorder, our initial research used both commercial and open source pathway analysis applications, including IPA™ **[Kramer et al 2013]**, pathway commons **[Cerami et al 2011]** and Panther **[Mi et al 2013]**. The fact that the majority of the SNPs identified elements that were located with introns and flanking regions of genes presented a challenge because function of many SNPs is still not well understood. However, we used pathway analysis of known biochemical interactions, and sought to further understand these pathways using disease and pharmacogenomic databases supplemented by extensive literature review.

Multi-Scale Epigenome Neuroanatomical Circuit Mapping

Our methodology also included manual examination of the imputed, functional SNPs and further classification of the functional SNPs in regulatory pathways based on neuroanatomical substrate and molecular interactions. Application of a neuroanatomical filter based on presumptive sites of lithium action in bipolar disorder was also included. Using a recent review about the neuroanatomical substrate summarizing the neuroimaging data, we filtered using this model–

> "Bipolar disorder can be conceptualized, in neural circuitry terms, as parallel dysfunction in prefrontal cortical (especially ventrolateral prefrontal cortical)-hippocampal-amygdala emotion- processing and emotion-regulation circuits bilaterally, together with an "overactive" left-sided ventral striatal-ventrolateral and orbitofrontal cortical reward-processing circuitry… A potential structural basis for these functional abnormalities is gray matter volume decreases in the prefrontal and temporal cortices, the amygdala, and the hippocampus and

fractional anisotropy decreases in white matter tracts connecting prefrontal and subcortical regions." **[Philips et al 2014]**.

Our expectation was that brain regions sampled and represented in data from both the epigenome roadmap consortium and data on RNA localization from other sources, including examination of the neuroanatomical distribution of RNAs from candidate regulatory elements from postmortem human brain, would indicate those brain regions showing regulatory events associated with lithium response in bipolar disorder. The human brain atlas from the Allen Brain Science Institute **[Sunkin et al 2013]** provides *in situ* hybridization and RNA microarray data on the distribution of specific mRNAs, while enhancer RNA localization data are available from the epigenome roadmap consortium **[Zhou et al 2015]**. We classified epigenome regulatory elements as to sites of *cis*-expression according to the following: (1) Mid-frontal cortex, as well as dorsolateral prefrontal cortex when available; (2) Angular gyrus, although there is less direct evidence to support an association, it is involved in the face processing pathway like the fusiform gyrus **[Dima et al 2013]**; (3) Anterior caudate, which contains the ventral striatum and nucleus accumbens; (4) Cingulate cortex; (5) Hippocampal formation, and (6) Inferior temporal cortex, which contains circuitry between the amygdala and hippocampus and prefrontal cortex and orbitofrontal cortex.

SYMBOL	HGNC NAME	FUNCTION / SIGNIFICANCE
ANK3	Ankyrin 3, node of Ranvier (ankyrin G)	Cytoskeletal adaptor protein that regulates sodium channel activity and sodium ion homeostasis at CNS GRIA2 synapses.
ARNTL	Aryl hydrocarbon receptor nuclear translocator-like	Transcription factor that forms dimer with CLOCK. Exhibits greatest circadian rhythmicity in human brain. The same SNPs have been significantly associated with lithium response in bipolar disorder.
CACNG2	Calcium channel, voltage-dependent, gamma subunit 2	Used to be known as 'Stargazin' – a transcription factor. Function in voltage-gated calcium channel in human brain not defined.
CACNA1C	Calcium channel voltage-dependent L type alpha 1C	Regulates heart contraction and fibrillation. Same SNP-detected enhancer responsible for Brugada syndrome: lithium-induced heart failure.
CDKN1A	Cyclin-dependent kinase inhibitor 1A	Regulates SWI/SNF mediated chromatin remodeling. Found in central glutamatergic neurons.
CREB1	Cyclic AMP-responsive element-binding protein 1	Ubiquitous distribution and function. SNP associated with SSRI antidepressant response and addiction, but data are not definitive.
GRIA2	Glutamate receptor, ionotropic, AMPA 2	GWAS showed association with lithium response in bipolar disorder and risk of schizophrenia.
GSK3B	Glycogen synthase kinase 3 beta	A kinase in the ATK / WNT signaling pathway. The same SNPs have been significantly associated with lithium response in bipolar disorder in many studies.
NR1D1	Nuclear receptor subfamily 1, group D, member 1	A transcription factor that regulates circadian rhythmicity. The same SNP has been significantly associated with bipolar disorder and lithium response in bipolar disorder.
SLC1A2	Solute carrier family 1 (glial high affinity glutamate transporter), member 2	Located in GRIA2 neurons in human CNS. The same SNP-detected enhancer associated with tremor, a side effect of lithium therapy. The same SNPs have been significantly associated with lithium response in bipolar disorder in many studies.

Figure 2-9: Reported Genes from the Liuthium Analysis

Ten Genes in Which Noncoding SNPs Were Found Within Introns and 5'UTRs. HGNC: HUGO Gene Nomenclature Committee.

Results

Filtered Set of Functional Lithium-Associated SNPs

Based on our hypothesis, and those of previous researchers **[McCarthy et al 2011, McCarthy et al 2010, Philips et al 2014, Strakowski et al 2012]**, that psychotropic drug response and psychiatric disease may share the same or overlapping sets of variants, we imputed, analyzed and annotated SNPs reported in GWAS that had been associated with bipolar disorder, other psychiatric disorders or side effects of lithium from the NHGRI GWAS catalogue. This phase 3 set included GWAS SNPs significantly associated with tremor, an adverse drug event associated with lithium response **[Thier et al 2012]**, sleep habits **[Parsons et al 2013]** and Brugada syndrome, which can be a consequence of lithium "un-masking' of ventricular arrhythmias related to sudden adult death syndrome **[Crawford et al 2015, Wright et al 2010, Darbar et al 2005]**. Following removal of spurious associations and epigenetic fine mapping according to the methods of Dunning et al **[Farh et al 2015, Edwards et al 2013]**, we characterized a set of 78 SNPs as bipolar disorder epigenome regulatory variants using our workflow.

Overlap Between Lithium-Response and GWAS Risk Variants

All 3 phases of this study were examined for SNPs shared in common, including phases 1 and 2 looking at lithium response candidates, as well as phase 3 GWAS studies that found SNPs in specific loci or genes carrying variants associated with bipolar disorder, a continuum of psychiatric disorders including bipolar disorder, risk variants for major depressive disorder and drug response,

and adverse events associated with lithium. We identified substantial overlap between SNPs from the published association studies on lithium response in bipolar disorder (phases 1 , 2) and the SNPs found in the bipolar, bipolar and major depressive disorder, bipolar and schizophrenia, bipolar as part of a continuum of psychiatric disease, and tremor genomewide association studies (phase 3). 27% of the phase 1 and phase 2 SNP sets, 217 SNPs, were also present in the phase 3 set. Of the 217 SNPs, 78 SNPs were annotated as active regulatory elements such as enhancers and promoters in a tissue-specific manner were located in 10 genes.

Comparison of Results with Those from Epigenome Roadmap Consortium

Through a process of LD pruning, we selected 19 functional SNPs that we characterized in more detail. Using data from the epigenome roadmap consortium browser **[Zhou et al 2015]** for selection of brain regions or left ventricle (for Brugada syndrome), followed by definition using ChromHMM **[Ernst et al 2012]**, ChromImpute **[Ernst et al 2015]** and the WUSTL epigenome browser **[Zhou et al 2011]**, we found high concordance between the results of our workflow and that of the epigenome roadmap consortium. The SNPs and the corresponding regulatory elements with tissue-specific localization for only mid-frontal cortex and hippocampus were provided by results from the epigenome roadmap consortium. Both mRNA and eRNA location is consistent, at least in human brain, with the tissue-specific assignment, as determined by corresponding data from the Human Brain Atlas from the Allen Brain Science Institute **[Sunkin et al 2013]** and from the epigenome roadmap consortium **[Zhou et al 2015]**. Not every SNP-detected regulatory element is reported for every brain region and peripheral tissue, just those considered important in the context of pathway analysis and RNA localization. The *CACNA1C* SNP associated with Brugada syndrome is expressed as a strong enhancer in the left ventricle.

RNA Localization

Most human brain region projects show some relationship to degenerative changes observed in bipolar disorder, and depending on the neurotransmitter circuits involved, may play a more central role than had been previously thought. Since our molecular mapping analysis had shown that *GRIA2* mRNA is regulated by enhancers, we compared the distribution of the expression of these genes with each other and with those regions compromised in bipolar disorder and thought to be sites of lithium response.

Both *GSK3B* and *CREB1* mRNAs show a pattern resembling either non-specific hybridization to neurons or expression like housekeeping genes. In contrast, although *SLC1A2* shows more uniform neuronal expression, it does resemble known distribution of glia or non-neuronal cell-types in these regions, although it has commonly been described as the "glial glutamate drug transporter" in the literature **[Thier et al 2012]**.

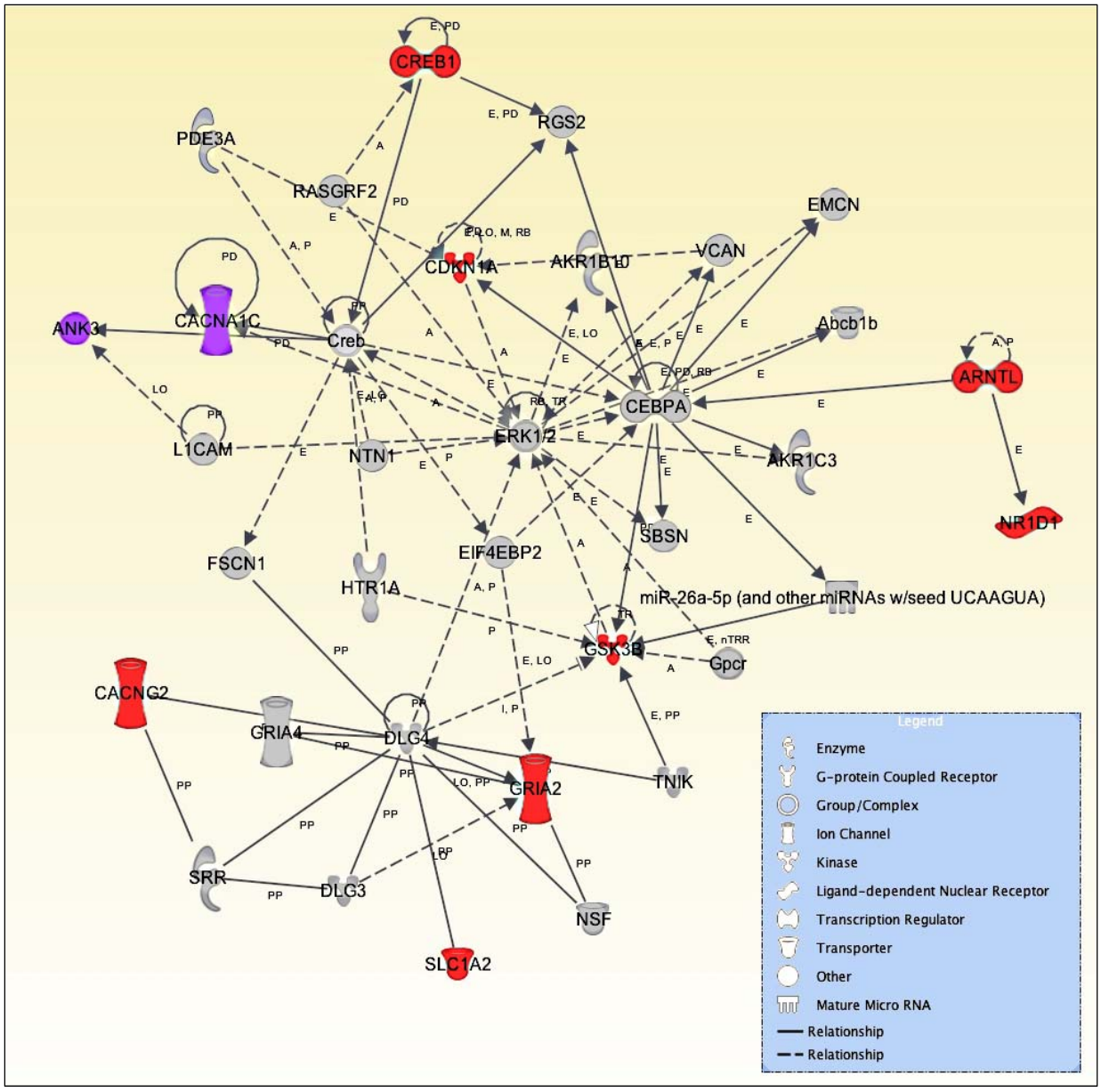


Figure 2-10. The Lithium Glutamatergic Pathway

Using the gene set shown in **Figure 2-9**, this shows an integrated glutamatergic pathway called by this software. This pathway is very significantly enriched for the "glutamate receptor" at $p<10^{-27}$ (Fishers exact test)**.** Colored objects in this regulatory network indicate the 10 genes that were found in this study, purple indicating ones that IPA™ called as bipolar risk genes (*ANK3* and *CACNA1C)*, although SNPs in other genes in this pathway have been associated with risk of bipolar disorder.

Pathway Analysis and Gene Set Enrichment

To understand the connections among our selected SNPs, we looked at the relationship between these genes using both open source and commercial pathway analysis software. **Figure 2-9** shows the 10 genes in which 19 noncoding SNPs were found to be significantly associated with lithium response in bipolar patients of European ancestry. **Figure 2-10** shows data from Ingenuity Pathway Analysis (IPA™; Qiagen) showing the enrichment of this gene set within a canonical glutamate receptor pathway. The same pathway was also found using open source pathway analysis software, including Panther network visualization based on Gene Ontology **[Mi et al 2013]**.

In addition, this pathway contains other genes of interest, such as *DLG4*, discs large homolog 4, a ionotropic glutamate receptor ligand, and nitric-oxide synthase regulator that acts in serotonin receptor signaling, in which mutations have been significantly associated with schizophrenia and bipolar disorder **[Frank et al 2011, Cheng et al 2010]**. In addition, the pathway also contains *HTR1A*, the 5-hydroxytryptamine (serotonin) receptor 1A, which is involved in behavioral reinforcement of reward **[Zhao et al 2014]** and, which mutations have been associated with schizophrenia. Another constituent of this pathway is the *GRIA4* glutamate receptor, the glutamate receptor ionotropic AMPA 4, in which mutations have been correlated with antipsychotic drug response in schizophrenia and bipolar disorder **[Makino et al 2003, Chiesa et al 2012]**. It is important to note that the gene product of *CACGN2*, the calcium channel voltage-dependent gamma subunit 2, originally called stargazin, is a primary regulator of *GRIA2* expression **[Selvakumar et al 2009, Hafner et al 2015]**.

A

GO cellular component	Number (#)	# Expected	Fold Enrichment	+ / -	P value
Dendrite	381	4	0.16	+	0.0122
Neuron part	981	5	0.40	+	0.0222
Neuron projection	797	6	0.33	+	0.00809

B

ID	Molecules in Network	Score	Focus Molecules
1	Abcb1b,AKR1B10,AKR1C3,**ANK3**,**ARNTL**,**CACNA1C**,**CACNG2**,**CDKN1A**,CEBPA,Creb,**CREB1**DLG3,DLG4,EIF4EBP2,EMCN,ERK1/2,FSCN1,Gpcr,**GRIA2**,GRIA4,**GSK3B**,HTR1A,L1CAM,miR-26a-5p,**NR1D1**,NSF,NTN1,PDE3A,RASGRF2,RGS2,SBSN,**SLC1A2**,SRR,TNIK,VCAN	27	10

C

Diseases or Functions Annotation	p-Value	Molecules	# Molecules
Neurodegeneration of brain cells	6.47E-11	ANK3,CREB1,GRIA2,GSK3B,SLC1A2	5
Neurodegeneration of cerebral cortex cells	3.92E-10	CREB1,GRIA2,GSK3B,SLC1A2	4
Behavior	8.78E-09	ARNTL,CACNA1C,CACNG2,CDKN1A,CREB1,GRIA2,GSK3B,SLC1A2	8
Neuronal cell death	2.04E-08	CDKN1A,CREB1,GRIA2,GSK3B,NR1D1,SLC1A2	6
Movement Disorders	2.32E-08	ANK3,ARNTL,CACNG2,CREB1,GRIA2,GSK3B,NR1D1,SLC1A2	8
Degeneration of neurons	3.33E-08	ANK3,CREB1,GRIA2,GSK3B,SLC1A2	5
Neurodegeneration of hippocampal neurons	6.04E-08	CREB1,GRIA2,GSK3B	3
Mood Disorders	6.79E-08	ANK3,ARNTL,CACNA1C,CREB1,GRIA2,GSK3B	6
Circadian rhythm	1.75E-07	ARNTL,CREB1,GSK3B,NR1D1	4
Bipolar disorder	3.90E-07	ANK3,ARNTL,CACNA1C,GRIA2,GSK3B	5

Figure 2-11. Gene Set Enrichment Analysis of Lithium Genes

(A) Subcellular compartmentalization from pathway commons **[Carbon et al 2009]; (B)** Pathway analysis using IPA™. All discovered genes (bold text) are contained within the same regulatory network. These genes are enriched in this pathway at $p < 1 \times 10^{-27}$ using Fisher's right-tailed exact test. **(C)** Determination of top 10 diseases and phenotypes associated with this regulatory network using IPA™**.**

Evaluation of Pathway Components

In this analysis, we identified multiple SNPs in some genes that had been proposed as candidates for mediation of lithium response and others that were identified by epigenomic mapping and pathway analysis. Within this set of SNPs that were filtered from a much larger dataset are contained CACNG2 (calcium channel voltage-dependent gamma subunit 2), also known as stargazin, which along with other AMPA transmembrane proteins, is a potent regulator of GRIA2 expression **[Selvakumar et al 2009, Hafner et al 2015]**. Also in this network are the genes *ANK3* (ankyrin 3 node of Ranvier) which encodes ankyrin G, a protein which regulates central glutamatergic synaptic organization in the human CNS **[Smith et al 2015]**, and *CACNA1C* (calcium channel voltage-dependent L type alpha 1C subunit), which is involved in the processing of episodic memory in human hippocampal neurons **[Krug et al 2014]**. Mutations in *ANK3* and *CACNA1C* have been significantly associated in GWAS with a variety of psychiatric disorders, including bipolar disorder, psychiatric drug response, and disruption of sleep disorders. A *CACNA1C* risk variant has also been found to be one of several mutations in ion channel genes that give rise to Brugada syndrome, which is a shortening of the QT interval in humans that leads to sudden adult death syndrome (SADS), which can be unmasked in patients that are treated with lithium salts **[Crawford et al 2015, Wright et al 2010, Darbar et al 2005]**.

There have been several molecular mechanisms suggested as the basis of lithium response in human brain, ranging from non-specific effects based on its status as an elemental metal acting in a nebulous manner, such as membrane stabilization, to a very specific mechanism that impacts intracellular WNT signaling and other pathways working through GSK3B (glycogen synthase

kinase 3 beta). The first evidence for a role of lithium acting at GSK3B was the observation that this serine/threonine kinase is directly inhibited by lithium in cultured cells **[Stambolic et al 1996]**. This led to several studies in which a promoter SNP in GSK3B was found to be associated with lithium therapy in bipolar disorder and depression **[Adli et al 2007, Benedetti et al 2005, Lin et al 2013]**, analysis of mechanisms of atrophy in bipolar disorders, and how lithium and GSK3B SNPs act in a similar manner to protect neural degeneration **[Benedetti et al 2013]**. These observational studies quickly led to the inclusion in lithium medication literature of GSK3B and related CREB pathways, which are widely distributed throughout all human cells that contain nuclei **[Valvezan et al 2012, Brown et al 2013, Chiu et al 2010, Can et al 2014]**. What is less widely recognized is that GSK3B is a potent regulator of ARNTL protein stability and circadian function through phosphorylation, as it does with other *CLOCK* genes **[Sahar et al 2010]**.

Since bipolar disorder is characterized by rapid cycling of mood, and by sleep disruption, it has always been an attractive hypothesis that members of the CLOCK gene family, whose proteins regulate circadian rhythmicity, must be key to dysregulation in this and other psychiatric disorders. In parallel, it has been of great interest to understand if lithium acts to establish normalization of circadian rhythmicity, which appears to be disrupted in major depressive disorder and other psychiatric disorders **[Crisafulli et al 2012]**. Extensive research has been dedicated to the role of CLOCK genes, bipolar disorder and lithium response **[McCarthy et al 2010, McCarthy et al 2011, Rybakowski et al 2014]**, although SNPs in these genes have not been found in any GWAS of bipolar disorder **[Welter et al 2014]**. However, in this study, SNPs in the *ARNTL* and *NR1D1* genes appear to contribute to enhance lithium drug efficacy in bipolar disorder.

This research clearly demonstrated that the *SLC1A2* (solute carrier family 1 member 2) gene, a sodium-amino acid symporter that acts in L-glutamate transport and neuroprotection, formerly termed the high affinity "glial" glutamate transporter, is expressed in neurons in human brain. Starting with the premise that adverse events that emerge from lithium treatment, such as fine tremor, can be better understood with knowledge of which variants have been significantly associated with essential tremor as a disorder in its own right, we included them in this study. Again, the strategy of evaluating GWAS SNPs associated with a drug's side effects in the context of examining pharmacogenomic response has been used previously by others **[Schulze 2012]**.

Discussion

This study implicates the involvement of a glutamatergic regulatory pathway in lithium response in bipolar disorder in individuals of European ancestry. The research builds on a foundation of published literature of candidate gene association studies that have examined lithium response and remission, including the identification of a SNP in the *GRIA2* (glutamate receptor ionotropic AMPA 2) gene found in a GWAS examining time to recurrence of lithium treatment in bipolar disorder **[Perlis et al 2009]**. The *GRIA2* gene encodes a transmembrane receptor that is part of a family of glutamate receptors that are sensitive to alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate (AMPA). It functions as a ligand-activated cation channel, and may form part of a channel that is assembled from 4 related subunits, GRIA1-4. The *GRIA2* gene has been previously associated with diagnosis and response to antipsychotic treatment in patients with schizophrenia **[Crisafulli et al 2012, Eastwood et al 1995]**, pharmacologic management of inflammatory

disorders involving chronic pain **[Kuttikat et al 2011],** and antidepressant response in major depressive disorder **[Chiesa et al 2012]**.

The results from the workflow we used in this analysis agreed with the results from annotation using data from the epigenome roadmap consortium. Although we used a SNP from a GWAS that examined time-to-lithium recurrence that did not achieve the stringent Bonferroni-type correction that has traditionally served as a threshold in GWAS to rule out false positives, this approach is increasingly being used to "rescue" GWAS associations in the context of pathway analysis. Thus, not only have recent GWAS have used a similar approach as the one used here to select less significant variants using pathway analysis **[Jia et al 2010, McGeachie et al 2014, Zhou et al 2013]**, but recent GWAS studies and guidelines emphasize the less stringent methods of FDR and Fisher's exact test as preferable to Bonferroni correction.

While it is true that our lithium response network does share some loci with results on the genetic etiology of bipolar disorder, such as the genes *ANK3* and *CACNA1C*, we feel strongly that it is not reducible to the genetic basis of the underlying disorder. For one, on a basic level, the math does not check out: some 80% of the genes in the network we report are not involved in bipolar etiology. But for another, it should be no surprise that some of the same genes whose mutations and regulation influence a disease may also play host to mutations (not necessarily the same ones) which influence the effectiveness of treatment. Examples of this phenomenon are forthcoming in disease, as for instance with BRCA1 and BRCA2, which play host to both mutations which predict the emergence of breast cancer, and also mutations that predict the effectiveness of breast cancer drugs. Our approach of focusing on those brain regions where lithium protects against neural

atrophy has been used by other investigators **[Philips et al 2014, Malhi et al 2013]**.

Of course, while these results admit of both counterarguments and affirmative defense against these counterarguments, such questions will ultimately be resolved by additional experiments in biological systems to examine the molecular components of lithium-mediated signal transduction which we have identified. While this analysis is based on tens of thousands of variants with demonstrated association with lithium response and adverse drug events in large human populations, this *in silico* analysis needs to be validated further in pharmacogenomic trials in human cohorts.

The current generation of genotype-based tests in neuropsychiatric pharmacogenomics have demonstrated clinical utility **[Mrazek 2010, Hall-Flavin 2013, Trinks et al 2014, Moaddeb et al 2013]**. Understanding the regulatory epigenome will provide the next generation of variants that will enable a new era of pharmacogenomic tests. These lithium variants and the methods that identified them are a promising method for finding candidate variants and loci of significance in epigenetic regulation of drug response.

Valproate: A Potent Initiator of Neurogenesis

Motivation

The mechanism of action of VPA as an anticonvulsant, mood stabilizer and analgesic in the human central nervous system (CNS) has not been adequately characterized. VPA and its derivatives exhibit a variety of effects which are cell- and tissue-specific, differ based on disease state, age, gender and ethnicity, and may be effective or deleterious. VPA is a weak blocker of sodium and calcium ion channels, and may inhibit key enzymes in the catabolism of gamma-aminobutyric acid (GABA), including ABAT (4-aminobutyrate aminotransferase) at physiologically relevant concentrations **[Ghodke-Puranik et al 2013]**. There is also evidence that VPA exerts its antiepileptic action through differential regulation of the *GRIN2B* (Glutamate Ionotropic Receptor NMDA Type Subunit 2B) gene **[Perucca 2002].** VPA is an effective inhibitor of histone deacetylases (HDACs), with an IC_{50} (0.4 mM) well within the therapeutic range of VPA (0.35–0.7 mM in serum). VPA causes robust chromatin decondensation, with a dspotent acetylation of core histones such as H3 and H4 that leads to activation of development gene expression **[Gottlicher et al 2001, Gurvich et al 2004].** VPA has been shown to be neuroprotective in animal models of traumatic brain injury (TBI) **[Halaweish et al 2015, Dash et al 2010]**, spinal cord injury **[Lv et al 2011]**, and in neurodegenerative disease **[Guo et al 2014].** In a swine model of TBI and hemorrhagic shock (HS) VPA decreases brain lesion size, improves neurologic recovery, and down-regulates genes associated with necrosis, apoptosis, and inflammation **[Dekker et al 2014]**. The HDAC inhibitor, sodium butyrate, with a mechanism of action similar to that of VPA, has been shown to activate neurogenesis in rodent brain following ischemic injury **[Kim et al 2009]**.

These results suggest different ways in which VPA may act in parallel to provide benefit in epilepsy, mood disorders, migraine, as well as recovery following trauma.

A recent study of the impact of VPA on the epigenome and gene expression in non-neuronal human cell lines demonstrated dose- and time-dependent up-regulation of TFs and cell type-specific histone modifications emblematic of poised, bivalent gene promoters active in cell fate specification and differentiation **[Halsall et al 2015]**. In the same study, the authors concluded that cells were actively buffering their homeostatic state in the presence of HDAC inhibitors to avoid hyper-acetylation that might lead to widespread and uncontrolled gene expression **[Halsall et al 2015]**. It appears that VPA-induced histone acetylation is not sufficient for chromatin decondensation, but rather a downstream effect of HDAC inhibition, suggesting that the drug suppresses the expression of proteins involved in maintenance of heterochromatin and/or uses chromatin remodeling proteins as intermediaries **[Yi et al 2013, Marchion et al 2005]**. VPA has been shown to direct the programming of fibroblasts into neurons **[Chu et al 2015]**. It is also routinely used to remove histone methylation in cellular domains such as the lamina-associating domain (LAD) located just interior of the nuclear membrane, a region containing heterochromatin, in which exposure to VPA abolishes H3K27 and H3K9 methylation **[Kelkhoff et al 2016]**. It is noteworthy that others have documented the presence of a "genomic storm" following TBI **[Xiao et al 2011]**. In addition, stem cell therapy has been heralded for treatment of trauma patients, including those with TBI **[Ahmed et al 2016]**, suggesting that cell fate reprogramming with development TFs might also be a useful approach in such cases.

Recent studies of the epigenomic control of gene expression that has identified distinct mechanisms through which chromatin interactions mediate transcriptional programs involving topologically-associating domains (TADs), enhancer-promoter loops and actively transcribed regions of the human genome characterized by the histone mark H3K27ac **[Kundaje et al 2015, Rao et al 2014, Gonzalez-Sandoval et al 2016]**. It has been shown that SNPs which disrupt the boundaries of TADs cause serious health problems **[Lupianez et al 2015**]; and causal SNPs exhibit significant allele bias in open chromatin **[Lee et al 2015]**. In addition, the greater the number of spatial connections a given enhancer or promoter maintains genomewide indicates both the potency of the regulatory element and its validity **[Ramani et al 2016]**, especially for CNS genes that are involved in coordinated transcriptional programs of neurogenesis and neuroplasticity **[Thakurela et al 2015]**.

Our hypothesis is that VPA acts through transcriptional activation and repression of specific genes resulting in chromatin-mediated neurogenesis and neuroplasticity in the adult CNS and inhibition of glial scarring. To determine the validity of this premise, we analyzed experimental and public data. We then reconstructed a gene regulatory network that mediates VPA's mechanism of action in human brain, and found that the drug exerts widespread effects in adult brain including a transcriptional program of neurogenesis and neuroplasticity, involving TFs that have been shown to program neuronal cell fate commitment and suppress oligodendrocyte cell fate.

Methods

Data

Figure 2-12 shows an overview of the experimental design, including public databases and sources used in this analysis**.** These included experimental results from a well-characterized swine model of TBI and hemorrhage **[Halaweish et al 2015, Dekker et al 2014],** that exhibits functional recovery following a single supra-therapeutic bolus of VPA evaluated for differential gene expression following VPA administration. Other primary data included microarray expression data from postmortem human brain tissue obtained from the Human Brain Atlas of the Allen Brain Science Institute **[Sunkin et al 2013]** and other sources **[Higgins et al 2015]**.

Experimental data from SK-N-SH cell lines

Publicly available date from SK-N-SH cells were used for two applications: (1) For determination of DNase I hypersensivity and allele bias, as they are included as examples in the deltaSVM machine learning algorithm **[Lee et al 2015]**, and (2) Evaluation of spatial interactions in the human genome using data from a high resolution Hi-C dataset **[Rao et al 2014]**. To discover how the VPA pharmacogenomic SNPs we selected produced robust pharmacogenomic stratification, evaluation of the chromatin interactions of regulatory elements they were located in was used to preliminarily map the putative VPA pharmacodynamic pathway in human brain.

Published literature on pharmacogenomic associations and VPA targets and pathways

Two hundred and fifty-four peer-reviewed published articles were retrieved between January 1 and October 1 2016, including gene association studies, as well as basic and clinical pharmacology reports. An automated Boolean search string was used consisting of "valproate OR valproic acid

OR sodium valproate OR divalproex sodium AND pharmacodynamics OR mechanism of action OR brain OR pathways AND pharmacogenomic AND bipolar disorder AND epilepsy AND migraine headache AND gene OR SNP AND association AND human." There are no published genomewide association studies (GWAS) that are focused on medication response or adverse events related to only valproic acid therapy in epilepsy, bipolar disorder or migraine. There are however, several GWAS that have investigated SNPs associated with human populations that have treatment-resistant or refractory epilepsy. These studies are made up primarily of patients who do not respond to a combination of anticonvulsant drugs or worsen on therapeutic regimens. Due to confounding related to polypharmacy these studies were excluded. For this study, we avoided anything but primary research studies in which we re-tested all of the parameters and statistical tests that were employed. This dramatically reduced the number of studies from 254 to 52 that could be included in this analysis.

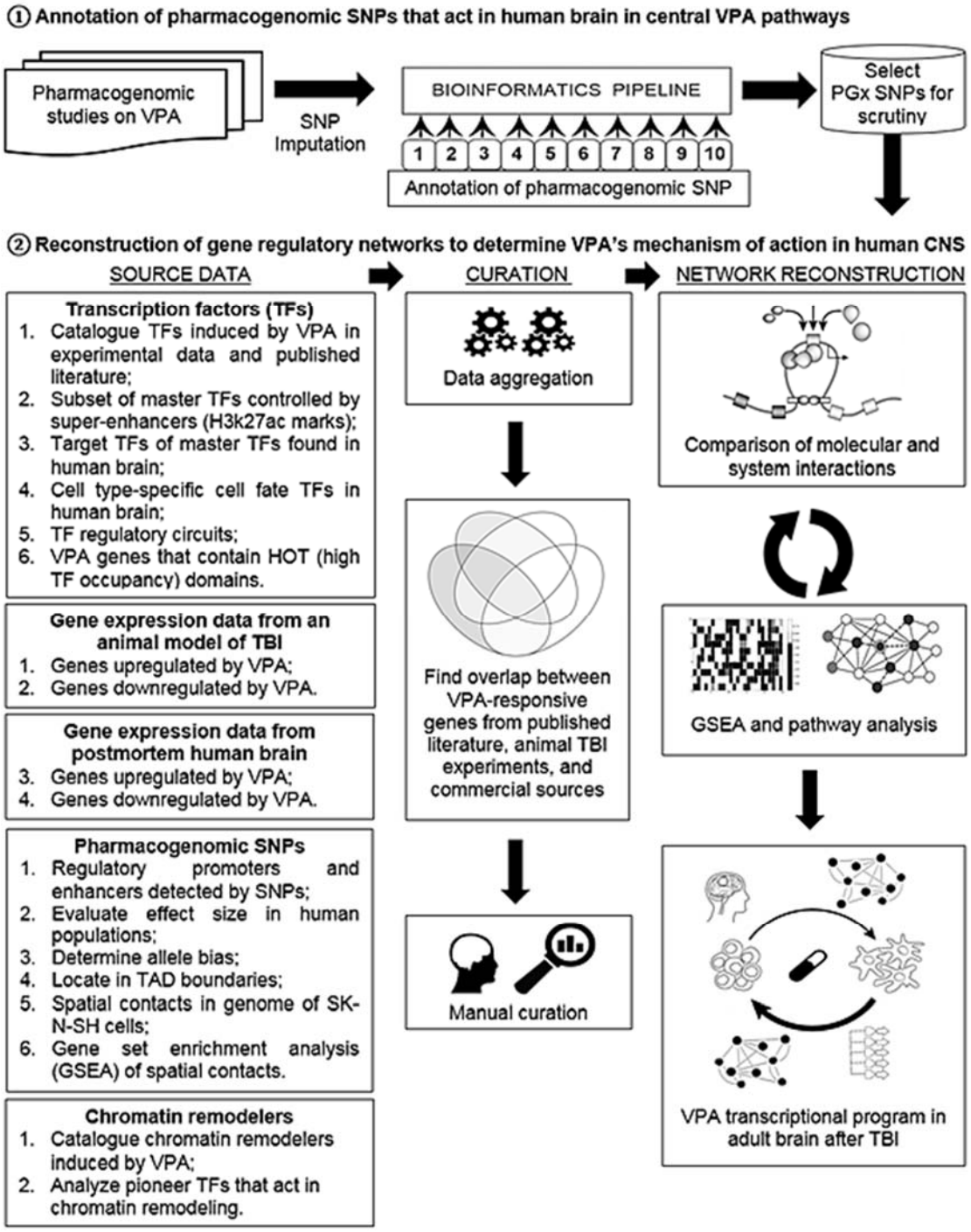


Figure 2-12. Diagram of Valproate Analysis Methods

(1) Shows steps to annotate putative VPA pharmacogenomics SNPS which were **(2)** mechanistically analyzed using integrative bioinformatics and regulatory network analysis

methods**.** In **(1)** numbers refer to the different annotation steps of the bioiformatis pipeline. Details can be found in the text.

Analysis

VPA Pharmacogenomic SNPs

Pharmacodynamic variants that stratify drug response in human populations have provided insight into CNS targets and pathways for neuropsychiatric medications that have been in clinical use for decades but have ill-defined mechanism of action in human brain **[Ghodke-Puranik et al 2013, Perucca 2002]**. The published literature on the pharmacogenomics of valproic acid response in human populations has been sparse, constrained by small sample size, confounded by an over-emphasis on exon variants, and a focus on pharmacokinetic genes. To date, genome-wide association studies (GWAS) have examined epilepsy and anticonvulsant-induced adverse events but have not included VPA in their analysis except in combination with other anticonvulsant medications.

To better understand human CNS pathways that are involved in the mechanism of action of valproic acid, we combined SNP imputation with bioinformatics analysis. SNPs that included in the analysis had to be located in genes expressed in human brain, were likely to be associated with known biological targets of VPA, could be functionally annotated, and were derived from primary research studies. For this analysis of potential pharmacodynamic pathways, we excluded pharmacokinetic genes and their variants such as the UGT and CYP super-families. To select VPA pharmacogenomic SNPs we combined SNP imputation from VPA pharmacogenomic association studies, and performed computational and bioinformatics analysis as described in detail in a previous publication **[Higgins et al 2015]**. DrugBank **[Drugbank 2015]** and the

pharmacogenomic mutation database (PGMD®; Qiagen) lists pharmacodynamic targets that have been associated with VPA **[Kaplun et al 2016]**. Studies were selected from the published literature as described in a previous study **[Higgins et al 2015]**. In addition, we interrogated the PGMD® (Qiagen GmbH) **[Kaplun et al 2016]** and DrugBank **[Drugbank 2015]** to gather additional information.

In **Figure 2-12.1**, bioinformatics analysis was used to identify the most probable causal SNP from the association regardless of the lead SNP(s) that were reported, and included the following bioinformatics analysis to determine the functionality of the variant: (1) Location in open chromatin as indicated by peaks of DNase I hypersensitivity, (2) Low to moderate methylation of any cytosine residues, (3) The presence of histone marks that indicate regulatory function (H3K27ac + H3K4me1 = enhancer; H3K27ac + H3K4me3 = promoter), (3) The location of the variant within the context of the gene or in an intergenic domain, (4) Whether the regulatory element has yet to be annotated as a molecular quantitative trait locus (eQTL, hQTL, etcetera), (4) Proteins, including transcription factors, bound to the regulatory element as determined by ChIP-Seq indicating its regulatory function, (5) Disruption by the pharmacogenomic SNP of transcription factor binding sites as indicated by alterations in the position weight matrix, (6) Association of the regulatory element with the requisite RNA species (e.g., bi-directional enhancer RNA=enhancer; mRNA= promoter), (7) Connectivity of the regulatory element with other elements in the genome as indicated by the Hi-C chromatin conformation capture method limited to SK-N-SH cells, which are the ENCODE Tier 2.5 neural surrogate cell line, (8) Transcriptional programming by factors which are responsible for determination of neuronal cell fate, (9) Determination of the allele bias of the allelic variant using the deltaSVM algorithm **[Lee et al**

2015], (10) examination of the neuroanatomical distribution of target gene expression data in postmortem human brain using both microarray and *in situ* hybridization data from the Human Brain Atlas of the Allen Brain Science Institute **[Sunkin et al 2013].** This process revealed a total of 3 pharmacodynamic SNPs that warranted further investigation, including rs3764028_G located in the promoter of *GRIN2B*, rs2857654_A which detects an enhancer and is located in *CCL2*, and rs2269577_G, an enhancer located 5' to *XBP1*. These were chosen because the objective was to provide a more detailed characterization of these SNPs and their corresponding regulatory elements to better understand VPA's mechanism of action in the human CNS (**Figure 2-12.2).**

For annotation, we focused on active enhancers, promoters and transcription start sites, so other chromatin states were not used in the present study, including flanking and poised regulatory elements, transcribed domains or repressed elements. Although potential regulatory variants may be repressed in one tissue or brain region but not another, we could directly assess function in any biological system in this study, so we relied on the data provided by the association studies. We harmonized our nomenclature with the results of the roadmap epigenome mapping consortium. In all cases, gene and protein symbols and definitions were consistent with HGNC nomenclature **(1)**, except as indicated by citation in the text. Gene symbols are from GENCODE except where indicated, and genomic location coordinates are from build 38 of the Human Genome.

SNP	GENE[1]	TYPE	deltaSVM SCORE[2]	DISEASE	EFFECT	EFFECT SIZE
rs2857654_A	*CCL2*	Enhancer	-2.275704	Epilepsy	Response in children	1.45 (1.06-1.99)
rs3764028_G	*GRIN2B*	Promoter	-5.097029	Epilepsy	Dose range	1.7553 (1.219-2.291)
rs2269577_G	*XBP1*	Promoter	4.710044	BPD	Response[3]	1.2754 (0.329-2.221)

Figure 2-13. Reported SNPs from the Valproate Analysis

Selected Pharmacogenomic Variants Selected for Allelic Variation in VPA Dose, Response and Pharmacodynamics in Human Populations. Noncoding SNPs that identified regulatory elements, including enhancers and promoters. SNPs included rs2857654 **[He et al 2013]**, rs3764028 **[Hung et al 2011]**, and rs2269577 **[Kim et al 2009]**. [1]RefSeq nomenclature; [2]deltaSVM is a machine learning algorithm that determines the causal nature of gene variants, including DNase I hypersensitivity and allele bias **[Lee et al 2015]**; [3]Total treatment response score, Kruskal–Wallis test for valproate prophylactic treatment response. BPD: Bipolar disorder, sample containing patients with BPD 1 and BPD 2.

Reconstruction of VPA Gene Regulatory Networks in the Human CNS

Figure 2-12.2 shows the method by which we reconstructed the central VPA pathway in the human CNS. It involved: Identification of databases relevant to modeling the VPA regulatory circuit, Data aggregation from public databases and in-house data, manual curation of the data in the context of a VPA central mechanism of action, preparation of curated data for model integration, iterative gene set enrichment analysis (GSEA) using different public and commercial software, and pathway analysis using IPA® and the STRING database of protein-protein interactions **[Szklarczyk et al 2014]**, as well as network analysis using weighted gene co-expression network analysis (WGCNA) in R **[Langfelder et al 2008]** and node-edge modeling systems modeling using Python **[Kestler et al 2008]**. In addition to applying redundant software analysis tools, we also used different open source and commercial databases, including IPA® **[Kramer et al 2013]**, Pathway Commons **[Cerami et al 2011]** and Reactome **[Croft et al 2014]** databases as well as manual curation of the scientific literature, to determine network interactions.

For reconstruction of VPA's gene CNS regulatory pathway, we used a hybrid model development method that combines GSEA and pathway analysis/network modeling software used in bioinformatics with a technique for constructing core regulatory circuitry which includes super-enhancers, TFs and auto-regulation, based on defined biological attributes of transcriptional regulation in the human CNS**.** During the course of the project, we discovered a number of other TFs that are regulated by VPA that directs cell fate, in addition to many genes whose was expression was regulated by ASCL1 (Achaete-Scute Family BHLH Transcription Factor 1), NEUROD1 (Neuronal Differentiation 1), MEF2C (Myocyte Enhancer Factor 2C), MEF2D

(Myocyte Enhancer Factor 2D), MYT1L (Myelin Transcription Factor 1 Like), and TBR1 (T-Box, Brain 1). Many TFs were tightly coupled in terms of target promoters, many of which were also involved in neurogenesis and neuroplasticity. As such, this convoluted super-program was investigated in greater detailed. Since many of these genes were also regulated by the same super-enhancers or enhancers, we gained better insight into underlying coordinated mechanisms of transcription induced by VPA. To do so, emphasis was placed on cellular reprogramming of neuronal cell fate and underlying mechanisms of coordination that supported VPA-mediated chromatin reorganization leading to functional recovery.

Results

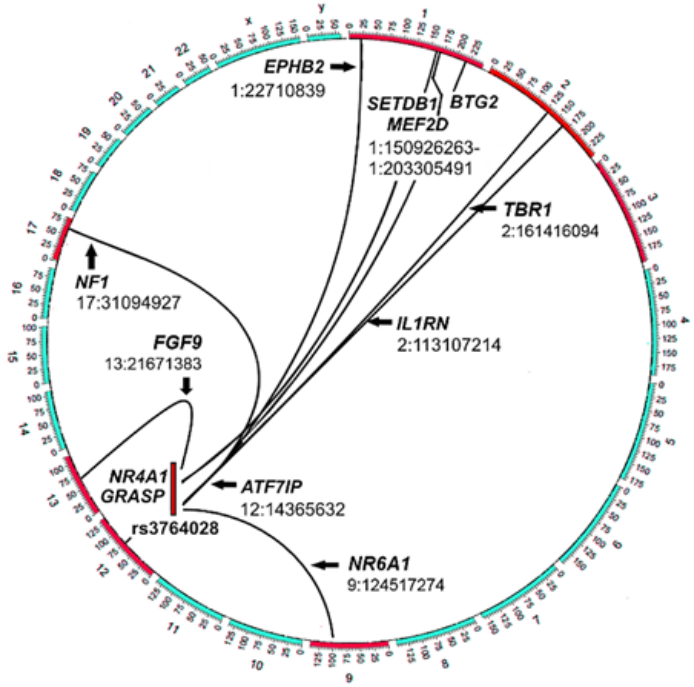
Selection of VPA Pharmacogenomic SNPs for Further Examination

Table 2 shows VPA pharmacogenomic SNPs that were selected for further analysis. The 3 SNPs stratify response to VPA in human populations, are located in regulatory domains (promoters or enhancers), and exhibit significant chromatin allele bias as measured by the deltaSVM algorithm (23). The latter is a feature of causal SNPs **[Lee et al 2015]**. They include:

1. The intronic SNP rs2857654_A located within the *CCL2* gene, which encodes the Chemokine C-C Motif Ligand 2, and is most significantly associated with cell movement and migration of cells and this enhancer interacts with VPA **[Kramer et al 2013]**;

2. The 5' SNP rs3764028_G located in the distal *GRIN2B* promoter is in a region associated with massive chromatin reorganization in brain in rodents and human cell lines. Spatial mapping in SK-N-SH cells shows that it forms an inter-chromosomal transcriptional hub (Figure 3A), contacting genes in *cis*- and *trans*- in SK-N-SH cells that are enriched for genes involved in neuronal differentiation and development of the CNS **(Figure 2-14)**;
3. The 5' SNP rs2269577_G located in the promoter of the *XBP1* (X-Box Binding Protein 1) gene, which functions as a transcription factor during endoplasmic reticulum (ER) stress by regulating the unfolded protein response. Required for cardiac myogenesis and hepatogenesis during embryonic development, and the development of secretory tissues such as exocrine pancreas and salivary gland. An enhancer co-regulates *XBP1*, *EWSR1* (EWS RNA Binding Protein 1), *CCDCC117* (Coiled-Coil Domain Containing 117), *KREMEN1* (Kringle Containing Transmembrane Protein 1) and *ZNRF3* (Zinc And Ring Finger 3) genes, which are co-localized within a TAD located on chromosome 22 **[Wang et al 2016]**. The location of this SNP 5' to the *XBP1* gene is shown in **Figure 2-14**.

A



B

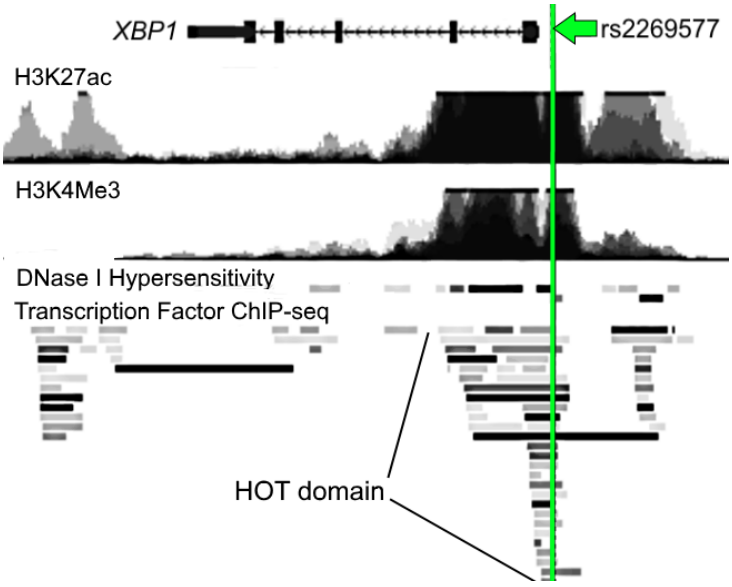


Figure 2-14. Spatial Contacts of Valproate SNPs

(A) Whole genome plot of transcriptional hub based on *cis*- and *trans*-contacts of the SNP rs3764028_G in the promoter of the *GRIN2B* gene based on data from Hi-C mapping of this SNP

in the human neuronal cell line SK-N-SH. **(B)** Relative location of the VPA pharmacogenomic SNP rs2269577 (green line), with histone marks H3K27ac and H3K4me4 indicative of an active promoter, location in a DNase I hypersensitivity region indicative of open chromatin, and overlapping a HOT domain containing many TFs **[Li et al 2016]**. This SNP is located withinh a TAD located on chromosome 22 (26,600,000-28,000,000) **[Wang et al 2016]** containing the *XBP1* gene as well as other genes. Inset. Screenshots from the UCSC genome browser, build hg18, taken from the UCSC genome browser in **(B) [Speir et al 2016]**.

VPA Modulates a Transcriptional Hub Consisting of 12 Genes in Spatial Contact with the *GRIN2B* Promoter

The 5' SNP rs3764028_G located in the distal *GRIN2B* promoter is in a region associated with massive chromatin reorganization in brain **[Thakurela et al 2015, Bharadwaj et al 2014]** in rodents and human cell lines. This SNP detects a promoter that maintains spatial contact, as determined by Hi-C maps of SK-N-SH cells, with 3 known enhancers including 2 super enhancers as well as a promoter with spatial contacts to genes associated with neuronal survival and plasticity. This replicates a result from a study in mouse brain in which researchers combined genome-wide analysis of data sets for chromatin accessibility (FAIRE-Seq) and the enhancer mark H3K27ac, in which they found a subset of genes associated with neuroprotection and plasticity that increased transcription in adult mouse brain following activation of the glutamate receptor **[Bharadwaj et al 2014]**. In our study, both gene set enrichment using Gene Ontology, as well as manual inspection of genes that maintain spatial contacts with the promoters and enhancers detected by our putative causal SNPs demonstrate selectivity for neuroplasticity and chromatin reorganization. In agreement with their findings, genes whose abundance increased included *NR4A1*, which regulates dendritic spindle density and organization in human brain following neuronal excitation, *BTG2*, a transcription factor that inhibits neural precursor cell proliferation and stimulates neuron cell differentiation and acts in histone arginine methylation, *ILRN*, which is also in contact with an enhancer associated with SNP rs2857654_A, and *NF1*, which differentially controls neural stem cell proliferation. In addition, we identified contacts with *GRASP*, part of a receptor complex scaffold that regulates G protein-coupled glutamate receptor signaling, *MEF2D*, which is regulated by NEUROD1, and *EPHB2* (EPH Receptor B2), a developmentally-regulated receptor tyrosine

kinase that functions in axonal guidance during development. This promoter also maintains long distance spatial contact with *SETDB1*, an H3-K9 histone methyltransferase that regulates epigenetic gene silencing to maintain stem cell pluripotency, a result that has been experimentally shown in rodent brain and cell lines by other researchers **[Thakurela et al 2015, Bharadwaj et al 2014]**.

The distal *GRIN2B* promoter maintains spatial interactions with genes that program neuronal cell fate in humans, and are highly responsive to VPA. These are *TBR1* (T-Box, Brain 1), which is up-regulated by VPA following TBI in our animal model. TBR1 is a TF that acts as a potent programmer of neurogenesis **[Mihalas et al 2016],** controls the differentiation of pyramidal cells in neocortex, and controls expression of the *GRIN2B* gene in developing cerebral cortex **[Notwell et al 2016].** The other is *FGF9* (Fibroblast Growth Factor 9), which is produced by developing neurons to maintain homeostasis within the surrounding milieu. It plays an important role in the regulation of embryonic development, cell proliferation, cell differentiation and cell migration, regulation of gliosis during repair and regeneration of brain tissue after damage and the differentiation and survival of neuronal cells **[Hadjab et al 2013].** Several of the genes that exhibit spatial proximity to the *GRIN2B* promoter in human SK-N-SH cells are also up-regulated by VPA in the swine model of TBI and hemorrhage. These include the following:

1. *FGF9* (Fibroblast Growth Factor 9): 1.74 fold ($\log_2$) (p-value = 0.005; t-statistic = 5.76);
2. *NR6A1*: 1.75 fold ($\log^2$) (p-value = 0.0000546; t-statistic = 15.77);
3. *TBR1*: 1.49-fold ($\log^2$) (p-value = 0.007; t-statistic = 4.73)
4. *MEF2D*: 1.11-fold ($\log^2$) (p-value = 0.005; t-statistic = 5.42).

Synthesis

With the successful completion of three variant discovery and analysis experiments in three separate drug-disease systems (and system agglomerations) in early 2015, it was clear within the Athey laboratory that such methods added significant value to the discovery of both variants and mechanisms for pharmacological phenotypes across the neuropsychiatric space.

However, at this point, the epigenome variant discovery effort was handicapped by serious problems, including a lack of consistency in methods between and within our variant discovery efforts, and a resulting lack of reproducibility, which were of concern to journals, peer reviewers, and our partners at Assurex Health. In addition to this, variant discovery experiments in this vein took a long time to conduct and to update in light of new information, and made extensive use of expert judgment in areas like neuroanatomy and tissue relevance, as well as thresholding, making them difficult to perform at scale. Finally, research then breaking in the area of epigenome informatics and the spatial genome was then suggesting new information modalities for evaluating regulatory variants, which could not be undertaken in the semi-manual mode we had then adopted.

The stage was then set for an automated pipeline which would synthesize the lessons of our manual variant discovery experiments into a broadly applicable featureset, to be run reproducibly on locally stored data, which would resolve these problems and lay a foundation for future development. I embarked upon the creation of what would eventually become the Pharmacoepigenomics Informatics Pipeline.

Chapter 3: The Pharmacoepigenomics Informatics Pipeline

This chapter described the Pharmacoepigenomics Informatics Pipeline, an integrative multiple omics variant discovery pipeline I created in order to make epigenome analyses in pharmacogenomics more powerful and easier to perform. In its initial iteration, which replicates closely the methods of the lithium analysis **[Higgins et al 2015]**, it has both replicated existing results and discovered previously unknown genetic and mechanistic basis for study phenotypes. In addition, the success of the PIP with warfarin, a system far outside of neuropsychiatry in a domain which has been the subject of a great deal of pharmacogenomics work which has borne little clinical fruit, is an indication that these epigenome methods and pipelines may add value in a variety of phenotype systems.

Methodological Guidance from Pre-PIP Analyses

The Pharmacoepigenomics Informatics Pipeline was conceived with the object of enabling epigenome analyses to discover variants for phenotypes, of the type demonstrated in the preliminary work, to be carried out with the benefit of three principal improvements:

Firstly, that they be easier to undertake and to maintain. The epigenome analyses in the "variants," **[Higgins et al 2015]** "lithium," **[Higgins et al 2015]** and "valproate" **[Higgins et al 2017]** papers, described in chapter two, each took months of effort, and involved manually downloading and processing relevant datasets, as well as consulting online resources. Any update to the analysis, e.g. including more variants or taking advantage of a new dataset, would require a similar scope of effort. To scale to larger experiments, be available to collaborators, and be used more widely, this type of analysis needed to be more automated, so that input files could be assembled and run with less investigator effort.

Secondly, that the methods be more standardized and the results more reproducible. The three epigenome analyses described in chapter two each had a separate featureset containing different methods, and each contained many manual steps, including consulting external databases that change in an unversioned manner. Thus, if such analyses continued to be carried out in such an ad-hoc manner, their results would not be directly comparable, not permitting all kinds of comparative measurements and judgments one might wish to make. In addition, they would not necessarily be reproducible, either by other groups or by one's own group over time.

Thirdly, that future such analyses be able to benefit from advanced features not contemplated in current analyses. As contemplated in chapter one, there was an awareness that the use of coimputation statistical approaches in scoring, as well as machine learning methods for variant effect and target gene identification, and advanced methods for processing spatial data, would enhance the discerning power of such analyses in the future, but many such methods are ill-suited for semi-manual processing and ad-hoc methods development.

With these three aims in mind, I undertook the formulation of a phased feature set for the Pharmacoepigenomics Informatics Pipeline, in three primary phases. First, a minimal featureset designed to replicate the methods and reproduce the results of the lithium analysis. Second, a tractable nearterm featureset within the context of my Ph.D. thesis; the version contained in this chapter. And third, a longer term vision for the PIP as the foundation of third generation pharmacogenomics, containing advanced spatial genome and machine learning methods, with coimputation scoring, to occur subsequently.

This construction was based on a number of important lessons from the prior analyses, including:

Population specific Linkage: All three analyses benefited from the use of linkage analysis to expand the set of prospective SNPs beyond merely the lead SNPs of the source GWAS, and located potent regulatory SNPs, with important target genes, which were not reported lead SNPs. The "variants" and lithium analyses benefited from the use of population-specific linkage analysis (using a similar genetic background to the original study populations) to uncover linked variants which did not exhibit linkage in a global population.

Joint analysis of GWAS: The lithium and valproate analyses benefited from the analysis of GWAS for related phenotypes, which frequently show joint hits. See the discourse on joint GWAS analysis in chapter 2.

Regulatory variants and chromatin state: All three analyses exhibited a predominance of regulatory variants over coding variants among the outputs of the analysis, including the observation that the outputs were predominantly located in areas of promoter and enhancer chromatin states with concordant chromatin marks, DNase accessibility, and TF binding.

Tissue specific analysis: All three analyses benefited from the use of tissue specific omics information to clarify the chromatin states around candidate variants in the particular tissue involved. Analyses based on less relevant tissues will inherently be less accurate.

Importance of getting datasets: All three analyses involved the manual collection of relevant datasets, a laborious and imprecise process.

Variant dependence: The "variants" analysis demonstrated conclusively both that regulatory variants identifiable as causal frequently exhibit allele-specific conformity with position weight matrices (PWMs) for transcription factors, and that they frequently exhibit concordant binding with the relevant transcription factors. Allele-biased PWM conformity was a portion of the scoring function in the lithium analysis.

Bimodal target genes with QTLs and spatial mapping: Although the lithium analysis only used sequence proximity as a means of identifying target genes, the "variants" and valproate analyses showed that the most potent effector variants had target genes throughout the genome, which could be identified by both molecular QTL mapping and spatial contacts.

Importance of pathway mapping: Both the lithium and valproate analyses benefited from using pathway connections to trim, rank, and organize putative candidate variants and their identified target genes.

Thus, this set of important conserved principles was identified as critical to the success of the new pipeline, and each was included in the design of the PIP featureset.

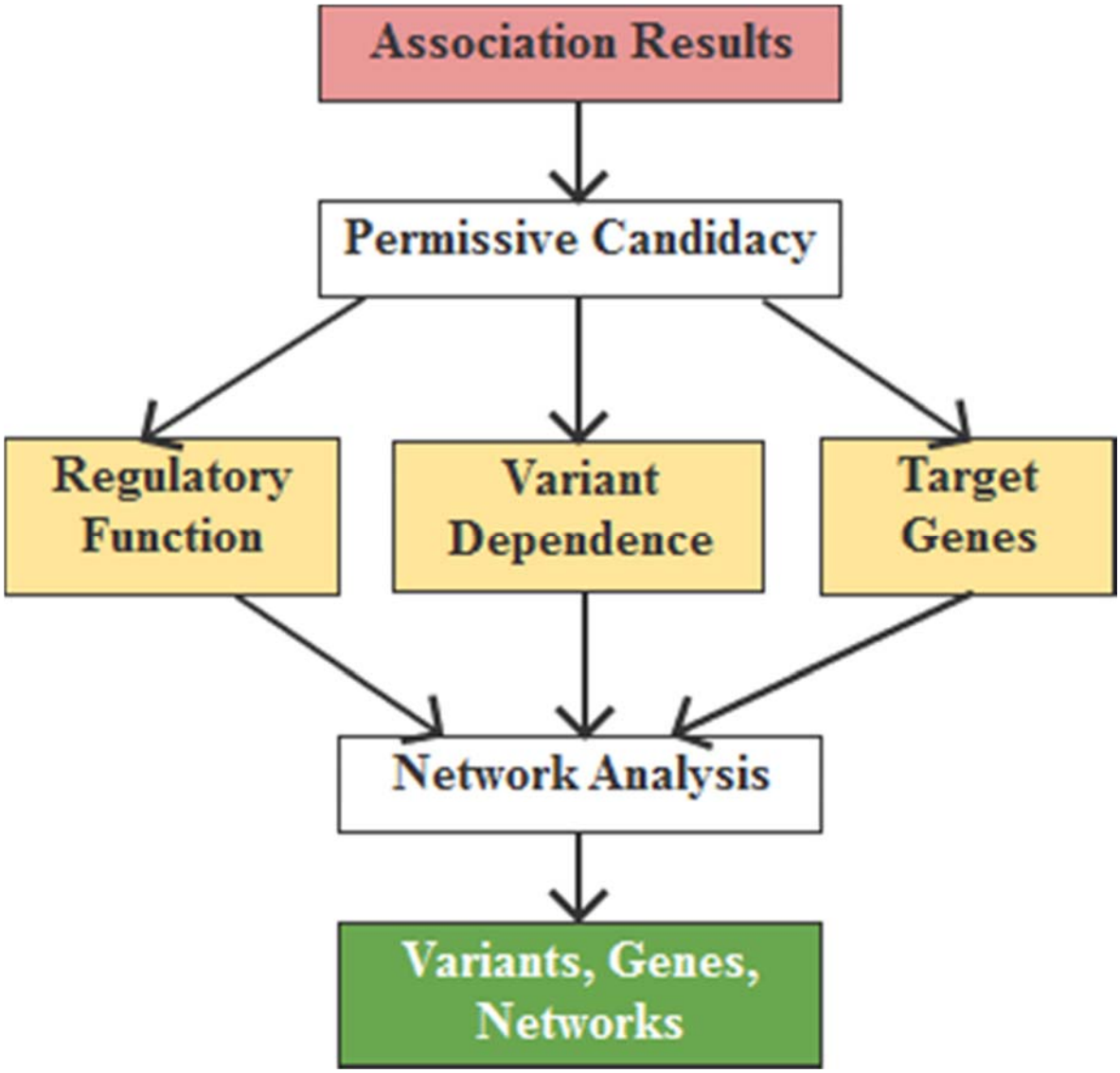


Figure 3-1: The Five Box Model

The Five Box Model

During the development of the PIP featureset in 2015, I began to conceptualize the function of such a multiple omics variant pipeline with what I refer to as the five box model **(Figure 3-1)**, reflecting a conceptual scheme of five general properties a variant may exhibit which provide evidence, individually and collectively, that it may be a causal regulatory variant in a phenotype. They are permissive candidacy, regulatory function, variant dependence, target genes, and network analysis.

By permissive candidacy, we mean that the variant has in some way come to the attention of a genome-wide screen: that it is located in a populations-specific linkage disequilibrium with a variant associated with a phenotype of interest, or that it regulates a gene whose mechanistic importance to the phenotype of interest has already been established.

By regulatory function, we mean that the variant is resident in a portion of the genome which is a regulatory element, a promoter or enhancer, in one or more of the particular tissues which are relevant to the drug-disease system.

By variant dependence, we mean that the function of this regulatory element must be dependent on the status of that variant, through the alteration of sequence features which help to determine the epigenome. Such sequence features may be a specific binding site for a transcription factor, but may also be a more general propensity score for an epigenome feature, as determined by an appropriate bioinformatics algorithm such as a learning machine.

By target genes, we mean the variant have identifiable target genes with which it is spatially and/or functionally associated, putatively whose expression it regulates.

And by pathway analysis, we mean that we may say of a collection of putative genes and variants for a phenotype or phenotype cluster, that taken as a totality they are associated with each other and with ontologic and/or network categories which are connected to the phenotype under investigation.

These five concepts may be evaluated in different ways under different circumstances: with different datasets, with different algorithms, manually and under automation. Nevertheless they are conceptually durable, and all the pre-PIP workflows, each extant version of the PIP, and our future plans for PIP-successor pipeline, fall within this overall orienting framework.

Motivation and Plan of Action

Pharmacogenomics is Poised for Transformation by the Epigenome

Pharmacogenomics variant discovery has prioritized the search for protein coding variants with genetic association and biochemical methods **[Black et al 2007]**. With the sequencing of the human genome, the initial hope for immediate discovery of highly penetrant coding variants for many phenotypes **[Collins et al 2001, Ganguly et al 2001]** did not bear out, both in pharmacogenomics and in many other fields **[Weinshilboum et al 2004]**. And with the advent of

GWAS (genome wide association studies), pharmacogenomic phenotypes began to be investigated with this powerful new tool **[Giacomini et al 2017]**. While most available pharmacogenomics tests were and are based on highly penetrant coding variants, GWAS studies have revealed that the bulk of genetic variation in drug response, like many other phenotypes, comes from noncoding regulatory variants which are cooperative and combinatoric **[Boyle et al 2012]**.

The development of the broader field of genomics and genomic regulation has been extremely fruitful in the last decade, but the deployment of GWAS in pharmacogenomics variant discovery has lagged deployment in other disciplines **[Giacomini et al 2017]**, epigenomic interpretation of GWAS results has been underutilized **[Nishizaki et al 2017]**, and the translation of GWAS results into clinical tests has been slower still in most areas **[Florez et al 2017, Giudicessi et al 2017, Fabbri et al 2016]**. Over the last decade, genome-wide methods have produced increasingly deep and broad atlases of both the genome and epigenome. Genomic atlases have included HapMap **[International Hapmap Consortium 2007]** and 1000 Genomes **[1000 Genomes Project Consortium 2015]**, while epigenome atlases have included ENCODE phases 1 **[ENCODE Project Consortium 2007]** and 2 **[Kellis et al 2014]**, the Roadmap Epigenome Mapping Consortium **[Roadmap Epigenomics Consortium 2015]**, the International Human Epigenome Consortium **[Stunnenberg et al 2016]**, and the upcoming Human Cell Atlas **[Regev et al 2017]**. Along with this, potent new techniques for probing the spatial and temporal genome have emerged, including time series omics analysis and spatial genome methods like Hi-C **[Rao et al 2014]**.

Analysis of data from these atlases and targeted experiments has produced a new paradigm of regulation in which the spatial genome and epigenome take a prominent role **[Roadmap**

Epigenomics Consortium 2015, Rao et al 2014, Cremer et al 2015]. In this paradigm, gene transcription is the result of the convergence of a series of spatial and functional factors. It takes place in chromatin loops at the spatial union of topologically associating domains (TADs) containing genic promoters with TADs containing controlling enhancers and co-regulated genes in a transcription factory **[Rao et al 2014]**. Such factories assemble prior to the initiation of transcription **[Hub et al 2017, Krijger et al 2017]**, at the edges of chromosome territories **[Rao et al 2014]**, in TADs localized to interior of the cell nucleus **[Cremer et al 2015]**. This process occurs with the aid of activating transcription factors and epigenomic marks **[Roadmap Epigenomics Consortium 2015]**, in a dynamic **[Chen et al 2015, Chen et al 2017]**, cell type specific **[Rao et al 2014]** manner, with different enhancers active in different cell types and tissue microenvironments, while some act in many or all cell types and tissues.

This model, which has developed along with increasingly powerful epigenomic measurement techniques **[Roadmap Epigenomics Consortium 2015]**, is yielding increasing predictive power and new insight in many areas. Genome-wide, multi omics epigenomes (i.e. those which combine a number of complementary genome wide epigenomic measurements) are now available on a genome wide basis in many cell lines and tissues **[Roadmap Epigenomics Consortium 2015]**, and even in multiple physiological conditions in many cases. As a result of this, the epigenome is increasingly regarded like the reference genome: as a resource to be consulted for systems and loci of interest **[Regev et al 2017]**, rather than an unknown quantity to be queried experimentally in specific contexts.

This transition is rapidly revolutionizing variant discovery with the emergence of integrated multi omics pipelines for variant discovery. These pipelines, which include HaploReg **[Ward et al 2016]** and RegulomeDB **[Boyle et al 2012]**, screen association regions from the genome for variants bearing important marks of epigenomic regulatory variants. These methods have allowed investigators to begin investigating at scale the over 90% of associated variants which are noncoding regulatory variants **[Boyle et al 2012]**, and to get new and increasing value out of GWAS studies whose results were initially enigmatic **[Fagny et al 2016, Tak et al 2015, Farh et al 2015]**.

The Pharmacoepigenomics Informatics Pipeline (PIP): Multi-Omics Variant Discovery for Pharmacogenomics

However, in pharmacogenomics these insights have not as yet been broadly applied. Despite recent work showing that an epigenome, 4D Nucleome based variant discovery and validation approach can yield value in excess of traditional methods **[Higgins et al 2015, Higgins et al 2015, Higgins et al 2017]**, this new era of "pharmacoepigenomics" **[Higgins et al 2015, Higgins et al 2017]** is still in its infancy. Indeed, much pharmacogenomics variant discovery still proceeds along traditional lines involving the search for coding variants to be designated as star (*) alleles, with a particular emphasis on pharmacokinetic (PK) genes for absorption, distribution, metabolism, and excretion (ADME). Such genes have formed the focus for test development for response, dosing, and adverse drug events (ADEs) and adverse drug reactions (ADRs).

To aid the wider dissemination of these advanced methods in pharmacogenomics, we have created the pharmacoepigenomics informatics pipeline (PIP), an integrative multi-omics variant discovery pipeline designed specifically for next generation pharmacogenomics and pharmacoepigenomics. It combines omics datasets (including epigenome compendia and phenotypically driven GWAS) with domain-specific knowledge (including key tissues and known significant genes). The PIP's variant discovery strategy is based on two separate workflows: a coding variant (CV) workflow based on traditional methods, and an expression regulatory (ERV) workflow that integrates bioinformatics algorithms and omics datasets to screen and organize regulatory variants. These variants and their target genes are organized together at the end of the analysis to provide the input for downstream statistical analysis, gene-set enrichment, and pathway analysis to clarify the genotype-phenotype relationships.

The ERV workflow is carried out with an overall model for combining multiple omics with preexisting knowledge about drugs and disease (i.e., relevant tissues and candidate variants) to discover variants and pathways that have a causal influence in chosen phenotypes. It begins with lead SNPs from association studies and expands them in a population-specific manner with linkage mapping. After this, tissue specific omics datasets are used to evaluate the regulatory function of the genomic regions around the variants, the dependence of that function on the status of the variants, and the identity of likely target genes. Finally, variant-gene pairs passing all these tests are filtered and organized with pathway mapping and gene set enrichment.

Nascent implementations of the ERV workflow have been used in several pharmacogenomics variant discovery experiments **[Higgins et al 2015, Higgins et al 2015, Higgins et al 2017]**, in a

number of drug disease systems including lithium for bipolar disorder, valproic acid (VPA) for traumatic brain injury, and citalopram and ketamine for major depression (unpublished data). The PIP is designed to introduce additional capabilities and further automation into such experiments.

The Glutamatergic Lithium Pathway

Since the publication of the lithium analysis, concordant work from other groups has shown that lithium response is controlled by a glutamatergic neuronal pathway operating in lithium-responsive regions of the human brain.

Variants in the *GRIA2* gene have demonstrated evidence for association with lithium response in bipolar disorder in several studies **[Perlis et al 2009, Hou et al 2016, Oedegaard et al 2016, Alda et al 2016]**. In contrast to glutamate-induced neurotoxicity mediated by the NMDA receptor family, knockout of GRIA2 shows that this molecule is responsible for synaptic transmission that plays a role in long-term potentiation in the hippocampus, which mediates consolidation of long term memory **[Jia et al 1996]**. Knockout of the *GRIA2* gene in a mouse model has also demonstrated the role of GRIA2 (GluR2) for stimulus-based reward conditioning **[Mead et al 2003]**. This conditioned association between environmental stimuli and reward is important in the control of appropriate behavioral responses in bipolar disorder in humans **[Murphy et al 2001]**. It has also been implicated in addiction, suggesting a link between bipolar disorder and substance abuse, which is common in patients with mood disorders **[Levin et al 2004]**.

There are numerous gene association studies and GWAS detailing SNPs in other genes in this pathway that are associated with lithium response and its adverse events in bipolar disorder. For example, mutations in the *CACNA1C* and *CACNG2* (aka *Stargazin*) genes have been significantly associated with lithium response in bipolar disorder **[Squassina et al 2011, Silberberg et al 2008]**, and these are presumed to be the voltage-dependent calcium channels associated with the GRIA2 receptor. Similarly, SNPs in the *NR1D1* (aka *Rev-erb-α*) gene have been significantly associated with lithium response in bipolar disorder in a Sardinian population **[Campos-de-Sousa et al 2010]**, and presumably are associated with disruption of circadian rhythmicity in these patients, a symptom of several mood disorders. In the context of adverse events associated with lithium, SNPs in *HTR1A* have been associated with weight gain **[Can et al 2014, Murphy et al 2013, Myosinski et al 2014]**, SNPs in *SLC1A2* have been associated with the type of tremor observed in bipolar patients who are on lithium therapy **[Thier et al 2012, Brandler et al 2013]**, and SNPS in the *CACNA1C* gene have been associated with Brugada syndrome **[Crawford et al 2015, Ou et al 2015].**

Several functional studies and a GWAS for lithium were published following the publication of the lithium analysis **[Higgins et al 2015]** in August of 2015. A genomewide association study (GWAS) for lithium-responsive bipolar disorder **[Song et al 2015]** found a significant association in European populations for SNP rs116323614, located within an intron of the *SESTD1* (SEC14 and spectrin domains 1) at $p = 2.74 \times 10^{-08}$. Following a search of the published literature and query of both Panther GO analysis and IPA®, SESTD1 was found to be significantly associated with a molecule in the pathway, DLG4, formerly known as PSD-95 **[Jie et al 2015, Lee et al 2015,**

Talpale et al 2014]. This image was taken from IPA® when DLG4 and SESTD1 were entered and the "connect" build function was used, following selection for expression in human brain:

"Calcium channel genes associated with bipolar disorder modulate lithium's amplification of circadian rhythms….In human fibroblasts, *CACNA1C* genotype predicted the amplitude response to lithium." **[McCarthy et al 2015]** Again, SNPs within another gene in this pathway, *CACNA1C,* seems to be significantly correlated with functional disruption of calcium flux associated with lithium response.

CACNA1C genotype alters the amplitude of the calcium current in human fibroblasts in response to lithium in bipolar patients **[McCarthy et al 2015]**. In addition, there are significant differences in the circadian expression of the *CACNA1C* gene in bipolar patients versus controls. It is of note that in the pathway described herein, both NR1D1 and ARNTL are transcription factors that are responsible for control of circadian rhythmicity.

Accordingly, the replication of the lithium pathway by the PIP is a demonstration of its discerning power.

The Need for Improved Warfarin Pharmacogenomics

Warfarin is an anticoagulant used for the prevention and treatment of venous thromboembolism in cardiac disease, post-operative recovery, and other contexts involving the need for coagulation control **[US FDA 2017]**. An inhibitor of the vitamin K epoxide reductase enzyme complex

(including *VKORC1*), it exhibits wide inter-individual response variation **[US FDA 2017]**. Dose requirements to achieve therapeutic international normalized ratios (INR) vary as much as 10-fold among patients **[US FDA 2017]**. Despite the recent availability of other oral anticoagulants, warfarin is still a commonly-prescribed anticoagulant, and until 2015 this venerable drug was the most popular prescribed oral anticoagulant with about 7 million office visit interactions resulting in a prescription **[Barnes et al 2015]**.

Warfarin has a complicated pharmacokinetics and pharmacodynamics, with its influences on the action of a number of clotting factors each exhibiting a different half-life **[US FDA 2017]**. It is widely acknowledged that warfarin dosing requirements and other phenotypes exhibit a strong patient-specific genetic element. Factor-of-ten differences in typical dosing requirements based on alleles in *CYP2C9*, the primary metabolic enzyme of warfarin, are present on the package insert label **[US FDA 2017]**, and data on association between warfarin requirements and other genetic loci have proliferated (see **Figure 3-3, Figure 3-8**). There have been multiple attempts to provide clinicians with genotype-guided dosing algorithms based typically on genotypes of *VKORC1* and *CYP2C9* **[Flockhart et al 2008]**. Despite the strength of variation in these two key genes, there is a "missing heritability" problem: identified loci in these genes account for only 30% to 50% of the predictive power implied by heritability estimates **[Flockhart et al 2008]**. Current generation tests use only variants in these two genes, and almost all use the same three SNPs: rs1799853, rs1057910, and rs9923231 (a promoter SNP for *VKORC1*).

Partially because of this, these tests have achieved only limited success despite intensive development and validation effort. Despite hopes that two large trials (EU-PACT **[Pirmohamed**

et al 2013] and COAG **[Kimmel 2013]**) would report good results, their appearance in the same issue of the NEJM did not bear out such hopes. COAG reported no difference in time within the therapeutic range (TTR), and EU-PACT reported a difference, but only compared to fixed starting doses with subsequent adjustment, not to initial dosing methods based on clinical indications that represent the real-world alternative to genetic dosing. Neither trial was powered to report a difference in bleeding and embolism events. Subsequently, a published meta-analysis of nine randomized controlled trials (RCTs) of warfarin pharmacogenomics dosing algorithms vs manual dose determination showed that current-generation tests using *VKORC1* and *CYP2C9* offer no improvement in TTR, percentage of patients with high INR, or bleeding and coagulation events **[Stergiopoulos et al 2014]**.

Since the conclusion of these trials, expert comment has indicated a consensus that such algorithms do not add clinical value relative to dosing on the basis of clinical indications, despite their predictive power **[Kimmel 2015, Johnson et al 2016]**. Although the recent GIFT RCT of genomic warfarin dosing found that genome-guided dosing provided a significant benefit in the composite endpoint of major bleeding, INR of 4 or greater, venous thromboembolism, or death **[Gage et al 2017]**, this generalizability of the study's results is limited. GIFT had narrow inclusion criteria: patients aged 65 years or older initiating warfarin for elective hip or knee arthroplasty. Those patients are at higher risk, and thus the generalizability of these results to larger patient populations and indications is debatable.

Despite their predictive power, such tests have failed to deploy in mainstream clinical practice. This experience suggests that new lines of research should be opened, such as the application of

advanced genomic methods, to discover new variants, recover missing heritability, and develop a new generation of tests which can add value relative to dosing by clinical indications.

Featureset

The Pharmacoepigenomics Informatics Pipeline (PIP)

The PIP uses lead variants from GWAS and candidate gene studies to find genetically linked permissive candidate variants (PCVs), using data from the 1000 Genomes Project for populations matched to the source studies. These variants are evaluated by two separate workflows: the expression regulatory variant (ERV) workflow for regulatory variants, and the coding variant (CV) workflow for coding variants. The ERV workflow evaluates the PCVs in disease-relevant tissues for DNA methylation, transcription factor binding, histone marks, DNase I hypersensitivity, chromatin state, quantitative trait loci (QTLs), and transcription factor binding site disruption using tissue-specific omics datasets. The CV workflow finds common non-synonymous coding variants within the pool of PCVs. Both sets of variants are mapped back to their host genes using RefSeq **[O'Leary et al 2016]**, and screened for expression in relevant tissues. The final output variants and their host genes are subjected to pathway analysis in Ingenuity Pathway Analysis® (Qiagen GMBH) **[Kramer et al 2014]**, offering an additional level of screening along with mechanistic insight for subsequent pharmacogenomic test development.

The methods of the PIP are shown in **Figure 3-2**. The PIP is based on the methods of our previous paper on lithium pharmacogenomics **[Higgins et al 2015]**. We performed two experiments with

this pipeline. First, a reproduction of the lithium experiment, with a restricted PIP feature set to replicate these methods, to validate the initial pipeline by reproducing our lithium pathway. Second, a new experiment in the pharmacogenomics of warfarin using the full feature set, to investigate whether the PIP may add value to a well-studied area of pharmacogenomics.

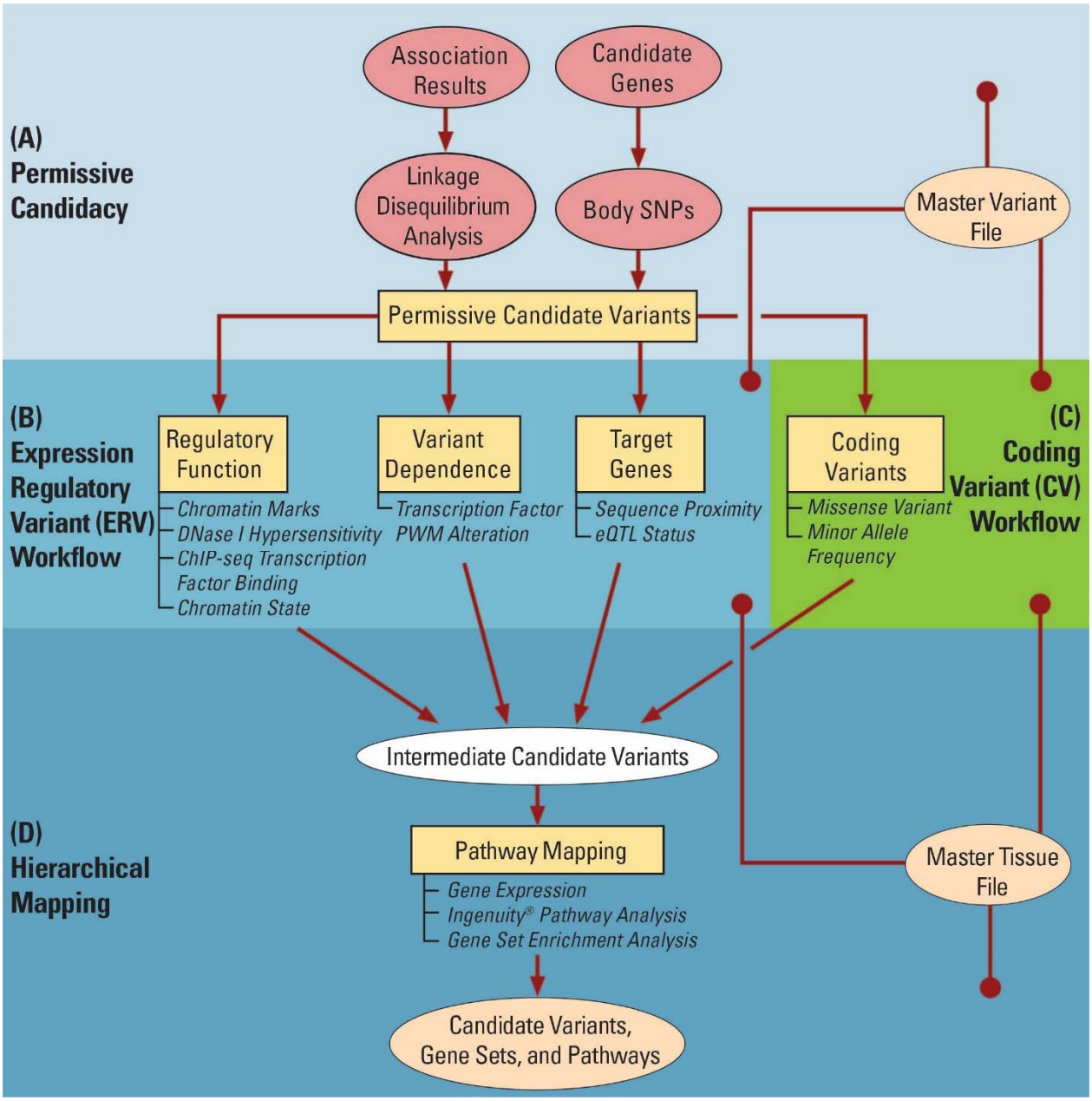


Figure 3-2: Schematic of the Pharmacoepigenomics Informatics Pipeline

(A): The Permissive Candidacy portion of the PIP. GWAS lead SNPs and known significant genes for the phenotype of interest are encoded by the user in a Master Variant File (MVF) and expanded by population-specific linkage and body SNP mapping (respectively) to generate a list of Permissive Candidate Variants (PCVs). **(B):** The Expression Regulatory Variant (ERV) Workflow. All of the PCVs are evaluated as putative regulatory SNPs using tissue-specific omics

data on phenotype-relevant tissues encoded by the user in a Master Tissue File (MTF). They are evaluated according to the regulatory function of the chromatin state segment in which they reside, the dependence of that regulatory function on the variant allele, and the presence of identifiable target gene relationships. **(C):** The Coding Variant (CV) Workflow. In the CV workflow, all the PCVs are analyzed to investigate whether they are non-synonymous coding variants with a meaningful minor allele frequency. **(D):** The Hierarchical Mapping portion of the PIP. Variants passing either workflow (Intermediate Candidate Variants) are associated with target genes, and those expressed in relevant tissues (per the MTF) are subjected to pathway mapping and gene set enrichment in Ingenuity Pathway Analysis (IPA). Gene sets associated with significant and related pathways are regarded as substantive, and the variants influencing them (regulatory and coding) are considered Candidate Variants.

Construction of Input Files for Lithium and Warfarin PIP Experiments

A PIP experiment begins with the input of associated lead variants in the form of a master variant file (MVF) which contains the refseq ID (rsID) of the variant, the populations in which the original association was derived using 1kG Phase 3 population codes, and the PubMed ID (PMID) of the study in which the association was derived.

PIP experiments were undertaken for lithium/BPD and warfarin/anticoagulation, beginning with the construction of MVFs. For the warfarin experiment, a literature review in September 2016 identified 23 GWAS on warfarin response and other pharmacological phenotypes of warfarin, venous thromboembolism risk, and baseline anticoagulant protein levels in healthy patients. Joint

analysis of studies on related phenotypes, and overlap between disease risk and pharmacogenomic variants, are known phenomena from previous work. In addition, we added 23 variants annotated by PharmGKB®, yielding a total of 204 SNPs. The input studies for the warfarin experiment are summarized in **Figure 3-3**.

The warfarin experiment began with input data from populations all over the world, including European, East Asian, South Asian, African, and American cohorts. Population specific associations from these groups were interpreted in the context of other population specific input data, from the worldwide genome catalog of 1000 Genomes. Population descriptions are taken from the summary annotations in the original papers and the GWAS Catalog, but in the PIP input files they are represented in 1000 Genomes format to replicate the detailed descriptions in the original papers as precisely as possible.

The various consortium data (ENCODE, REMC) and ChromHMM **[Ernst et al 2012]** chromatin states are imported for relevant tissues, with the relevant tissues being supplied in the format information for each consortium separately, for a unified set of tissues, in a Master Tissue File (MTF). The MTF is designed for each experiment to contain the set of tissues which are relevant to the particular drug-disease system under investigation, so that tissue specific omics data may be used in the experiment. Information on the relevant chromatin marks and transcription factors was imported in a series of TF and Chromatin mark whitelist files containing metadata codes in the formats of each of the consortia.

MTFs were constructed for both experiments. The warfarin MTF included the human liver (site of warfarin metabolism), vasculature (site of action) and small intestine (site of absorption, a known variable in response, and metabolism). These were represented by a mixture of cell line and tissue samples in the ENCODE, REMC, and GTEx datasets, but a complete epigenome with methylation, DNase accessibility, all core histone marks, and many TF binding tracks was available in all tissues.

With the MVFs and MTFs generated for the lithium and warfarin experiments, the PIP was used to conduct its analysis.

Permissive Candidate Generation and Variant Annotation

Generation and annotation of Permissive Candidate Variants (PCVs) occurs in the following manner:

- Linkage disequilibrium mapping is performed in the PLINK software package **[Lu et al 2017]**, version 1.9 **[Chang et al 2014]**, with the May 2 2013 (latest) release of 1000 Genomes consortium Phase 3 data. For each of the SNPs in the MVF, PLINK computes LD coefficients and outputs a set of all cis and trans variants that achieve R^2>=0.8 linkage with the input variant. These variants constitute the set of permissive candidate variants (PCV).

- Chromatin state annotation is performed with the version of ChromHMM results available in the Washington University (WUSTL) epigenome browser **[Zhou et al 2013]** as of April 2016. Chromatin mark information on a number of indicative marks, along with DNA accessibility measurements, TF binding, and methylation (currently not used in scoring) are assayed using ENCODE and REMC data in the relevant marks and the relevant tissues as rendered in the Master Tissue File and whitelist files.

- TFBS disruption data is assessed using TFM-Scan **[Liefooghe et al 2006]** on 23-bp-per-side hg19 sequences on reference and alternate alleles from the source SNPs, using integer-converted data from the PWM library used in HaploReg **[Ward et al 2016]**, and requiring a threshold-crossing-or-10-point difference in score between reference and alternate alleles for a qualifying PWM, as a measure of significance.

- QTL data were imported from eight sources: 1) The HaploReg QTL database **[Ward et al 2016]**, 2) SeeQTL **[Xia et al 2012],** 3) MuTHER **[Grundberg et al 2012],** 4) GTEx **[Carithers et al 2015]**, 5) Shi et al meQTLs **[Shi et al 2014]**, 6) McClay et al meQTLs **[McClay et al 2015]**, 7) Banovich et al meQTLs **[Banovich et al 2014]**, 8) GeneVar **[Yang et al 2010]**. Gene annotations were retrieved from the latest build of RefSeq in hg19.

- Gene expression data was retrieved from the latest build of GTEx **[Carithers et al 2015]**, and from the Allen Brain Atlas **[Sunkin et al 2013]**.

- The coding variant workflow located all non-synonymous coding variants among the PCVs using dbSNP data, then screened them for minor allele frequency with 1000 Genomes data on all human populations, using a MAF cutoff of .01.

Variant Scoring, Expression Testing, and Pathway Analysis

Scoring proceeded according to **Figure 3-3** above, with SNPs being required to pass all of the separate thresholds in each of the individual portions of either one of the two workflows (ERV and CV) to proceed. Then, their host genes must pass gene expression testing in relevant tissues in order to reach the final gene list. All of these steps are undertaken on the Flux computing cluster at the University of Michigan. Finally, the gene list is analyzed with IPA.

Lithium PIP Experiment: (Mostly) Same Methods as Original Experiment

The two experiments were conducted with slightly different feature sets. The warfarin experiment used the full feature set, while the lithium experiment used a restricted feature set designed to match the semi-automated analysis from our 2015 paper, including no use of gene bodies as candidate region inputs, no CV workflow, expression analysis performed manually with Allen Brain Atlas *in situ* hybridization data on brain tissue, and the use of bio-chronicity analysis to screen genes.

The lithium MVF was constructed to contain precisely the set of input variants originally used in the lithium experiment previously published, a total of 108 SNPs. The lithium MTF included the same tissues used in the original lithium experiment; BPD-relevant brain regions plus the human liver. The lithium MTF does not feature GTEx data because the restricted featureset for this experiment used manual gene expression analysis with Allen Brain Atlas in-situ data in order to replicate the original lithium experiment. The tissue files for both experiments are summarized in **Figure 3-4**.

Despite this effort, the lithium experiment feature set of the PIP does not fully match the original experiment **[Higgins et al 2015]**. Several obstacles were encountered in which subtle differences in methods were necessitated by database deprecation, ambiguities in the documentation of available literature methods, and ambiguities in versioning of online resources. These obstacles included:

LD mapping methods. In the original paper, LD mapping was performed in HaploReg v4.1 **[Ward et al 2016]**, which uses the Phase 1 cohort of the 1000 Genomes project **[1000 Genomes Project Consortium 2015]** as its LD mapping cohort. However, this tool was developed after the Phase 3 analysis results became available for this cohort, and it is unclear if they used the Phase 1 variant calls, or Phase 3 variant calls. We elected to use the Phase 3 cohort and population stratifications with the Phase 3 variant calls, because they are the most accurate.

Switch in QTL databases. The GeneVar **[Yang et al 2010]** QTL database we used in the original analysis is no longer available in its full form. We have contacted the original authors and they

have not responded. Because of this, we have been forced to seek supplemental databases in addition to the partial GeneVar database still available, as described below.

PWM integer approximation. The PIP uses the same database of PWMs that was used by HaploReg **[Ward et al 2016]** during the initial Lithium analysis, and scores them against the flanking sequences using the same software, TFM-Scan **[Liefooghe et al 2006]**. However, the PWM matrices available for bulk download are in probability format, and TFM-Scan presents warnings about accuracy when using probability matrices. It expects frequency matrices. We do not know if HaploReg used the probability matrices and (possibly) suffered inaccuracies, or if they have copies of the matrices in natively-integer format which they have not made available, or if they converted the probability matrices to integer. We have elected to convert the probability matrices to frequency matrices, but because of rounding this is imperfect and may introduce small errors.

PMID	Phenotype Class	Phenotype	Sample Size	Population Ancestry*	SNPs
18535201	Response	Warfarin maintenance dose	555	European	3
19300499	Response	Warfarin maintenance dose	1641	European	4
20833655	Response	Warfarin maintenance dose	1952	Japanese	5
23755828	Response	Warfarin maintenance dose	965	African American	1
26265036	Response	Warfarin maintenance dose	2967	Brazilian, European, Japanese, African American	16
22443383	ADEs	Hemostatic factors and hematological phenotypes	951	European	9
23381943	ADEs	End-stage coagulation	2100	European	23
24357727	ADEs	Thrombin generation potential phenotypes	3224	European	4
19278955	Disease/Background	Venous thromboembolism	4884	European	1
20212171	Disease/Background	C4b binding protein levels	352	European	1
20303064	Disease/Background	Activated partial thromboplastin time	1431	European	3
21980494	Disease/Background	Venous thromboembolism	2652	European	4
22216198	Disease/Background	Anticoagulant levels	397	European	8
22701019	Disease/Background	Factor XI	997	European	0
22703881	Disease/Background	Prothrombin time	3569	European	2
22703881	Disease/Background	Activated partial thromboplastin time	11851	European	9
22672568	Disease/Background	Venous thromboembolism	5787	European and other	5
23267103	Disease/Background	Coagulation factor levels	3250	European	7
23509962	Disease/Background	Venous thromboembolism (SNP x SNP interaction)	4291	European	37
23650146	Disease/Background	Venous thromboembolism	51266	European	8
25240745	Disease/Background	Mitochondrial DNA levels	386	European	21
25772935	Disease/Background	Venous thromboembolism	65734	European	10
22550155	Disease/Background	Platelet thrombus formation	241	European, African American	6

Figure 3-3: GWAS Input Studies for the Warfarin Experiment

23 associations from 22 GWAS **[Cooper et al 2008, Takeuchi et al 2009, Cha et al 2010, Perera et al 2013, Parra et al 2015, Oudot-Mellakh et al 2012, Williams et al 2013, Rocanin-Arjo et al 2014, Tregouet et al 2009, Buil et al 2010, Houlihan et al 2010, Germain et al 2011, Athanasiades et al 2011, Sabater-Lleal et al 2012, Tang et al 2012, Heit et al 2012, Desch et al 2013, Greliche et al 2013, Tang et al 2013, Lopez et al 2014, Germain et al 2015, Edelstein et al 2012]** met the criteria for inclusion in the warfarin experiment, and their significant associations were rendered into the MVF and organized into phenotypic classes, along with 23

additional variants annotated by the Pharmacogenomics Knowledge Base (PharmGKB®) **[Whirl-Carrillo et al 2012]**. The 23 additional variants included all variants annotated by the PharmGKB as of September 2016, with a significant association with warfarin response, and were not already included in the list of variants identified from GWAS. Population descriptions are taken from the summary annotations in the original papers and the GWAS Catalog, but in the PIP input files they are represented in 1000 Genomes format to replicate the detailed descriptions in the original papers as precisely as possible. Although they represent a variety of ancestries, European ancestry predominated among 17 out of 22 (77%) of the study populations.

(A) Lithium			
Tissue	**ENCODE**	**REMC**	
Frontal lobe	Brain_Mid_Frontal_Lobe	H1-hESC:SK-N-SH	
Insula	Brain_Cingulate_Gyrus	H1-hESC:SK-N-SH	
Temporal cortex	Brain_Inferior_Temporal_Lobe	H1-hESC:SK-N-SH	
Cingulate cortex	Brain_Cingulate_Gyrus	H1-hESC:SK-N-SH	
Amygdala including_ Amygdalo-hippocampal transition area	Brain_Inferior_Temporal_Lobe	H1-hESC:SK-N-SH	
Hippocampal formation including Parahippocampal gyrus	Brain_Hippocampus_Middle	H1-hESC:SK-N-SH	
Anterior caudate and Putamen	Brain_Anterior_Caudate	H1-hESC:SK-N-SH	
Thalamus, right	Brain_Anterior_Caudate	H1-hESC:SK-N-SH	
Fusiform cortex, Angular gyrus	Brain_Angular_Gyrus	H1-hESC:SK-N-SH	
Motor cortex, Substantia nigra, Cerebellum	Brain_Substantia_Nigra	H1-hESC:SK-N-SH	
Hypothalamus	Fetal_Brain_Female: Fetal_Brain_Male	H1-hESC:SK-N-SH	
(B) Warfarin			
Tissue	**ENCODE**	**REMC**	**GTEx**
Liver	Adult_Liver:HepG2_Hepatocellular_ Carcinoma	HepG2:liver:hepatocyte:right lobe of liver	Liver
Vasculature	Aorta: HUVEC_Umbilical_Vein_Endothelial _Cells	endothelial cell of umbilical vein:dermis blood vessel endothelial cell:pulmonary artery endothelial cell:endothelial cell of coronary artery: aorta:thoracic aorta endothelial cell:vein endothelial cell:lung microvascular endothelial cell	Artery - Tibial: Artery - Coronary: Artery - Aorta
Small Intestine	Fetal_Intestine_Small:Small_Intestine	small intestine:Caco-2:jejunum	Small Intestine - Terminal Ileum

Figure 3-4: Tissue Files used in the Lithium and Warfarin Experiments

The tissue file coding used in the lithium (A) and warfarin (B) PIP experiments. The first column contains natural language names for the tissues, while subsequent columns contain colon-delimited consortium ontology codes for related tissues and cell lines. The lithium experiment does not have

tissue codes for GTEx because annotation of gene expression mapping was done manually with Allen Brain Atlas *in situ* hybridization data **[Sunkin et al 2013]** in this experiment, as described herein.

The Lithium PIP Experiment

More Significant Pathway, More Genes and Variants

To demonstrate the ability of the PIP to replicate the results of previous integrative omics methods in pharmacoepigenomic variant discovery, we first undertook a PIP experiment to replicate the findings of our 2015 paper on lithium **[Higgins et al 2015]**, as a positive control. The featureset of the PIP used for this experiment was restricted to match that of the previous experiment, as described in the methods.

The lithium experiment yielded a total of 1727 PCVs, of which a total of 78 passed the entire set of filters. These results differ from and expand on the results of the original lithium analysis. Whereas the original experiment identified 19 SNPs in 10 genes, the new experiment identified 17 out of 19 original SNPs, and all 10 original genes, plus more SNPs in these and two more genes for a total of 78 SNPs in 12 genes. These are the 10 genes from the original lithium paper, with the addition of *HTR1A* and *TNIK*, both of which appeared in the glutamatergic lithium pathway called by IPA in the original lithium paper.

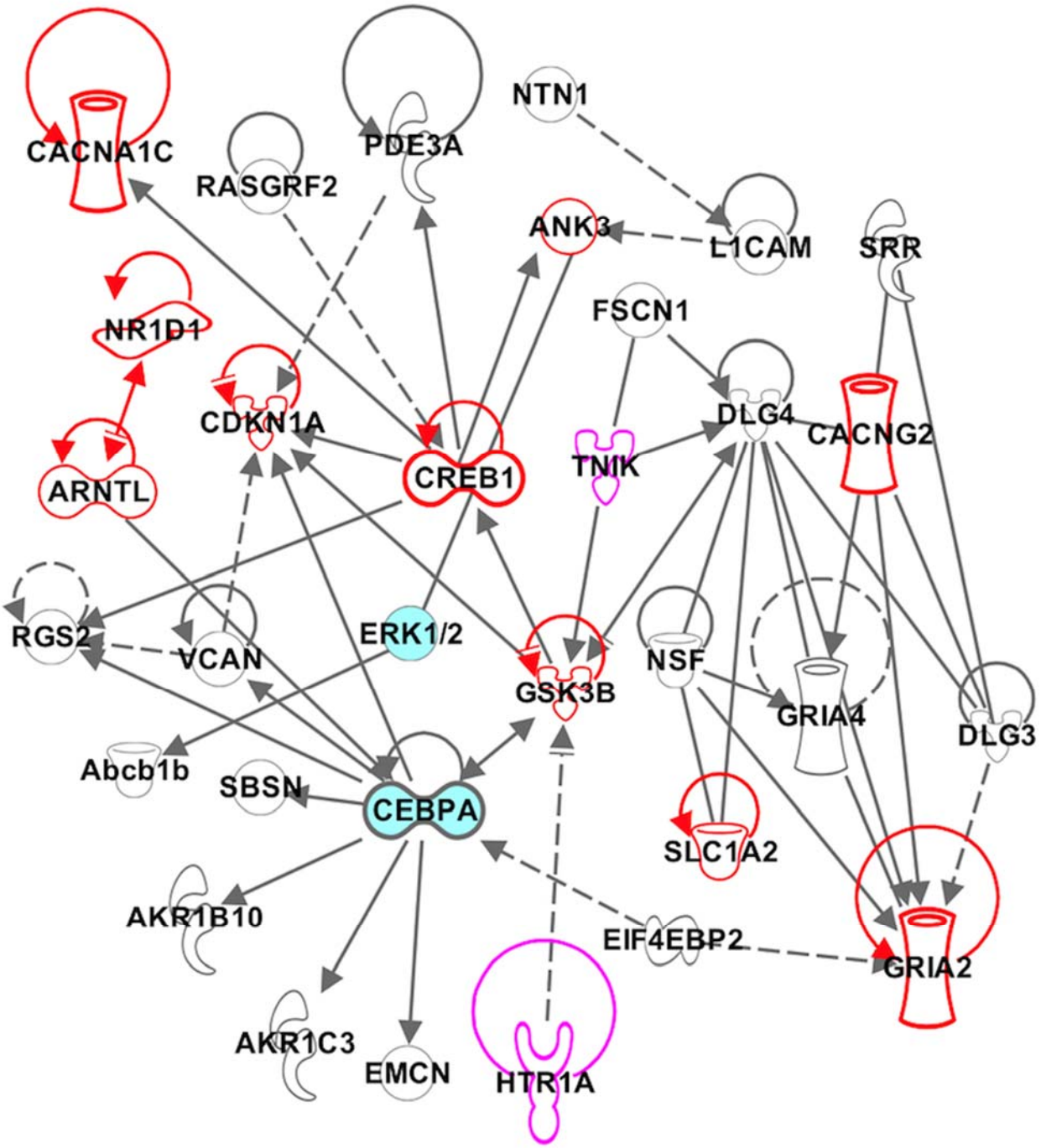
Because of the factors mentioned in the methods, as well as imprecisions in the manual steps in the original lithium experiment, the lithium results differ subtly from those of the original experiment. This underscores the importance of using versioned, locally stored versions of all databases in bioinformatics experiments to allow reproducibility. In addition, however, the

automated analysis recovers additional genes from the same pathway in the lithium experiment, emphasizing the utility of automated workflows with advanced functionality.

Thus we reproduce the glutamatergic pathway for lithium response with the PIP, as reflected in **Figure 3-5 and Figure 3-6**. IPA calls the same pathway for both the 10 genes discovered in the original experiment, and the expanded 12-gene set. The pathway also includes 22 additional genes not discovered in either experiment but used to connect them in the pathway mapping functionality of IPA. The two pathways differ only in IPA's inclusion of a group of miRNAs in the old pathway and not in the new, which is the result of an IPA database change. The PIP recovers the same pathway as the original experiment.

In addition to this, in the period of time between the execution of the 2015 lithium experiment and the publication of this manuscript, other groups have reported connections between lithium response and ADME and some of the genes we reported in this network **[Crawford et al 2015, Mitjans et al 2015]**.

Using substantially the same input data as the original lithium experiment **[Higgins et al 2015]**, the PIP pipeline successfully identified the same glutamatergic pathway in the human brain, and identified additional genes which are part of the same pathway. This constitutes a validation that the PIP replicates the discerning power of our previous semi-automated methods, and may function as a foundation for future development.


CACNA1C
PDE3A
NTN1
RASGRF2
ANK3
L1CAM
SRR
NR1D1
FSCN1
CDKN1A
DLG4
CACNG2
ARNTL
CREB1
TNIK
RGS2
VCAN
ERK1/2
NSF
GSK3B
GRIA4
DLG3
Abcb1b
SBSN
CEBPA
SLC1A2
AKR1B10
EIF4EBP2
GRIA2
AKR1C3
EMCN
HTR1A

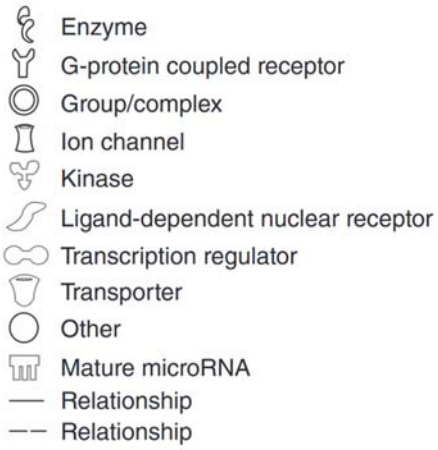
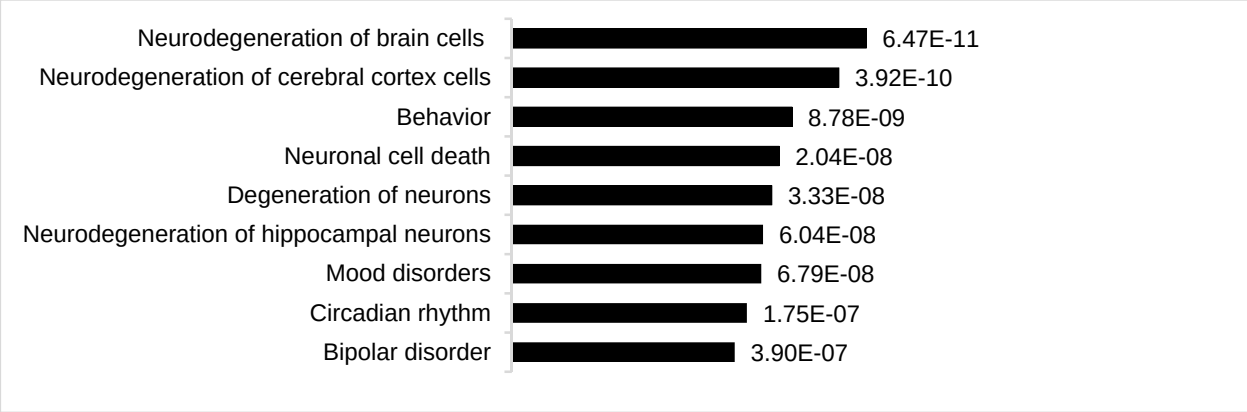


Figure 3-5: PIP Reproduction of the Glutamatergic Lithium Pathway

The same lithium pathway discovered by both manual and PIP experiments. The ten genes discovered by both experiments (*ANK3, ARNTL, CACNG2, CACNA1C, CDKN1A, CREB1, GRIA2, GSK3B, NR1D1, SLC1A2*) are in red; 2 genes (*HTR1A, TNIK*) discovered by the PIP experiment only are in purple. The PIP discovers 78 candidate variants including 61 not discovered in the original lithium experiment. Gene classes and relationship types are designated with line types and gene glyphs as shown in the legend and described in the IPA® documentation **[Kramer et al 2014]**.

A)



B)

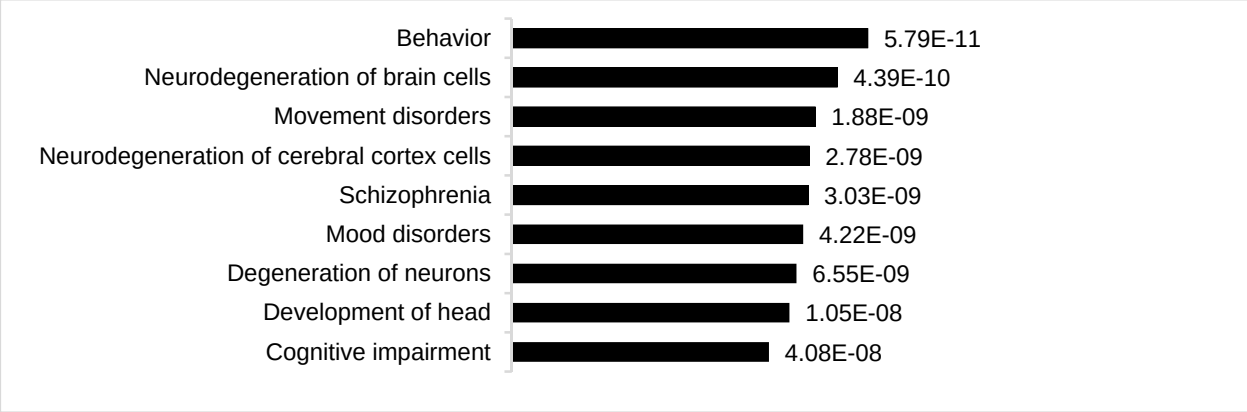


Figure 3-5: Lithium Pharmacogenomic Gene GSEA Results

(A): Top diseases and functions detected by IPA for the ten genes detected in the original lithium experiment. **(B):** Top diseases and functions for the twelve genes detected in the PIP lithium experiment. Although there is significant overlap in the top nine hits for the two gene sets (five categories) there are some significant differences, demonstrating that the extra discerning power of the automated methods may be adding value relative to the semi-manual experiments.

The Warfarin PIP Experiment

Known and Novel Variants Form an Enhanced and Unified Pathway

With the methods thus validated, we proceed to the results of the warfarin experiment in the PIP. The input variants from the warfarin study yielded a total of 4492 PCVs, which were analyzed for each of the criteria in the ERV and CV workflows. Of these PCVs, there were significant associations between the sets of SNPs passing some related modules in the PIP, and some modules were, in this experiment, significantly stricter than others. However, all the modules filtered out a meaningful number of SNPs as part of the overall process. This process is summarized in **Figure 3-7.**

The ERV and CV workflows identifies a total of 223 SNPs, which were then mapped to their host and target genes, which were subjected to the expression test with GTEx data on tissues from the MTF. Of these 223 SNPs, 87 were located in 41 genes passing the expression test. Notable among them is the presence of classic and previously known warfarin response genes including *CYP2C9*, clotting factors and *VKORC1*.

Next, we performed pathway analysis in IPA® on this set of 41 genes. There are only four genes whose variants are used in current generation warfarin pharmacogenomics tests: *VKORC1, CYP2C9, CYP4F2,* and *GGCX*. Despite the predictive power and clinical utility of these variants and the tests constructed with them, IPA does not connect them into a pathway, and previous warfarin literature has not conceptualized warfarin response in pathway terms. Despite this, and

despite the typical pattern wherein IPA® discovers many pathways of questionable relevance, this analysis yielded only two pathways, one of them with 31 genes and a particularly striking p-value of 10E-65 (as calculated by IPA). The gene list for this pathway is shown in **Figure 3-8**, while the pathway network diagram is shown in **Figure 3-9.** It densely unites previously known warfarin response genes (including those not previously used in test design) with new genes.

This pathway may be considered the "canonical" warfarin pathway based on the published literature base. It consolidates the bulk of discovered genes, particularly all but one of those previously known for warfarin response, including *VKORC1*, warfarin-metabolizing *CYP* enzymes, and a collection of clotting factors. This includes the FGA and FGG fibrinogen subunits. Despite the fact that they operate as a complex and cannot function separately, FGG had been previously identified for warfarin response but FGA had not, whereas the PIP identifies both genes, which are located in the same TAD and are co-regulated.

Among previously known genes in this pathway, over 75% are highest-expressed in liver, while 57% of newly discovered genes are highest-expressed in the intestines or vasculature. Notably, a mixture of coding and regulatory SNPs are identified for both known and new genes.

The second pathway contains only one previously known warfarin gene (and that one, *BCKDK*, of disputed significance), does not show coagulation-related GSEA results, and its central elements are not of relevance. It may be considered an artifact.

Next, we processed the genes in the warfarin pathway with GSEA in IPA and GO. The results of the GSEA are shown in **Figure 3-10.** Warfarin and coagulation related terms are prominent among the results, and many of the genes contributing to these identifications were not previously identified as warfarin-related.

The PIP has identified many known genes for warfarin phenotypes, and added both new coding and regulatory SNPs for known genes, as well as new genes.

In addition to this, we investigated the warfarin input and output variants with DeltaSVM **[Lee et al 2015, Ghandi et al 2014, Ghandi et al 2016]**, evaluating their propensity to differentially induce DNase accessibility in the HepG2 liver cell line. Although the result SNPs contained a number of strong candidates by DeltaSVM, these candidates were not enriched relative to the candidates among the input set, nor were the mean and variance of the DeltaSVM scores significantly different between inputs and outputs. This implies that our existing variant dependence algorithms (PWM alteration) are mostly independent of DeltaSVM's predictive power, and that adding Delta SVM or similar methods to the variant dependence portion of the pipeline (with appropriate changes to the scoring system) would add value to the pipeline.

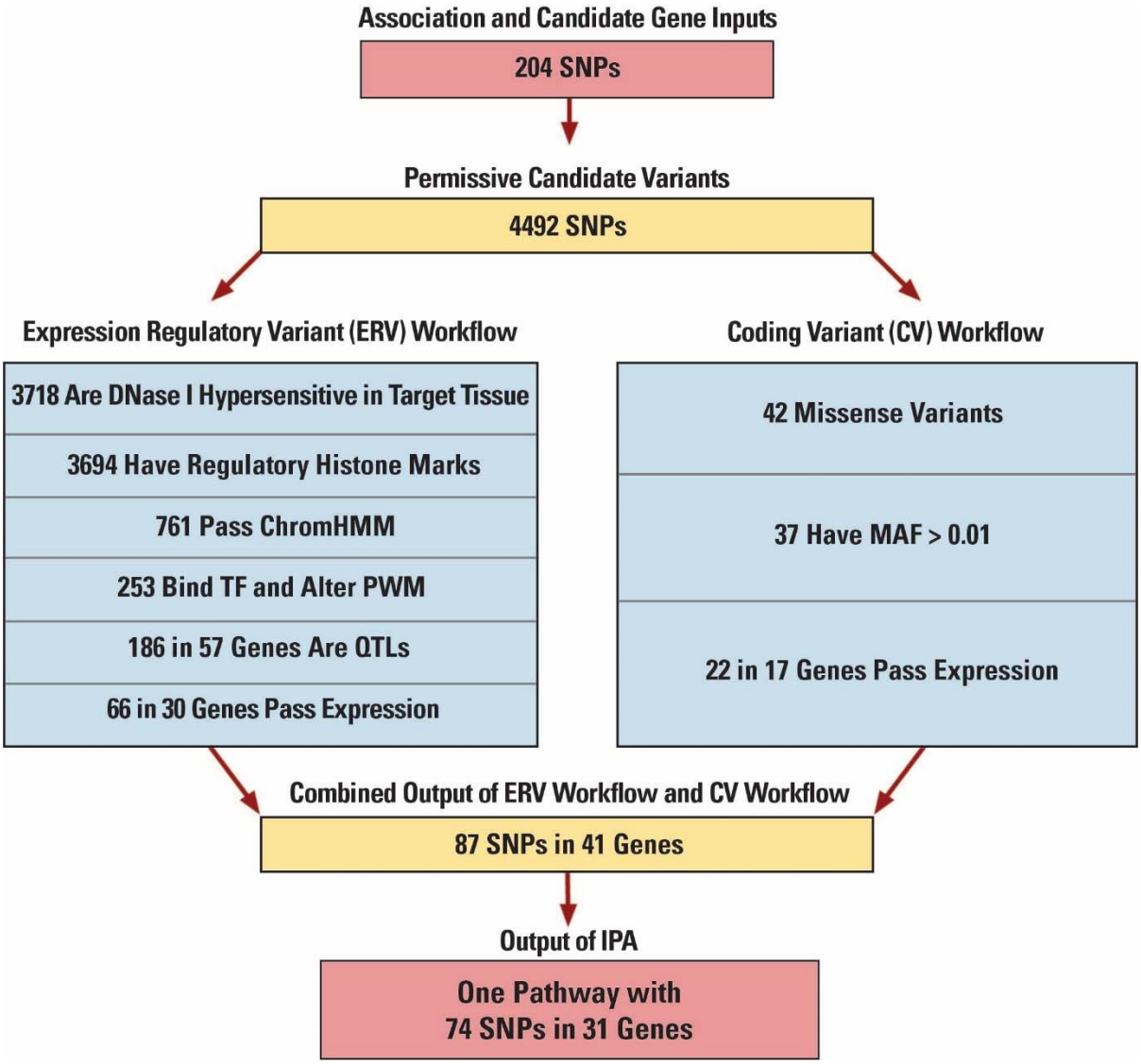


Figure 3-7: Disposition of PCV SNPs in the Warfarin PIP Experiment

The disposition of SNPs within the workflows of the PIP during the warfarin experiment. 190 input SNPs expand to 4492 PCVs. After evaluating all of these PCVs, the ERV workflow outputs 186 regulatory SNPs mapping to 57 genes, of which 30 (bearing 66 SNPs) were expressed in relevant tissues in GTEx datasets. The CV workflow outputs 37 SNPs in 27 genes, of which 22 in 17 genes pass the expression test. These 87 genes in 41 SNPs were subjected to IPA analysis,

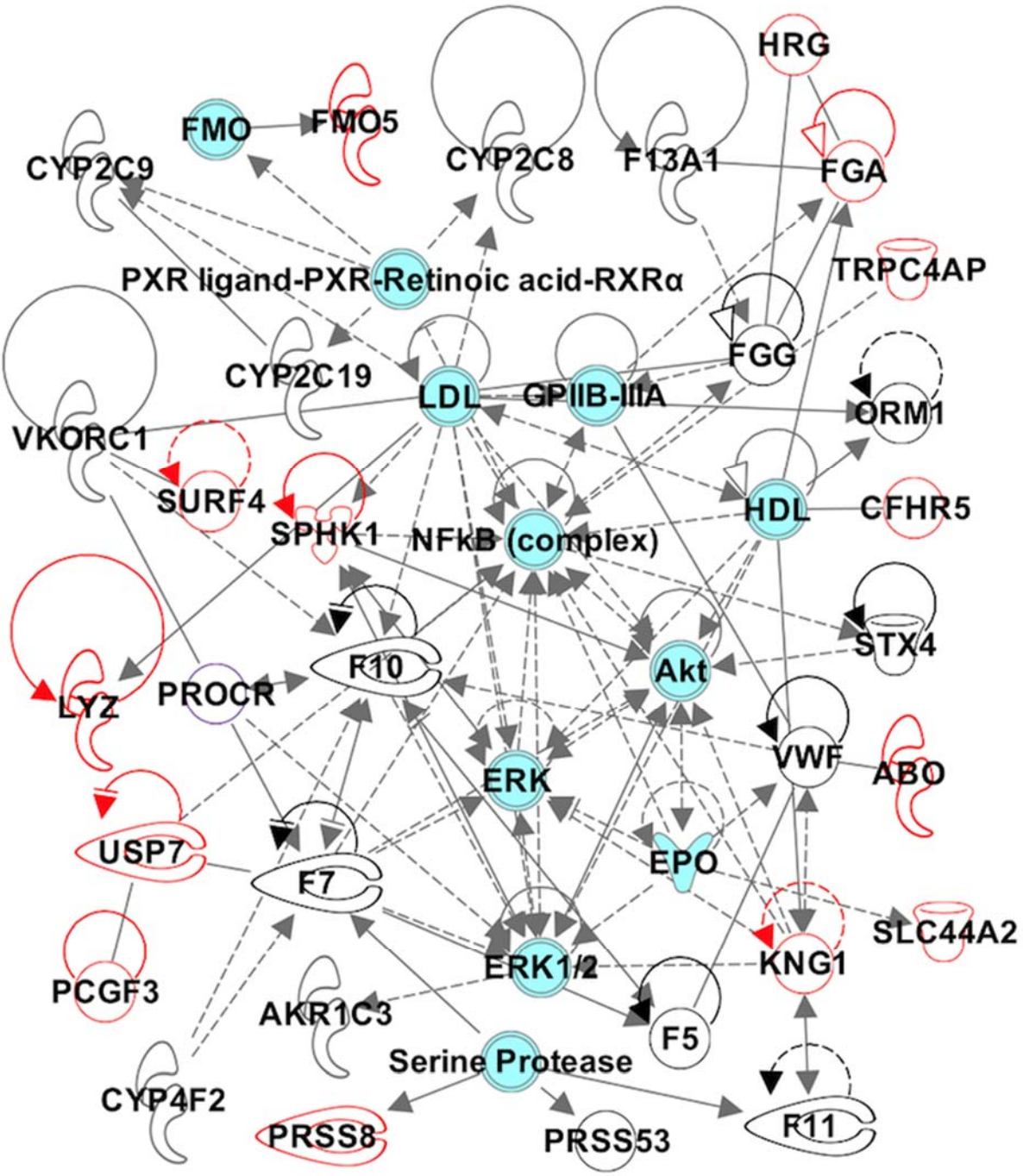
revealing a combined network of 31 genes bearing a total of 74 SNPs (53 regulatory and 21 coding), as shown in **Figure 3-8** and **Figure 3-10**.

Gene	Function	Maximum Expression Level (RPKM)	Tissue of Highest Expression	Candidate SNPs Altering Protein Sequence or Gene Expression
AKR1C3	Steroidogenic enzyme	250	Liver	rs12775913 rs346803 rs346797 rs762635 rs76896860
CYP2C19	CYP Enzyme	170	Liver	rs3758581
CYP2C8	CYP Enzyme	2000	Liver	rs10509681 rs11572080
CYP2C9	CYP Enzyme	950	Liver	rs1057910 rs1799853 rs7900194
CYP4F2	CYP Enzyme	140	Liver	rs2108622
F5	Coagulation Factor	86	Liver	rs6009 rs11441998 rs2026045 rs34580812 rs749767 rs9378928 rs7937890
F7	Coagulation Factor	110	Liver	rs7552487 rs6681619 rs8102532 rs491098 rs6046
F10	Coagulation Factor	100	Liver	rs11150596 rs11150596
F11	Coagulation Factor	77	Liver	rs2165743 rs11252944
FGG	Fibrinogen Subunit	12300	Liver	rs8050894
ORM1	Acute Phase Plasma Protein	16100	Liver	rs10982156
PRSS53	Serine Protease	37	Liver	rs7199949
VKORC1	Reductase Enzyme	150	Liver	rs2884737 rs9934438 rs897984 rs17708472
STX4	T-SNARE	33	Small Intestine	rs35675346 rs33988698
F13A1	Coagulation Factor	170	Vasculature	rs5985
PROCR	Serine Protease	150	Vasculature	rs867186
VWF	Plasma Glycoprotein	150	Vasculature	rs75648520 rs55734215 rs12244584 rs1063856
CFHR5	Plasma Glycoprotein	120	Liver	rs674302
FGA	Fibrinogen Subunit	10400	Liver	rs12928852 rs6050
FMO5	Metabolic Enzyme	73	Liver	rs8060857 rs7475662
HRG	Plasma Glycoprotein	1300	Liver	rs9898
KNG1	Coagulation Hormone	540	Liver	rs710446
SURF4	Membrane Protein	160	Liver	rs11577661
ABO	Blood Typing Enzyme	52	Small Intestine	rs11427024 rs6684766 rs2303222 rs10888838 rs13130318 rs12951513
LYZ	Lysozyme	870	Small Intestine	rs8118005
PCGF3	Polycomb Complex Subunit	36	Small Intestine	rs76649221 rs9332511 rs6588133
PRSS8	Serine Protease	280	Small Intestine	rs11281612
TRPC4AP	Ion Channel	72	Small Intestine	rs11589005 rs8062719 rs889555 rs36101491 rs7426380 rs6579208 rs77420750 rs73905041
SLC44A2	Solute Carrier	180	Vasculature	rs3211770 rs3211770 rs3087969 rs2288904
SPHK1	Inflammatory Kinase	45	Vasculature	rs683790 rs346803
USP7	Ubiquitin Protease	32	Vasculature	rs201033241

Figure 3-8: Output Genes and SNPs of Warfarin PIP Experiment

Genes passing all of the scoring thresholds in the warfarin PIP experiment, along with their known functions, highest observed expression levels in relevant GTEx tissues, and SNPs mapping to them detected in the warfarin experiment. Known warfarin response genes **[Malatkova et al 2016, Wadelius et al 2007, Chen et al 2014, Bare et al 2011, Marsh et al 2006, Bader et al 2016, Wang et al 2013, Eriksson et al 2016, Lind et al 2012, Lip et al 1995, Patillon et al 2012, Bargal et al 2016, Kimmel 2010, Undas et al 2005, McDonald et al 2009, Lee et al 2012, Scordo et al 2002, Kim et al 2012, Chang et al 2015, Limdi et al 2008, Kudzi et al 2009, John et al 2017]** are highlighted first, in black. Below, shown in red, are detected genes not previously known

to be associated with warfarin response. Among the SNPs, regulatory SNPs are identified in blue, while coding SNPs are shown in green.


HRG
FMO
FMO5
CYP2C9
CYP2C8
F13A1
FGA
PXR ligand-PXR-Retinoic acid-RXRα
TRPC4AP
FGG
CYP2C19
LDL
GPIIB-IIIA
ORM1
VKORC1
SURF4
SPHK1
NFkB (complex)
HDL
CFHR5
STX4
R10
Akt
LYZ
PROCR
VWF
ABO
ERK
USP7
EPO
F7
SLC44A2
KNG1
ERK1/2
PCGF3
AKR1C3
F5
Serine Protease
CYP4F2
R11
PRSS8
PRSS53

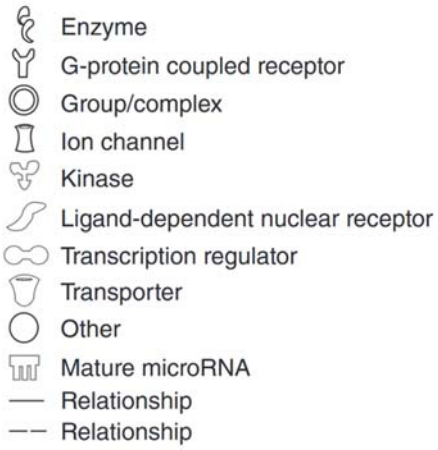
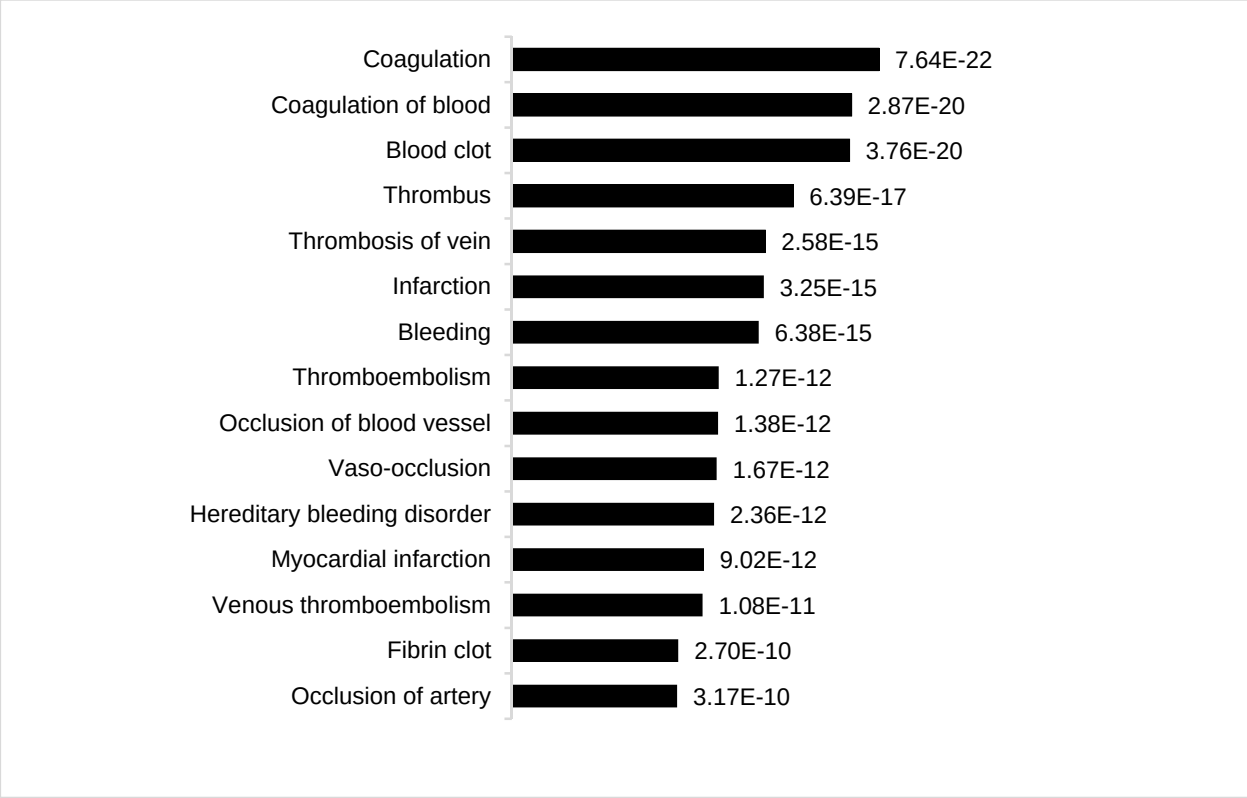


Figure 3-9: The Warfarin Pharmacogenomic Pathway

Pathway mapping results of warfarin response genes. IPA connects a warfarin pathway discovered by the PIP, with thirty one genes and a p-value less than 1E-65. Genes include: *ABO, AKR1C3, CFHR5, CYP2C8, CYP2C9, CYP2C19, CYP4F2, F5, F7, F10, F11, F13A1, FGA, FGG, FMO5, HRG, KNG1, LYZ, ORM1, PCGF3, PROCR, PRSS8, PRSS53, SLC44A2, SPHK1, STX4, SURF4, TRPC4AP, USP7, VKORC1, VWF*. Previously reported warfarin response genes are in black, and new genes are in red. Gene classes and relationship types are designated with line types and gene glyphs as shown in the legend and described in the IPA® documentation **[Kramer et al 2014]**.

A:



B:

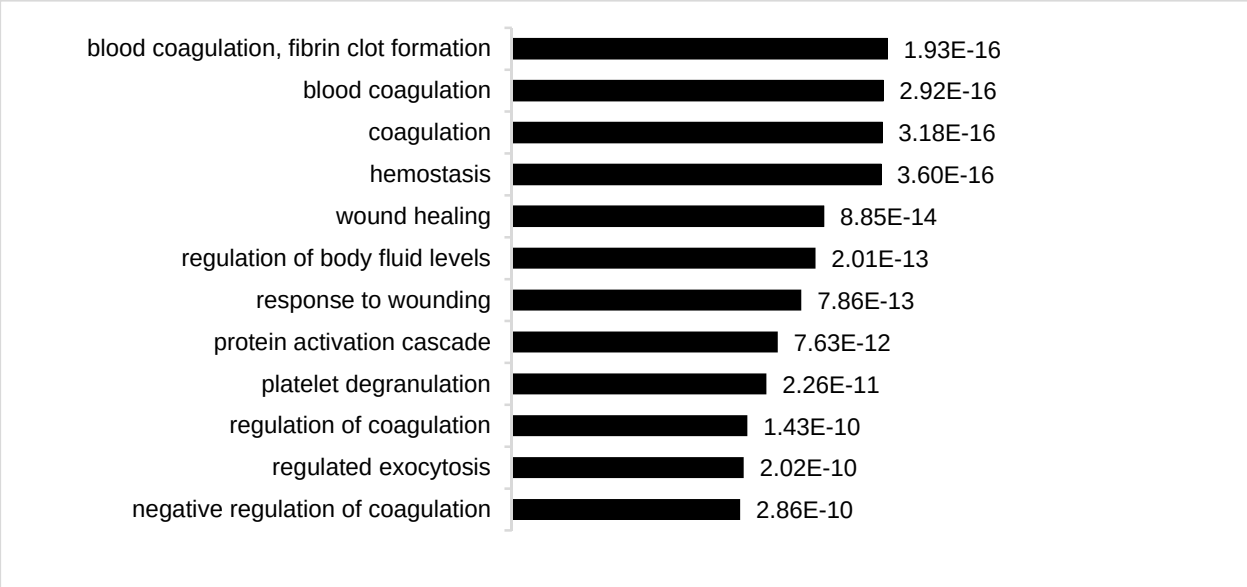


Figure 3-10: Warfarin Pharmacogenomic GSEA Results

(A): Top IPA GSEA hits for the warfarin gene pathway. **(B):** GO GSEA results for the warfarin pathway. Warfarin-related and coagulation-related gene sets are prominent among the results for the pathway in both GSEA systems, including bleeding, coagulation, clotting, thrombosis, hemorrhage, and related gene sets. Many of the genes contributing to the identification of these gene sets were not previously discovered.

Runtime and Computational Performance

The PIP is Scalable

For the warfarin experiment, total runtime and its division into the various components of the workflow, are summarized in **Figure 3-11**. The total experimental runtime on the University of Michigan Flux cluster was about 14,300 node hours (each node using two six-core Intel Xeon X5650 processors running at 2.67 Ghz, with 48GB of RAM). Of this, the bulk was devoted to the initial LD calculations, QTL analysis, and the retrieval of epigenome data from REMC datasets.

Although we anticipate substantial runtime improvements from code optimization in subsequent versions of the PIP, it is clear that this version is already computationally scalable for use in pharmacogenomics variant discovery in many systems. When sufficient cluster resources are available, it can finish running after about six hours of computation. Although REMC and QTLs are the largest contributors to total runtime, PLINK is a large contributor to parallel runtime because it is parallel on the level of input SNPs, not on the level of PCVs. With a sufficient number of nodes PLINK accounts for a third of total runtime, though not total computation.

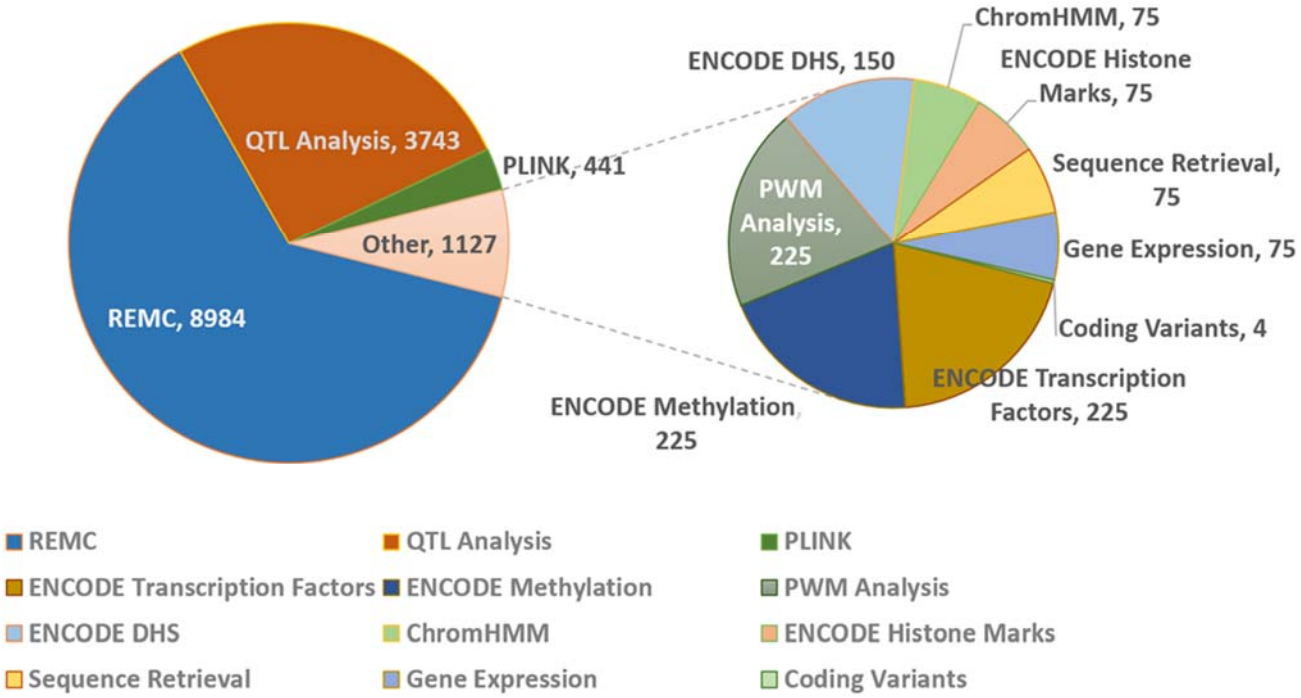


Figure 3-11: Warfarin Experiment Runtime Distribution

The distribution of runtime among experimental modules in the PIP Warfarin experiment, in node-hours. Over 90% of total runtime was composed of just three pieces: Epigenome Roadmap data analysis, QTL analysis, and PLINK linkage calculations. Of the remaining computational modules, shown in the smaller pie chart on the right, ENCODE data lookups and PWM analysis constitute the largest remaining steps. The CV workflow was computationally minimal.

Discussion

Warfarin Pharmacogenomics

The discovery of a number of previously unknown candidate genes and SNPs for warfarin response, dosing, and ADEs, including a unifying pathway with known variants, is a potential direction for a significant improvement in warfarin pharmacogenomics. This includes both additional variants controlling known warfarin response genes, and also putative new genes with associated variance. Although the previously known variants in VKORC1 and CYP2C9 are almost certainly the most potent pharmacogenomic variants for warfarin phenotypes, accounting by themselves for 30% to 50% of the heritability of response, the additional predictive power of these new variants may make the difference in providing unambiguous clinical utility for such predictive models, which has so far evaded the field of warfarin pharmacogenomics. And in the context of other anticoagulants, including warfarin analogues, these loci may play a larger role than for warfarin itself. These variants should be included in the construction of next generation predictive models for warfarin and anticoagulation pharmacogenomics, and the mechanistic investigation of the newly discovered pathway may also be valuable.

The genes discovered in this experiment cluster into a pathway. Thirty one genes and 74 SNPs, including 17 previously known genes **[Malatkova et al 2016, Wadelius et al 2007, Chen et al 2014, Bare et al 2011, Marsh et al 2006, Bader et al 2016, Wang et al 2013, Eriksson et al 2016, Lind et al 2012, Lip et al 1995, Patillon et al 2012, Bargal et al 2016, Kimmel 2010, Undas et al 2005, McDonald et al 2009, Lee et al 2012, Scordo et al 2002, Kim et al 2012,**

Chang et al 2015, Limdi et al 2008, Kudzi et al 2009, John et al 2017], cluster into a coagulation pathway mainly made up of known genes, including clotting factors, *CYP* enzymes, and vitamin K epoxide reductase. This may be considered the canonical warfarin pathway, although it appears not to have been explicitly described previously. Thus, the PIP has systematized and extended the previous literature into a more integrated picture. Because of the extensive pleiotropy observed with pharmacogenomics genes and variants, systematizing evidence is very important for integrative test design **[Denny et al 2013]**.

It is noteworthy that in the warfarin experiment, the PIP was successful in rediscovering both key pharmacokinetic (PK) and pharmacodynamics (PD) genes.

While pharmacogenomics has traditionally used more variants in PK genes, as discussed in the introduction, PD variants have proved to be extremely potent in the instances where they do find clinical use, including VKORC promoter variants used in current-generation warfarin genetic tests, which are the most potent variants used in such tests. Thus it is heartening to see extensive discovery of PD genes in the PIP, including extensive discovery of regulatory variants for PD genes. Discovery methods like the PIP may be able to aid the transition to make more use of PD genes in pharmacogenomics, as suggested previously **[Higgins et al 2015, Higgins et al 2015, Higgins et al 2015]**.

PK genes discovered for warfarin include *CYP2C9* and other key metabolic enzymes for warfarin. However, for PK genes, the PIP mostly discovered only coding variants, and not regulatory variants. We suspect this may be the case for two reasons. First, the *CYP* loci have a complicated

structure of homology with many similar genes in close proximity, which tends to confound both hybridization, linkage, and sequence alignment **[Danielson 2002, Hoffman et al 2002, Zanger et al 2013, Nelson et al 2013]**, all of which are essential steps in the PIP and the omics assays which provide its source information. Secondly, however, the current version of the PIP locates target genes by sequence proximity, rendering it unable to find distal regulatory elements like those which control many facultative metabolic enzymes. It is anticipated, therefore, that subsequent PIP enhancements will enable better discovery of regulatory variants for PK genes.

Warfarin is a major anticoagulant, but in recent years a host of other anticoagulants including other vitamin K analogue agonists as well as novel oral anticoagulants (NOACs) have been gaining in acceptance **[Barnes et al 2015]**. Many of these medications share known mechanistic elements with warfarin, and also have known differences in mechanisms **[Sangkuhl et al 2011, Wanat et al 2013]**. Various drugs that modulate the coagulation system are used for other purposes **[Wanat et al 2013, Harter et al 2015]**, including clopidogrel, which is routinely prescribed on a chronic basis to reduce the risks of stroke and heart disease. And some SNPs discovered in the PIP experiment for warfarin have also appeared in association experiments for other anticoagulants, especially vitamin K antagonists. For example, rs2108622, a coding SNP in CYP4F2 which the PIP identifies as a warfarin response SNP, is also a response SNP for acenocoumarol **[Teichert et al 2009]**. A joint analysis of the pharmacoepigenomic properties of this broader collection of coagulation-system-modifying drugs, integrating and comparing pathways and mechanisms, could be of real benefit to extend the breadth of future anti-coagulation pharmacogenomics testing.

More broadly, these results indicate that the epigenome-and-3D-nucleome-informed paradigm of pharmacogenomics that we described in our 2015 papers **[Higgins et al 2005, Higgins et al 2015, Higgins et al 2015]** may generalize beyond neuropsychiatry into cardiovascular pharmacogenomics and other domains. The genes and variants discovered with the PIP should be the subject of follow-up research in both pharmacogenomics test design and future anticoagulant drug development. Automated pharmacoepigenomics-based variant discovery and investigation methods like the PIP should continue to be developed, with potential future application in pharmacogenomics over the coming years.

Validation and Extension of the Pharmacoepigenomics Informatics Pipeline (PIP)

Our successful reproduction of the results of the lithium experiment with the PIP demonstrates that reproducibility in bioinformatics experiments depends closely on adherence to important reproducibility principles. Reproduction of previous lithium results was complicated by elements of the original experiment like the inclusion of manual steps, consulting resources from external databases without retaining them or documenting precisely which tracks were consulted, and the use of databases whose contents change in an un-versioned manner. These lessons have been incorporated into the PIP with design features like the use of versioned, locally stored databases, the selection of resources and records in an algorithmic and reproducible manner, and the retention of intermediate results and code. This ensures that PIP experiments are reproducible. Remaining challenges, including the use of the un-versioned IPA database for pathway analysis, should be rectified in subsequent PIP refinements. Conversely, however, it is clear from our results that the

automated pipeline offers better functionality than the semi-automated pipeline it was designed to reproduce.

It is possible to make pervasive use of machine learning in gauging the variant dependence of enhancer function. The current pipeline gauges variant dependence only with PWM measurements, but a number of machine learning algorithms have recently been published for gauging variant function, including DeltaSVM **[Lee et al 2015, Ghandi et al 2014, Ghandi et al 2016]**, methods of Nishizaki **[Nishizaki et al 2017]**, and recently GKM-DNN **[Zhang et al 2017]**. We have performed testing with DeltaSVM indicating that its predictive power is "orthogonal" to that of PWMs, and would add value, both by allowing rejection of variants that pass the current scoring system but do not show variant dependence with DeltaSVM, and rescue of variants that barely miss the current thresholds but exhibit strong variant dependence in DeltaSVM. The integration of some of these published methods would strengthen the PIP.

In addition, methods like the PIP may help alleviate a vexing issue in current pharmacogenomics and other genetic testing: inapplicability of some test results across populations. Sometimes genetic associations discovered with association testing may hold in the population from which the study populations were drawn (predominantly European, although increasingly Asian populations are being used in research conducted in Asia), but not in other populations **[Huizinga et al 2004, Cooper et al 2008, Visscher et al 2017]**. In this case, only three of 23 studies included African populations. This has been an issue of comment both in the literature **[Condit et al 2003, Yasuda et al 2008, Urban et al 2010, Ortega et al 2014]** and in the lay press **[Stein 2011, Arthur 2017]**. Strategies for surmounting it have included including population-specific lead SNPs in test

design, designing separate tests for different populations, including population-linked SNPs in test design, and various combinations of these methods **[Daneshjou et al 2016]**. Warfarin response and coagulation phenotypes in particular are an area in which certain populations, including African Americans, have exhibited different response profiles and different GWAS lead SNPs from European populations **[Heit et al 2012, Heit et al 2017]**.

The PIP and similar methods may help to resolve this problem. One mechanism by which such an effect may arise is when a GWAS lead SNP is only a tagging variant for an underlying effector SNP, and the LD relationship between these SNPs is population dependent. In such cases, the PIP, by evaluating the population of population-specific LD partners of the lead SNP, and finding effector SNPs within this population, will sometimes enable tests to be designed on the basis of population specific GWAS which will function predictively in a population-agnostic manner.

Methods like these may generalize in other drug disease systems. PIP-style methods may identify trans-ethnic causal regulatory variants and recover missing heritability by interpreting association studies in the light of disease mechanisms and massive cohort omics. These high quality novel variants may help to enable the development of new pharmacogenomics tests with greater predictive power and clinical benefit.

Chapter 4: The H-GREEN Hi-C Compiler

Motivation: Distal Target Gene Finding for the PIP

The initial version of the PIP had a notable weakness: it relied on sequence proximity to find the target genes of regulatory variants. This decision resulted mainly from two things: a desire to replicate the featureset of the lithium analysis, and the lack of availability of automated methods for distal target gene identification. Nevertheless, this is important both because of widespread evidence that important phenotypes are mediated by genome-wide spatial and functional networks uniting genes from multiple chromosomes in a transcriptional program with unifying regulatory elements, but also because of the demonstrated value of distal target gene analysis in the "variants" and valproate variant analyses.

The construction of a multimodal target gene suite for a future version of the PIP rests on several pillars: spatial contact data, molecular QTLs, and machine learning. The value of both spatial contacts and QTLs in identifying enhancer control relationships, and the explosive proliferation of both types of data, have made the deep and comprehensive use of these data modalities in a genome wide tool like the PIP plausible. Moreover, new modes of analysis treating enhancer elements and genes as elements on a bipartite graph on a genome wide basis, along with high throughput

CRISPR screens for enhancer activity, have offered the possibility of using such high throughput contact data, along with the "ground truth" of CRISPR screens, in the context of network mathematics, to make target gene prediction amenable to matrix densification methods in machine learning.

All the elements of this emerging vision of multimodal target gene discovery are in place, except for one. Proximal target gene analysis with integrative analysis of Hi-C, epigenome, and QTL data is available in TargetFinder. Hi-C, CRISPR, and QTL datasets are increasingly available. Bipartite graph analysis tools are increasingly robust. Matrix densification is a well-studied area of active ongoing research in machine learning. However, there are no currently available automated tools for distal (tens of MB to trans) target gene analysis with Hi-C data.

For this reason, in 2015 I conceived and began the development of a new Hi-C compiler specifically designed for this challenge. Unlike other extant compilers, it uses functional elements of DNA as its compiling bins, maximizing the discernment of functional element mediated connections in the sparse off-axis regions of squared genome space. It was and remains my intention that these compiling methods be integrated into a PIP-style pipeline, and used with Hi-C data from relevant tissues to discover the long range spatial contacts of potent enhancer loci to identify their distal target genes.

Introduction:

The 3D Spatial Hierarchy of Chromosome Organization and Stratification of Drug Response

Noncoding regions of the human genome contain the infrastructure that regulates gene expression, and mutations within elements such as enhancers and promoters are critically associated with disease risk and drug response. Evidence that most SNPs (single nucleotide polymorphisms) in genomewide association studies (GWAS) and candidate gene studies that are significantly associated with disease risk, phenotypic traits and drug response have been mapped to enhancers emphasizes the importance of studying domains not found within the exome **[Zhang et al 2014, Raj et al 2008, Fukaya et al 2016]**. Regulation of gene expression depends upon the interaction between enhancers, transcription factors, chromatin remodeling proteins, histone modifications indicative of active chromatin, the ability of DNA-binding proteins to access a casual allele as indicated by DNase I hypersensitivity, TAD boundary proteins, and gene promoters within a TAD. The frequency of interactions of these elements within the spatial and temporal chromatin architecture of a TAD is much higher than are inter-TAD interactions **[Bonev et al 2016, Higgins et al 2017, Cloney et al 2016]**. The boundaries of TADs contain insulator proteins, including CTCF and cohesin **[McCarthy et al 2012, Bunney et al 2015]** and are timing transition regions **[Yan 2015]**. Mutations in TAD boundaries have been shown to cause disease in humans ranging from autosomal dominant Autosomal Dominant Adult-Onset Demyelinating Leukodystrophy (ADLD) **[Kino et al 2010],** to a variety of developmental limb malformations that disrupt enhancer-promoter interactions **[Pope et al 2014]**.

Differences in drug response and AEs among individuals represent an aggregate of pharmacodynamic and pharmacokinetic variation, but are most often attributable to altered regulatory and biotransformation pathways which are limited to a few tissues. Since variation in

genomic DNA sequence, which should be roughly the same across all tissues, cannot account for the bulk of the genetic contribution in pharmacogenomics, it is likely that cell type-specific differences in chromatin organization and interactions may be responsible **[Goel et al 2009, Chrousos 1995, Konturek et al 2011].** Previous studies using Hi-C demonstrate that the spatial organization of TADs appear to be invariant across tissues, but that TADs are either active or inactive based on developmental state, pathology and cell type **[Bonev et al 2016, Goel et al 2009, Chrousos 1995, Konturek et al 2011, Chen et al 2015, Dai et al 2016]**. During development, the binding of cell type-specific master transcription factors within subsets of TADs help determine cell fate **[Tjong et al 2016]**. Several drug classes, including HDAC inhibitors and chemotherapeutics, work by inducing genes that encode master transcription factors or other proteins that are important in determination of cell fate, leading to re-activation of developmental transcriptional programs **[Tjong et al 2016, Bharadwaj et al 2014]**.

Even in cases where an exon variant is linked to an AE, regulatory domains distant to the gene that harbors the mutation impact drug response. For example, 60% of patients with melanoma contain the V600E variant at amino acid position number 600 on the BRAF protein, in which the valine is replaced by glutamic acid. The chemotherapeutic drug vemurafenib interrupts the BRAF/MAPK (B-Raf Proto-Oncogene, Serine/Threonine Kinase-Mitogen-Activated Protein Kinase) pathway, and this drug is commonly used to treat patients with melanoma. However, drug resistance developed in almost all patients taking vemurafenib, as tumor cells switch to an alternate survival pathway. Using knock-out experiments in human cell lines and interruption of regulatory elements using CRISPR-Cas9, it was found that the *CUL3* gene is a major contributor to drug resistance in these patients **[Thakurela et al 2015, Hung et al 2011]**. Bombardment with small guide RNA-

mediated mutations in domains surrounding *CUL3* found an enhancer located 22 kb from the TSS of the gene, which was found to be responsible for the regulation of the gene. This could be identified as the culprit regulatory domain responsible for drug resistance to vemurafenib **[Hung et al 2011]**, leading to therapeutic strategies. For many genes, which encode ADME proteins, enhancer looping with promoters is not uncommon, as exemplified by the *SLC2A13* (Solute Carrier Family 2 Member 13) gene **[Chen et al 2015]**.

Although many regulatory interactions are constrained in *cis*, many long-distance interactions, including active inter-chromosomal spatial contacts, play a role in determination of pharmacogenomic response **[Chen et al 2015, Hung et al 2011, Zhang et al 2014]**. Inter-chromosomal spatial contacts are common **[Bush et al 2016, Denny et al 2010]**, and AEs associated with psychotropic drugs can be explained by these *trans*-interactions in some cases **[Zhang et al 2014, Hebbring et al 2015]**. Complementary methods such as Hi-C which employing deep sequencing for visualization of enhancer-promoter and promoter-promoter loops **[Higgins et al 2015], or** ChIA-PET **[Hebbring et al 2015]**, in tandem with super-resolution microscopy **[Bush et al 2016],** may enable closer examination of these functional pharmacoepigenomic interactions. It is probable that not all enhancer-promoter or promoter-promoter interactions involve chromatin loops **[Denny et al 2010].**

Hi-C sequencing and similar methods mapping contacts into squared genome space have emerged as the preeminent mode of gauging spatial organization of chromatin at a genome wide scale. Since the publication of the original Hi-C methods and results, experiments using these methods have proliferated in many biological systems at ever-increasing depths of sequencing, and resulted

in fundamental biological insights about the spatial structure of the genome, how it is regulated, its causative relationship with gene expression, and the ways that such connections operate in unusual tissue and disease contexts. In addition to this, there has been a proliferation of methodological enhancements and related methods, including in-situ Hi-C **[Rao et al 2014]**, capture Hi-C **[Mifsud et al 2015]**, ChIA-PET **[Li et al 2014]** and Hi-ChIP **[Mumbach et al 2016]**, Micro-C **[Hsieh et al 2015]**, SPRITE **[Quinodoz et al 2017]**, and Genome Architecture Mapping **[Beagrie et al 2017]**.

Hi-C works by reconfiguring DNA sequences in the human genome to associate by spatial proximity rather than genomic contiguity, using fixation, cutting, religation, and then digestion. When the reconfigured DNA is sequenced with paired end sequencing, the paired reads correspond to spatially proximal elements which may be distal in linear sequence space, or on separate chromosomes. The millions to billions of read pairs making up a Hi-C dataset are compiled into a contact frequency matrix in squared genome space, showing the contact frequencies for all pairs of loci in the genome.

Chief among the discoveries flowing from Hi-C has been the discovery of Topologically Associating Domains (TADs) **[Dixon et al 2012]** in the human and other genomes. These self-associating regions have been identified as potent spatial and functional elements in the human genome. The division of approximately 80% of the human genome into approximately 2500 TADs is remarkably robust, being largely conserved between cell types in the human body **[Rao et al 2014]**, between different humans **[Ruiz-Velasco et al 2017]**, and under disease states **[Rao et al 2014]**. In fact, syntenic regions of related genomes (e.g. mouse) often share the same TAD

structure as the related regions of the human genome **[Krefting et al 2017, Nora et al 2013]**. TADs also function as replication domains **[Pope et al 2014]**. Moreover, TADs mediate long range spatial interactions **[Rao et al 2014]**: the contact frequency in any given portion of the squared genome will more closely correlate with a more sequence-distant portion which is in the same TAD pair than a sequence-proximal portion spanning TAD boundaries. (**Figure 4-1**)

While the portion of a genome which composes a TAD is relatively invariant, TADs differ from one cell type and biological condition to another in their degree of transcriptional activity. This differentiation between the "A" and "B" compartments is connected with the sign of the dominant eigenvector of a genome-wide Hi-C matrix **[Dekker et al 2013]**, the degree of spatial openness as observed by both Hi-C, imaging, and biochemical experiments **[Roadmap Epigenomics Consortium 2014]**, the degree of gene expression **[Roadmap Epigenomics Consortium 2014]**, the presence of active histone marks **[Roadmap Epigenomics Consortium 2014]**, the presence of activating transcription factors **[Roadmap Epigenomics Consortium 2014]**, replication timing **[Pope et al 2014]**, and the extent of long range and interchromosomal contacts **[Rao et al 2014]**. Genes located in the same TAD tend to be co-regulated, and they are regulated partially by their TAD context.

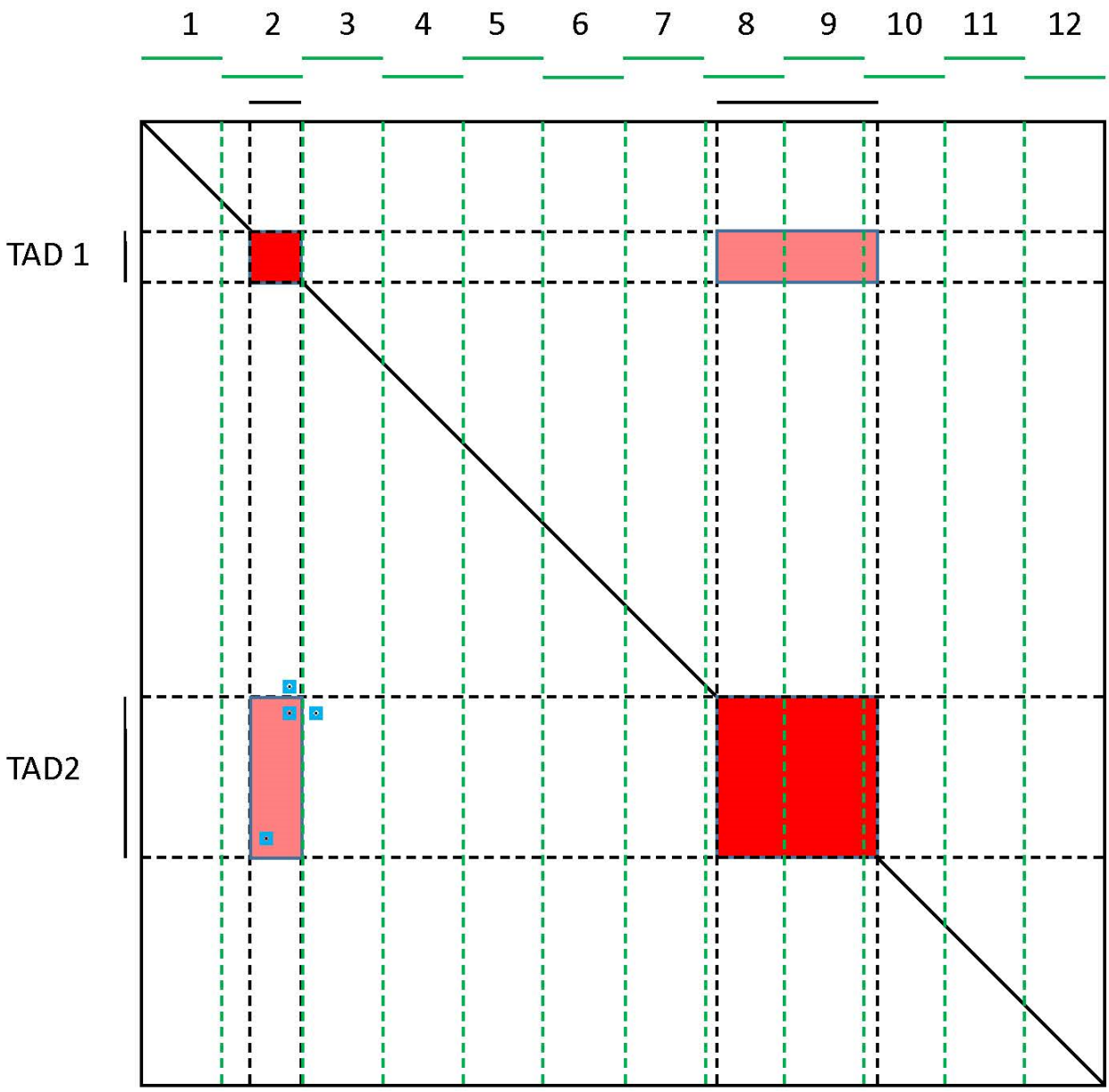


Figure 4-1: Schematic Representation of Hi-C Compiling Concepts

A schematic chromosome containing two TADs, designated TAD1 and TAD2, shown on top and at left. Along the diagonal the intra-TAD TAD pairs (1,1) and (2,2) are shown in red, while the inter-TAD interaction area of TAD pair (1,2) is shown in pink at the top right and bottom left. Also shown are TAD bins for TAD1 and TAD2, designated by the horizontal and vertical black dotted lines. Also shown are fixed bins 1-12, designated at top and by green dotted lines. It may

be seen that TAD1 is contained within Bin2, which also contains sequence area outside of TAD1, while TAD2 is divided into Bin8, Bin9, and Bin10, two of which also contain sequence area outside of TAD2. Accordingly, the TAD pairs (1,1), (1,2), and (2,2) are represented by one, three, and nine binpairs. Also shown are four positions in TAD pair (1,2), designated in blue, with the two positions in (1,2) having more correlated signal intensities, even though two positions in adjacent TAD pairs are closer in sequence space to the top right position in (1,2).

Moreover, the sequence context governing these regulatory interactions is beginning to illuminate under sustained investigation. High resolution Hi-C experiments have discovered a hierarchy of super- and sub-TADs running all the way down to the level of "loop domains" **[Rao et al 2014]** which spatially unite the proximal enhancers of genes with their intra-TAD proximal enhancers. These contacts, and those above them in the hierarchy, are largely governed by the presence of convergent pairs of CTCF sites **[Nichols et al 2015, de Wit et al 2015]** which form a CTCF-cohesin anchor binding loci together by means of a loop extrusion mechanism **[Sanborn et al 2016]** which has been verified by biochemical and imaging methods **[Fudenberg et al 2016, Sanborn et al 2016]**. Perturbations of these sequence elements by CRISPR in vitro, or by disease-related mutations in vivo, as for example in enhancer hijacking in cancer **[Northcott et al 2014, Weischenfeldt et al 2017]**, have the predictable effects on Hi-C maps, the epigenome, and gene expression **[Wutz et al 2016, Sanborn et al 2016]**.

Genes are often regulated by enhancer elements which are very distant from them in sequence space, referred to as distal-cis, or on different chromosomes, referred to as trans. Some recent opinion has argued that these interaction categories should be considered together, as sufficiently distal cis interactions are no longer mediated by sequence proximity and may be considered effectively trans interactions **[Wong et al 2017]**. In any event, long distance enhancer interactions, including those of "super enhancers" which regulate many genes on many chromosomes, are validated, and such validated interactions often appear in Hi-C contact matrices **[Rao et al 2014]**. There is increasing interest in using Hi-C maps to screen for such interactions.

However, current methods are not well tuned for this task. Functional elements from TADs to chromatin state segments have become central to the interpretation and use of Hi-C maps in biological science, and Hi-C read pairs are mapped into squared genome space at the resolution of cut-sites (in the case of traditional Hi-C, 4-cutter or 6-cutter restriction enzymes) or at base pair resolution (in the case of Micro-C **[Hsieh et al 2015]**). Despite this, the compiling of Hi-C read pairs into interaction matrices has largely proceeded by chopping the genome into evenly-sized pieces contiguous to each other, i.e. fixed binning **[Lieberman-Aiden et al 2009, Dixon et al 2012, Rao et al 2014]**. This method has the advantage of being easy to implement and producing matrices which are relatively easy to normalize, as well as sparing the researcher the task of justifying a choice of bins. It suffices for many purposes, particularly those analyzing on-axis measurements corresponding to proximal cis interactions.

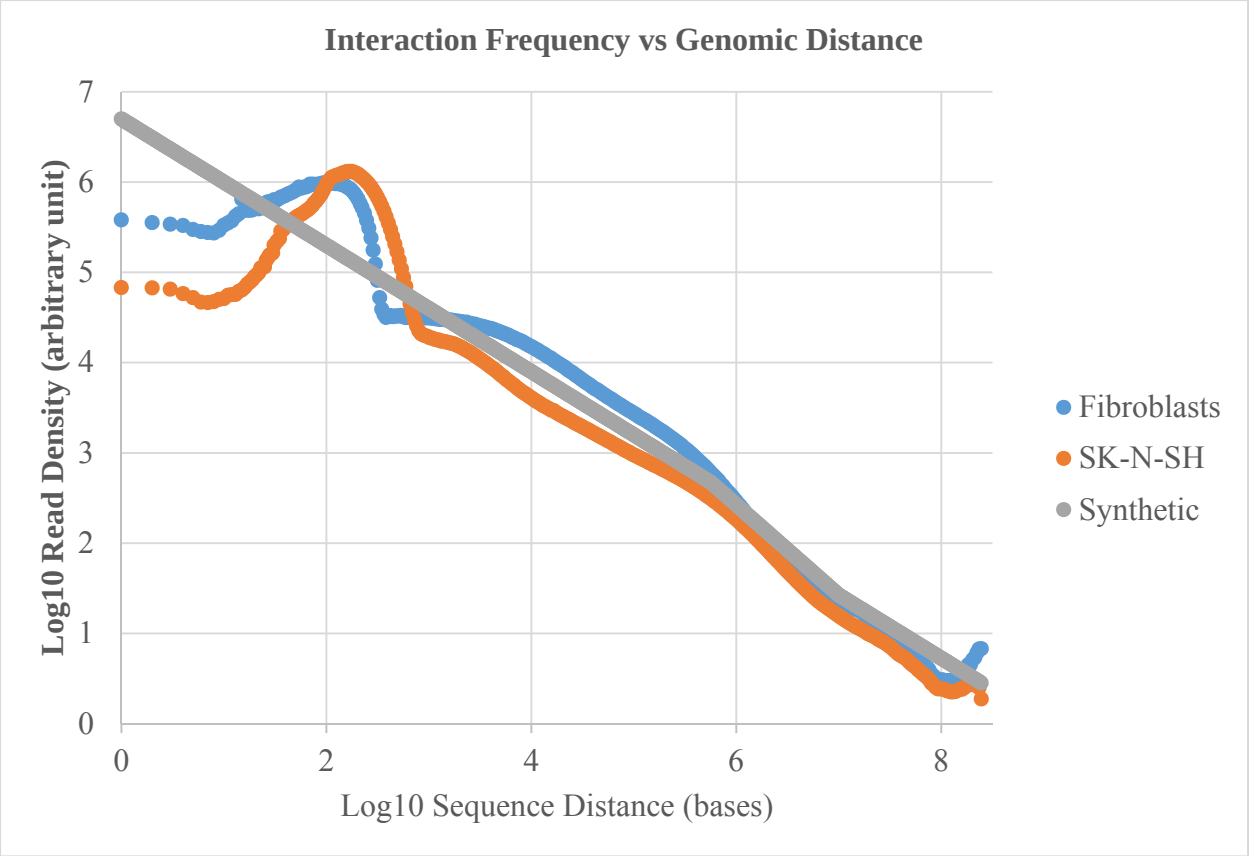


Figure 4-2: Normalization Curves in Hi-C Compiling

Shown are three distance normalization curves, representing the contact density (Y axis) as a function of sequence distance (X axis), on log10 based scales. The Y axis is denoted in arbitrary units representing multiples of the contact density in the trans (inter-chromosomal) domain. Shown in gray is a power law spline density function used for normalization in some Hi-C compiling methods, while empirical density curves derived from two real datasets (fibroblasts and SK-N-SH neural cells) are shown in blue and orange. The two empirical curves both show humps corresponding to the mononucleosome fragment, followed by monotonic curves with different slopes in different distance regimes, and a hump corresponding to the end joining of Chromosome 1. However, they differ substantially from each other and the synthetic curve, sometimes up to threefold, in a manner which differs by distance regime.

However, the method of fixed binning is poorly suited for the detection of long distance interactions between enhancers and their target genes. Trans and distal cis interactions suffer from severe data sparsity, as read pairs in a linear genome are mapped into a squared genome whose area is over nine million squared megabases. Moreover, read densities vary by five orders of magnitude with genomic distance and the majority of measured interactions are concentrated along the axis. Thus with fixed bins, fine resolutions will result in a genome-wide matrix over 99.9% of whose entries are empty, while coarse resolutions will abuse the functional elements which mediate long range contacts, chopping them into pieces and combining them with adjacent sequence regions. (**Figure 4-1**)

Accordingly, there is a need for Hi-C compiling methods designed to detect long range cis and trans interactions mediated by functional elements. To answer this need we present H-GREEN (Hi-C compiling with Genomic and Regulatory Elements and Empirical Normalization), a Hi-C compiler designed for distal contact detection. It compiles aligned reads into functional-element bins, normalizes them against an empirical density function, and detects contacts with statistical testing and comparison control.

Because the distal contacts sit in a regime wherein there is substantial variability in the distance contact density of different Hi-C maps, and a functional-element-based compiler will be forced to normalize bins of variable size and shape, it is essential that the normalization package of such a compiler make use of dataset-specific empirical distance normalization curve. (**Figure 4-2**)

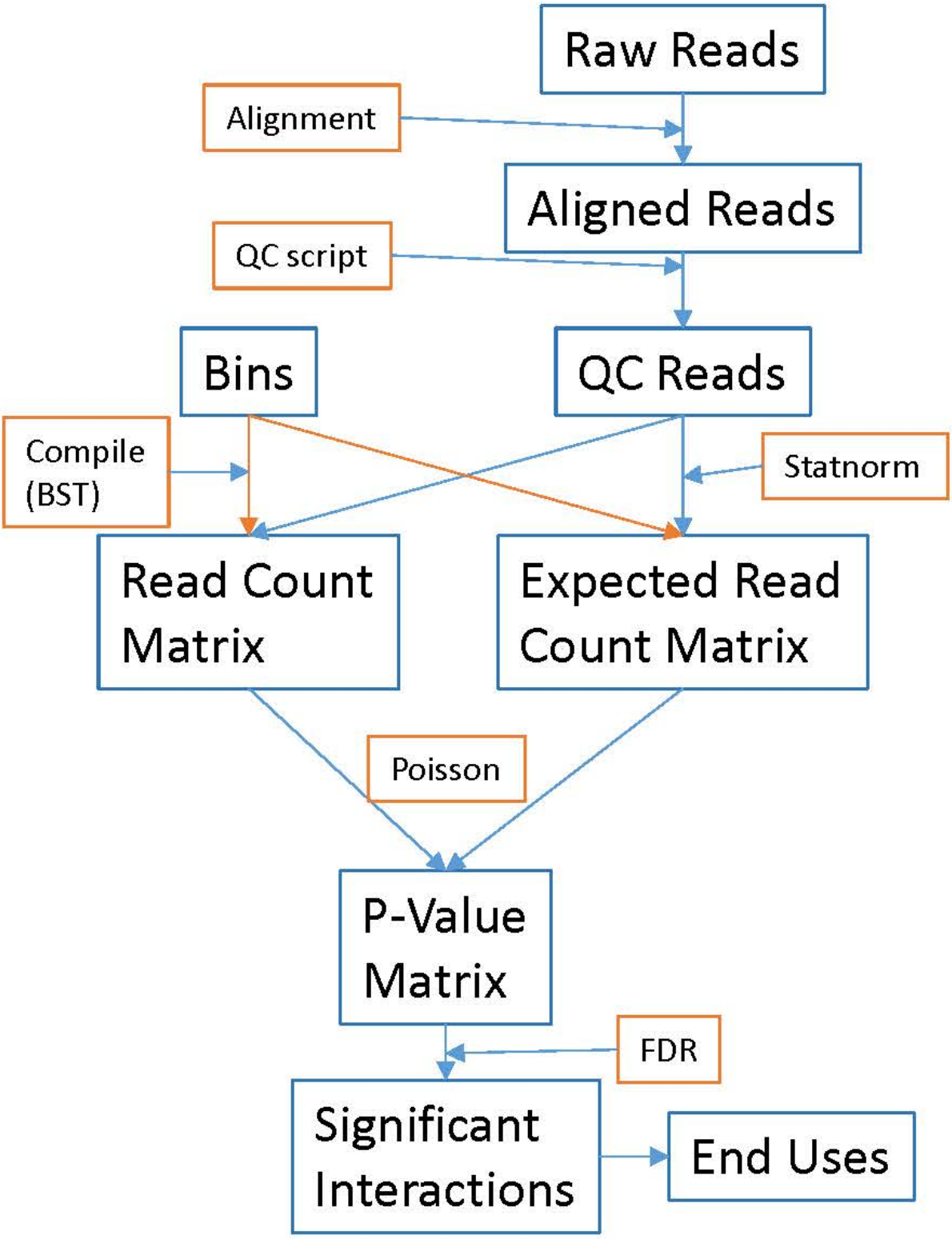


Figure 4-3: H-GREEN Methods Diagram

The steps in the H-GREEN Hi-C compiling algorithm, beginning with alignment and quality control of reads, followed by the separate compiling of real and expected read count matrices, followed by statistical normalization and the availability of significant enriched and depleted interactions for end uses.

Hi-C Compiling with Genomic and Regulatory Elements and Empirical Normalization

System and Methods, Algorithm, and Implementation:

The H-GREEN algorithm is summarized in **Figure 4-3**. It proceeds by way of alignment, quality control, compiling, integration, statistical testing, and result output.

Raw paired end reads imported in FASTQ files **[Cock et al 2010]** are split into their single-ended components and then aligned with Bowtie **[Langmead et al 2012]** using standard settings, with multiprocessing and reordering.

Then, the pairs are selected wherein both reads aligned singly with a quality score exceeding a threshold which would present only a .05 probability of misalignment in either read, and the unique read pairs satisfying this criterion are used in the subsequent analysis and recorded in a CSV file.

Aligned, quality-controlled reads from published experiments are sometimes published in the form of text spreadsheets with a ".nodup" **[Lieberman-Aiden et al 2009]** file extension instead of raw reads, and in addition, sometimes they are published together but the needs of an experiment with H-GREEN may dictate the use of the same reads and alignments as in another analysis. For this purpose, there is a script that converts "nodup" files directly into quality controlled read files.

Then, alignment proceeds. The bins are imported from a BED file of user designation, either a single set of bins for both axes, or two separate sets of bins, one for each axis. The quality

controlled read pairs are mapped into their host bins with a binary search tree, **[Booth et al 1960]** wherein the ordering relation for the node indices is an ordering on genomic positions, as opposed to integer indices. The read pairs wherein both reads are located in a bin for the relevant axis are mapped into the matrix.

Separately, an empirical background density function is generated for each single base pair distance in cis squared genome space, by dividing the number of read pairs at each distance by the amount of squared genome space at that distance. The same is done for trans interactions. Then, the background density function is integrated over the rectangular region in squared genome space corresponding to each bin pair, to produce an expected read count for this read pair based on its size, shape, and position.

Then, the expected and observed read counts for each bin pair are used to generate a p-value with a two tailed Poisson distribution. The collection of p-values are then subjected to multiple comparison control with the false discovery rate methods of Benjamini and Hochberg **[Benjamini et al 1995]**, and significant interactions (enriched and depleted) with a fold change of at least twofold are reported as significant interactions in a CSV spreadsheet.

The H-GREEN algorithm is implemented in Python, with external calls to Bowtie for alignment, with the exception of the binary search tree for compiling itself, which is implemented in C++. With TAD bins, the use of a BST in C++ offered a fifteenfold performance improvement over a ranked list search, and a hundredfold improvement over a naïve list search.

On a 6th generation Intel Xeon processor, for one lane of Illumina Hi-Seq 2500 worth of data, the computational time involved in executing H-GREEN with TAD bins on both axes is approximately 84 core hours for alignment, one core hour for quality control, two core hours for compiling, 120 core hours for integration, and one core hour for statistical control and data output. Notably, all phases of the process are data parallel so that a suitable implementation could run H-GREEN quickly. Where n is the number of read pairs and k the number of bins in a square matrix, the complexity of each step is approximately O(n) for alignment and quality control, O(n*log(k)) for compiling, O(k) for integration, and O(k^2) for statistical control and data output.

Superior Detection of Long Range Contacts

For a first evaluation of H-GREEN's discernment of distal contacts, we compared it directly with HOMER **[Benner et al 2018]**, a widely used Hi-C compiler, on a published dataset of human fibroblasts **[Chen et al 2015]**. With H-GREEN, we aligned the raw reads and compiled them into a matrix with TAD bins **[Dixon 2012]** on both axes. A set of enriched long range interactions was defined as those cis interactions at a range of greater than 10 megabases which were enriched twofold or more and passed comparison control, squared genome wide. We then compared these interactions with the detected interactions by HOMER in published matrices of normalized interaction frequencies for cis interactions in all human chromosomes, at one megabase resolution. In defining TAD-TAD interactions in the HOMER-compiled published data on the same experiment, we were presented with the complexity introduced by the fact that unlike H-GREEN, HOMER uses fixed bins, and many such bins may intersect with the same TAD boundary. To give HOMER maximal opportunity to detect such contacts, we defined a TAD pair as interacting

if any pair of bins containing parts of both TADs was enriched twofold in the HOMER-produced normalized matrix.

We chose a 1MB compile not because this method would be optimal for the purpose of TAD contact detection with traditional methods; it's well understood that higher resolution compiles with subsequent bin grouping are more optimal for detecting distal contacts **[Rao et al 2014]**. It is likely that the use of these post-compile methods with a higher resolution compile in HOMER would be capable of detecting many or most of the additional contacts H-GREEN detects, and rejecting those H-GREEN does not detect. However, this compile resolution is most comparable to the TAD compile in terms of the overall number of bins in the genome and average size of bins, among widely used compiling resolutions in fixed-binning Hi-C compiling. It therefore represents the closest direct comparison between fixed binning and H-GREEN's TAD binning methods.

At these settings, HOMER detects 18,220 contacts, and H-GREEN detects 17,720. **(Figure 4-4)** Of these, only 5,648 are common to the two methods. Contacts detected by H-GREEN and not by HOMER, of which there are 10,193, may be attributed to the superior discerning power of H-GREEN, which gauges all the signal from the TAD pair under inspection and does not dilute with signal from boundary regions and adjacent TADs. Contacts detected by HOMER but not by H-GREEN, of which there are 12,572, may be attributed to bins passing the thresholds which overlap one TAD pair due to a signal from an adjacent TAD pair. Among such pairs, 82% fail the fold-change cutoff in H-GREEN, 90% fail the FDR cutoff, and 72% fail both cutoffs. Among such pairs, 92% have a neighbor TAD pair which did have an H-GREEN detected contact. The non-neighbor discordant HOMER contacts number 1,054.

If indeed long range contacts detected in low resolution compiles with HOMER and similar methods are susceptible to false positives in this vein, then other methods without this problem should be preferred. Fortunately, however, the wide recognition that traditional methods are poorly suited for detecting distal contacts in sparse data has made such analyses uncommon, so the presence of such erroneous results in the literature is likely quite limited. H-GREEN and similar methods are more suited to such analyses.

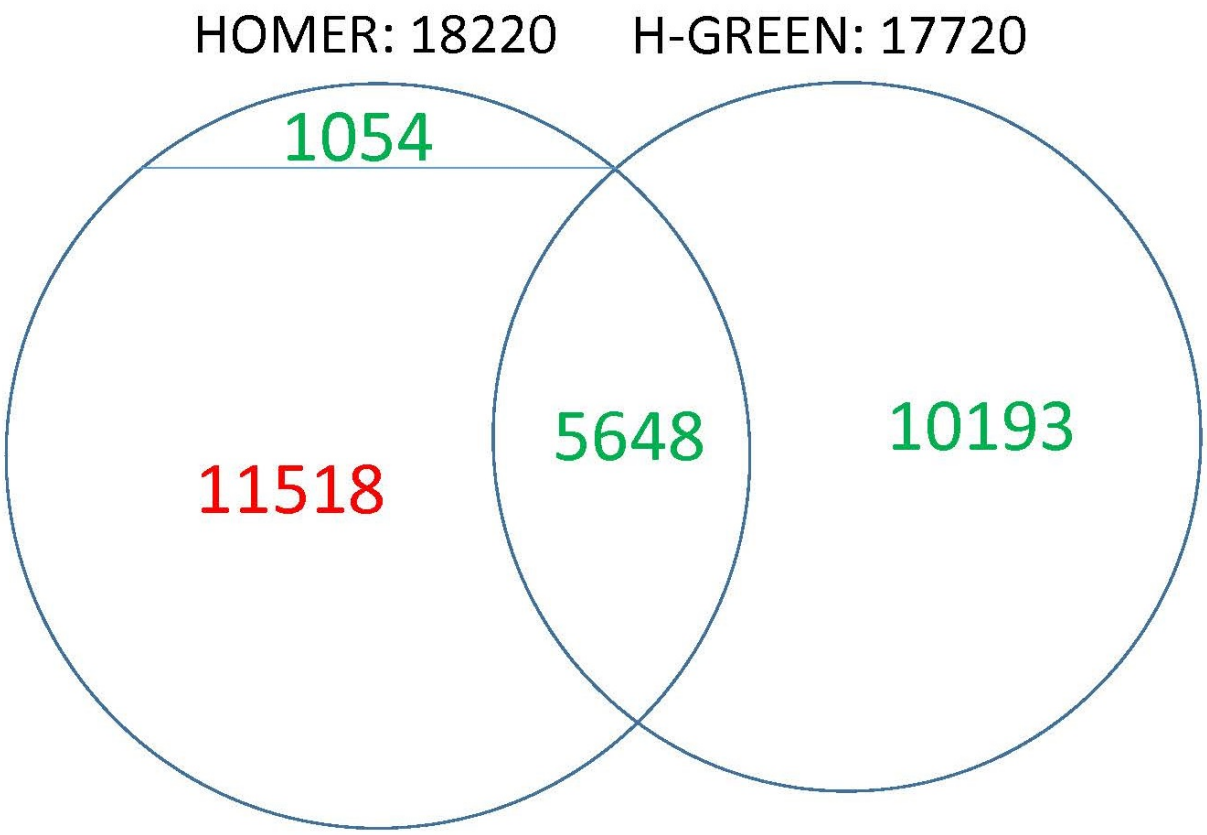


Figure 4-4: Contact Discernment: H-GREEN vs. HOMER

Direct comparison of contact discernment by H-GREEN and HOMER in human fibroblasts. Pictured are TAD-TAD contacts detected by H-GREEN and HOMER, divided into four categories: those concordant between the two compilers, those discordant and detected by H-GREEN only, and those discordant and detected by HOMER only, divided into those attributable to bleedover of neighboring signal, in red, and those not so attributable, in green.

rsID	Drug	Chr	Begin	End	TAD Num	Contacts	Gene	Candidate Contacts
rs3822158	Ketamine	chr4	157080000	159360000	693	5	*GLRB*	0
rs17161937	Ketamine	chr7	98880000	100320000	1063	388	*CYP3A43*	4
rs34547970	Ketamine	chr9	136840000	137800000	1343	172	*GRIN1*	4
rs12248560, rs77957608, rs61886222	Ketamine	chr10	94240000	95440000	1418	206	*CYP2C9, CYP2C19*	5
rs71311904	Ketamine	chr11	27480000	29360000	1472	73	*BDNF*	2
rs3764028	Both	chr12	13320000	14720000	1567	3	*GRIN2B*	0
rs7980427	Ketamine	chr12	108800000	109360000	1636	46	*DAO*	2
rs1543927	Ketamine	chr15	74040000	75000000	1832	218	*PPCDC*	3
rs7184748	Ketamine	chr16	9160000	10920000	1862	48	*GRIN2A*	0
rs2857654	Valproate	chr17	33720000	35360000	1977	47	*CCL2*	1
rs11083595	Ketamine	chr19	38400000	40920000	2112	682	*CYP2B6*	9
rs2269577	Valproate	chr22	28040000	29080000	2245	134	*XBP1*	2
rs28439297	Ketamine	chr22	41920000	43480000	2258	609	*CYP2D6*	8

Figure 4-5: Table of Neurogenesis-Gene-Containing TADs for Ketamine and Valproate Response

The TADs involved in neurogenesis in the ketamine and valproate SNP analyses, including the rsID of the enhancer SNPs, the drug for which these SNPs is an association hit, the genomic coordinates of the TAD, its number in the Dixon et al TAD numbering scheme **[Dixon et al 2012]**, the number of genome-wide TAD-TAD contacts in the SK-N-SH cell line, notable genes in the TAD, and the number of TAD-TAD contacts within the set of neurogenesis-related TADs. Pharmacokinetic TADs are shown in yellow; pharmacodynamics TADs in blue.

Automated Construction of Functional Interaction Networks

After verifying that H-GREEN is capable of detecting long range contacts not detected by traditional methods in the same datasets at comparable resolution, we proceeded with an exploratory experiment in a different biological system, to directly gauge the usefulness of TAD-TAD contact data in mapping regulatory circuits.

We assembled a list of thirteen TADs which play host to genes and variants involved in the pharmacokinetics and pharmacodynamics of two drugs, ketamine **[Soumier et al 2016, Huang et al 2016, Clarke et al 2017]** and valproic acid **[Higgins et al 2017]**, for which there is evidence of xenobiotic induction of neurogenesis in the adult human brain **(Figure 4-5)**. In a TAD-TAD compile of a published dataset of Hi-C on SK-N-SH neural cells **[Guo et al 2015]**, we looked for contacts involving these TADs.

Of the thirteen, all exhibited some distal contacts detected by H-GREEN **(Figure 4-5)**, from three contacts up to several hundred throughout the genome. Notably, TADs containing pharmacokinetic loci, e.g. the CYP genes which metabolize these drugs, appear to harbor the largest number of contacts. In addition, of the thirteen TADs, ten have connections to other TADs within the same set of thirteen, forming a densely interacting spatial network of functionally related TADs **(Figure 4-6)**. The P-value for such a density of contacts, an enrichment of over tenfold relative to the background, is less than 1e-14.

H-GREEN is capable of finding distal contacts that are suggestive of functional connections, and may function as a means of finding target genes of regulatory variants.

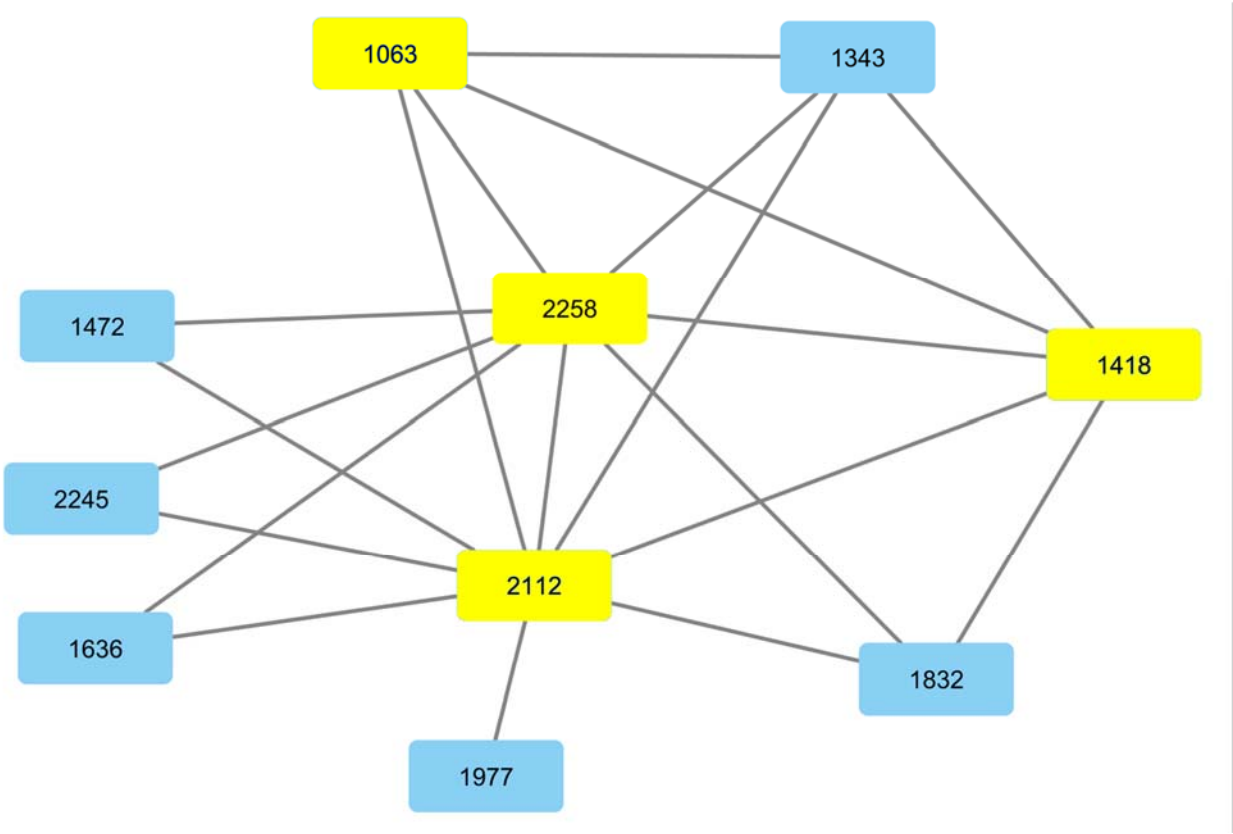


Figure 4-6: Interaction Network of Neurogenesis TADs in SK-N-SH Cells

The significant enriched interactions between neurogenesis-related valproate and ketamine response TADs, as called by H-GREEN in SK-N-SH cells. Ten out of thirteen neurogenesis TADs join into a network, at p<10^-14. The two TADs in the center of the network (2112 and 2258) contain metabolizing enzymes for ketamine and valproate which are expressed in the human brain. Pharmacokinetic TADs are shown in yellow; pharmacodynamics TADs in blue.

Discussion

H-GREEN is capable of finding spatial contacts suggestive of functional regulatory interactions, and of finding connections that cannot be found with traditional methods in the same datasets at comparable resolutions. Accordingly, these methods should be broadly used to find long range spatial contacts in Hi-C datasets, particularly for target gene detection for regulatory variants.

This can be done either by compiling a dataset with one set of loci of interest along an axis, e.g. the chromatin state segments containing association hits, and a genome wide TAD collection on the other, or by selecting the loci of interest out of a squared-genome-wide compile, as we have done in the present work.

In particular, the PIP will benefit substantially from the use of H-GREEN to find distal target genes. As noted in Chapter 3, the current PIP only detects proximal target genes and has no module to detect distal target genes, which is one of the most notable weaknesses of the current PIP featureset. Suggestive but inconclusive results show that this deficiency may be more critical in the detection of regulatory variants for pharmacokinetic SNPs relative to pharmacodynamics SNPs, as PIP experiments to date have consistently detected almost exclusively coding variants for pharmacokinetic genes, despite the existence of regulatory networks for PK genes. The discovery of a spatial contact network uniting regulatory variants for PK and PD genes together using H-GREEN thus provides a hopeful note that the incorporation of H-GREEN and other distal target gene methods into the PIP will address this critical need, as contemplated in Chapter 5. The finding of long range contacts for regulatory SNPs and the construction of spatiotemporal

regulatory contact networks has the potential to become a significant element of the standard methods for investigating the mechanism of regulatory variants, both across pharmacogenomics and in the broader field of human genetics as a whole.

It will serve to discuss another noteworthy Hi-C analysis package which attempts, like H-GREEN, to dispense with fixed binning in Hi-C analysis and thereby to analyze long range contacts, but which takes a radically different approach. The SHAMAN package **[Cohen et al 2017]** uses a sparse matrix at base pair resolution, and then generates a randomized matrix satisfying distance frequency and marginal coverage criteria sampled from the real matrix. It uses this randomized matrix to compare to the real one, generating p-values which are compared with FDR statistics to address random error in Hi-C matrices, similar to H-GREEN. However, the p-values are generated from the Kolmogorov-Smirnov D statistics for the density of the K-nearest-neighbor cluster around each of the individual read pairs in the database. Pairs with a significantly dense K nearest neighbors may be considered enriched. The selection of the K value for a particular experiment thus represents a significant tradeoff between resolution and statistical power, much like the selection of bin sizes in H-GREEN or in traditional Hi-C compiling.

This approach shares many advantages with the H-GREEN approach: independence of fixed-binning artifacts, sensitivity to low density, and natural scaling of statistical power with read density. It is therefore certainly superior to coarse fixed binning for genome wide analysis, and probably more useful than H-GREEN for intra-TAD and other proximal work. However, for distal cis contacts, it has one important disadvantage relative to H-GREEN, because it does not account for the mediation of contacts by large sequence elements. The K nearest neighbors of a particular

read pair may not be significantly enriched, while the entire TAD pair in which the read pair sits may be enriched. For a suitable value of K, these will be approximately concordant, but SHAMAN provides no way to choose such a K, which will in any event vary genome wide. In addition, a read pair adjacent to a TAD pair with strong clustering may "stow away" on the sequence-close dense read pairs, generating neighbor-spillover contact detections, just in the manner of fixed binning.

The developers of SHAMAN acknowledge these problems. They attempted to detect long range contacts and noted that they were mediated by larger domains: "Using the D scoring scheme, it is possible to screen for contact enrichment hotspots throughout the entire Hi-C matrix without prior assumptions. We applied this approach to the high depth lymphoblasts and ertirhroid maps, deriving 462 and 1304 non-overlapping contact enrichment hotspots, respectively ranging in distance from 4Mb to 100Mb (Methods). We observe that many of the identified hotspots were characterized by a broad pattern of enrichment, associating together genomic elements at a much larger scale than observed above for CTCFs or TSSs…This confirmed the broad nature of nearly all long-range contact hotspots we identified, showing contact enrichment at distance range of 100kb around the center, and in many of the cases larger distances are observed. Highly localized contacts between elements genomically separated by more than 4MB are very rare."

It may be possible, in the fullness of time, to attain a synthesis of the methods of SHAMAN and H-GREEN by adapting the SHAMAN algorithm to use a distance metric function which is not sequence-agnostic, and accounts the distances between neighboring pairs according to their

insulating score in local Hi-C mapping. This would tend to ensure that the nearest neighbors of any position fell into their local genomic elements, so that the algorithm would be more capable of detecting contacts mediated by these elements, as opposed to detecting nearby contacts of unrelated elements, or missing such contacts.

It is possible to envision many applications of H-GREEN and similar methods which would necessitate some enhancements to the current H-GREEN package. While the TAD bins used in the present experiments do not vary significantly by cutsite density or GC content, the use of smaller bins, e.g. gene binning or chromatin state segment binning, which do vary significantly, would require these elements to be accounted for in the normalization package in order to avoid major errors. This would be a straightforward extension using methods already used to adjust for these factors in traditional normalization. Coverage normalization is probably inadvisable as it would adjust away significant variation in the overall extent of genomic spatial contact which is observed to exist.

In addition, with the multiple sampling distribution of the Poisson distribution, it is possible for the H-GREEN statistical framework to be used to detect differential contacts between two Hi-C datasets in different biocellular systems or biological conditions. This could include, e.g., a survey of a large number of cell types based on published Hi-C maps. This type of bodywide atlas of spatial connections will be particularly important in light of upcoming compendia of tissue Hi-C data such as the upcoming Human Cell Atlas **[Regev et al 2017]**.

H-GREEN was designed to overcome data sparsity and make maximal use of sparse signal for contact detection. It may be that a version of these methods with a distance correction method optimized for this use may be useful in analyzing single-cell Hi-C data.

Hi-C data is used for a large and growing number of purposes, and with the added discerning power of H-GREEN and similar methods, surveys of long range contacts, both for specific loci and on a squared-genome-wide basis, may be added to this collection, to the potential benefit of many areas of bioscience research.

Chapter 5: The Pharmacophenomic Atlas

Pharmacogenomics in the Age of Omics Atlases, Biobanked EMRs, and Artificial Intelligence

The Pharmacoepigenomics Informatics Pipeline and associated tools and methods have added value in the identification of causative enhancer variants for phenotypes of interest, and the identification of their target genes for future mechanistic work. The variants, genes, and networks discovered for warfarin, lithium, valproic acid, and other drug-disease systems could be used directly in traditional test design, and other phenotypes can be investigated with these tools.

However, this featureset is not the last word. Advances in a number of areas, including the underlying biology of enhancer function, artificial intelligence, omics atlases, biomedical ontologies, and genotyped medical records from biobanks and clinical trials have opened the doorway to much more powerful PIP-style pipelines, which are more sensitive to interactions which cannot be detected by the current PIP, and more thorough in pruning out interactions which are not as promising. More, they have opened the door to a future wherein such pipelines are used pervasively in parallel against thousands of biomedical phenotypes, with results that can be used in routine medical practice.

This chapter describes an evolved PIP featureset which is based on current cutting edge methods in all the subdomains of biomedical inference from which the PIP draws, and some notions of how to score variants and genes and pathways in an evolved PIP with machine learning. It describes an orienting framework for how to use the output of PIP-style pipelines in biomedical genetic test design. And finally, it concludes with a vision for the use of such a pipeline with biobank datasets and biomedical ontologies to create a genome-wide, phenome-wide pharmacophenomic atlas of predictive models for thousands of phenotypes, which could be then be implemented directly in a clinical decision support system.

The availability of genetic prediction on a routine basis for medically important phenotypes, shortening the translation cycle on genetic discovery, can improve patient care.

An Evolved PIP Featureset

The underlying scientific domains of epigenome regulation, spatial genomics, and population genetics on which the PIP is based have not stood still since the PIP featureset was laid down in 2016. They have continued to advance. And while a future evolved PIP featureset will still be based on the overall Five Box Model of regulatory variant discovery, every portion will be affected by these significant discoveries. In addition to this, the universe of data from which a future PIP-style pipeline can draw has expanded vastly.

This will begin with data input: unlike the current PIP, which treats all variant inputs and all tissues uniformly, an evolved PIP would separately take tissue inputs for a collection of related phenotypes, and separate inputs of the relevant variants and populations for each phenotype, as well as information about the degree of relatedness of phenotypes to each other. In addition, the directionality of the variant effect is important: an evolved PIP would track which of the alleles of the SNP had which effect on the phenotype, and in the context of the variant dependence portion of the ERV workflow, the directionality of effects on TF binding and enhancer function will be gauged as well. The concordance of this information within the scores for each SNP, and between the common regulatory SNPs for a gene, will function as important information in the context of scoring.

Such a PIP would have the opportunity to draw on a wealth of data sources on which the current PIP does not draw. In addition to the resources of the current PIP, new and not-previously-available datasets which would add significant value would include:

- Primary GWAS and genotyped cohorts for relevant phenotypes
- Expanded collections of genomes for linkage analysis under the auspices of various biobanks and genomics initiatives
- Expanded omics atlases with deeper omics on more tissues and cell types, under the auspices of the International Human Epigenome Consortium (IHEC) **[Stunnenberg et al 2016]** and the upcoming Human Cell Atlas **[Regev et al 2017]**
- Expanded atlases of TFBS motifs and binding data, under the auspices of MotifDB **[Shannon et al 2018]** and the omics atlases

- Collections of Hi-C data from the Hi-C laboratories (currently highly decentralized), and soon in the form of a bodywide atlas from the Human Cell Atlas
- Expanded collections of molecular QTLs (currently highly decentralized) under the auspices of various biobanks, clinical trials, consortia, and individual analyses
- Libraries of validated enhancers
- The updated pathway mapping libraries of IPA **[Kramer et al 2013]** and other pathway mapping methods

A conceptual featureset for such a pipeline is shown in **Figure 5-1**, and described in more detail below.

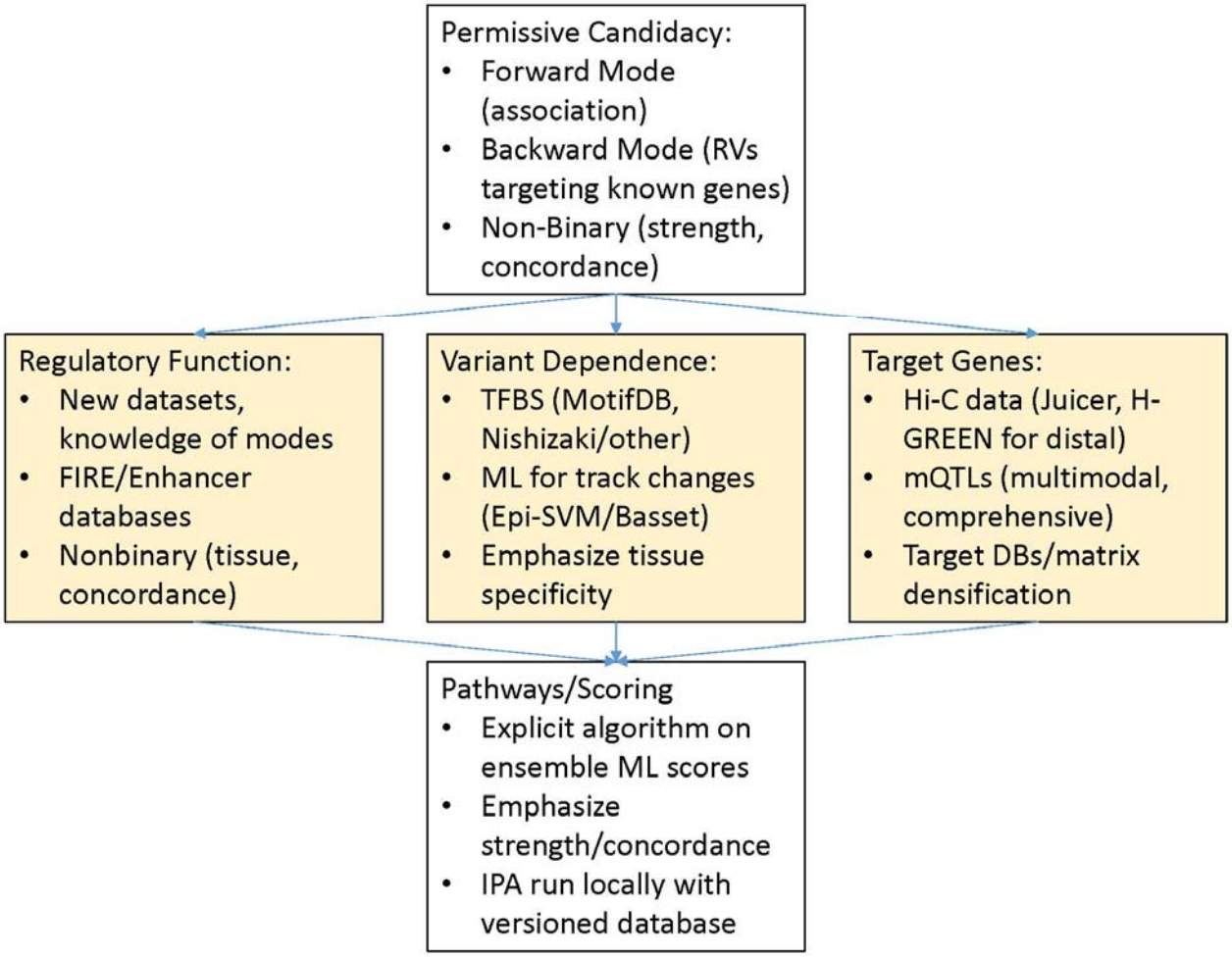


Figure 5-1: Conceptual Featureset for a Next-Generation PIP

Schematic visualization of the envisioned featureset of a next-generation PIP, organized according to the five major elements of the five box model of regulatory variant discovery.

Permissive Candidacy

Currently, all PCV identification in the PIP proceeds in a "forward" mode, proceeding from the phenotype toward a locus by association. This is currently done with GWAS lead SNPs by population-specific linkage analysis.

This forward mode of PCV discovery can be expanded. For one, linkage analysis with 1000 Genomes **[1000 Genomes Project Consortium 2015]** populations can be enhanced with the use of more relevant population groups as the number of analyzable genomes continues to expand, and the granularity of population genetics mapping continues to refine. For another, while linkage analysis has been the best approach for secondary analysis of GWAS for which only topline results are available, primary analysis of GWAS and CGAS can offer a more granular approach, when such data are available. Rather than using linkage, the collection of associated variants can be used together, in sum.

Moreover, however, a second mode of PCV identification, the "backward" mode, can take on a more prominent role. In the current PIP, we identify body SNPs of known genes of mechanistic importance in a phenotype as PCVs. In fact, however, what is desired is not body SNPs, but those SNPs which alter the sequence or expression of these genes, i.e. both coding SNPs and regulatory SNPs for this gene, regardless of their position in the genome. Thus, the evolved featureset can take gene inputs, just like the current one, and look for not only coding variants, but also tissue-specific regulatory variants for these genes, no matter where they may be in the genome, using the same target gene tools as the ERV workflow to find PCVs.

Moreover, although the current PIP treats all PCVs equally, PCV status need not be a binary. The strength of a PCV relationship may be modulated by a number of factors. For example, currently all affiliated phenotypes are weighted equally; in the Warfarin analysis, this included not just response and adverse events, but also disease risk and background phenotypes. In an evolved featureset, we would wish to give more weight to variants for more directly related phenotypes.

But for another, the strength and directionality of identification matters. For example, a variant exhibiting a lower p-value in a source GWAS, or stronger linkage to a lead SNP, is probably a stronger candidate. So, too, is a "backward mode" variant with a stronger target gene relationship, or a relationship with a more validated target gene. In addition, a variant identified through both "forward" and "backward" methods may be considered a particularly strong candidate.

Regulatory Function

In the current PIP, variants are evaluated for regulatory function in a rigid way: they must have one a promoter or enhancer chromatin state in a relevant tissue, and each of a collection of relevant histone marks in a relevant tissue, and be located in accessible chromatin in a relevant tissue. However, this status is awarded in a simple binary manner, and with no requirement that these tissues be concordant with each other or with other tissue-specific information.

In an evolved featureset, in addition to using larger datasets with higher quality and more relevant data (e.g. on tissues of more granularity), the criterion could enforce concordance between the

tissues for each omics modality. It would also serve to look for membership in databases of validated enhancers, and particularly potent categories of FIREs and super enhancers. It would place more emphasis on the features whose relevance has been emphasized in recent work, particularly chromatin accessibility, Hi-C contacts, TF binding, and chromatin state, over individual histone marks and DNA methylation, which have been downplayed.

In addition, of course, this status need not be binary (as in the current PIP), but could scale with the strength of the appearance of regulatory status, the level of inter-tissue concordance, and the degree of relevance of the tissues to the phenotypes for which the PCV attained PCV status.

Variant Dependence

The current PIP uses only one modality for variant dependence analysis, and this modality is both limited, and poorly evaluated. This is PWM analysis, using the Kellis et al library of PWMs **[Ward et al 2016]** along with the software TFM-Scan **[Liefooghe et al 2006]** to look for alterations. The rationale for this type of analysis is strong: variants altering TFBS are frequently potent enhancer variants **[Higgins et al 2015]**, and PWMs are a potent means of gauging TF binding affinity **[Stormo et al 1982]**. In addition, the "variants" analysis **[Higgins et al 2015]** showed that concordant changes in TFBS affinity for enhancers located at ChIP-seq-validated binding sites were a good mark of regulatory variant function.

However, the current methods in the PIP are weakened by an obsolete database of PWMs, as well as a crude and unmaintained algorithm for evaluating conformity. In addition, there is a format

conversion from probability to frequency matrices which introduces some imprecision (**Chapter 3**). In addition to this, although the current PIP tests for conformity and tests for TF binding, it does not value concordance between these measures.

An evolved PIP featureset would use a comprehensive versioned and updated PWM library in its native format, along with compatible and maintained code for PWM conformity. MotifDB **[Shannon et al 2018]** is one possibility. However, it may also be advisable to consider abandoning PWMs entirely in favor of a machine learning based method for gauging TFBS occupancy as a function of sequence, such as the Nishizaki algorithm **[Nishizaki et al 2017]**, which can gauge both sequence effects and cell type effects (by using epigenome data).

A more significant advance would be the use of machine learning methods to gauge variant effect on both epigenome tracks, chromatin state, and enhancer function. Several machine learning applications have been developed for predicting the impact of non-coding SNPs in GWAS on phenotypes; however fewer than 40% of GWAS publications from 2015 utilized these tools **[Nishizaki et al 2017]**. **Figure 5-2** lists some examples of deep learning software that score features, such as DHS for prioritization of regulatory function and protein annotation of chromatin loops, to predict functional enhancer-promoter interactions and drug-target inference.

SOFTWARE [REF]	SOURCE CODE	DESCRIPTION
DeepSEA [Zhou et al 2015]	http://deepsea.princeton.edu	Predicts the non-coding variant effects *de novo* from sequence by directly learning a regulatory sequence code from large-scale chromatin profile data, enabling prediction of chromatin effects of sequence alterations with single-nucleotide sensitivity.
DeepBind [Alipanahi et al 2015]	http://tools.genes.toronto.edu/deepbind/	Predicts potential transcription factor binding sites (TFBS) and RNA binding protein (RBP) binding sites; both *in vitro* and *in vivo*, outperforming 26 previously tested algorithms.
deepnet-rbp [Xu et al 2015]	https://github.com/thucombio/deepnet-rbp	Predicts RBP binding sites taking (predicted) RNA tertiary structural information into account.
Basset [Kelley et al 2016]	http://www.github.com/davek44/Basset	Predicts DNA accessibility, simultaneously learning the relevant sequence motifs and the regulatory logic with which they are combined to determine cell-specific DNA accessibility. Predictions for the change in accessibility between variant alleles are greater for Genome-wide Association Studies (GWAS) in SNPs that are likely to be causal relative to nearby SNPs in linkage disequilibrium with them.
DanQ [Quang et al 2016]	https://github.com/uci-cbcl/DanQ	Uses the same features and data as the DeepSEA framework, outperforming DeepSEA for 97.6% of the targets.
DeepChrome [Singh et al 2016]	https://github.com/QData/DeepChrome	Predicts gene expression from histone modification signals and enables the visualization of the combinatorial interactions among histone modifications via a novel optimization-based technique that generates feature pattern maps from the learned deep model.
TFImpute [Qin et al 2017]	https://bitbucket.org/feeldead/tfimpute	Predicts cell-specific TF binding for TF-cell line combinations using a multi-task learning (MTL) setting to use information across TFs and cell lines.
Rambutan [Schreiber et al 2017]	https://github.com/jmschrei/rambutan	Predicts Hi-C contacts at 1 kb resolution using nucleotide sequence and DNase I assay signal as inputs. Predicted contacts exhibit expected trends relative to histone modification Chromatin Immuno-Precipitation-Sequencing (ChIP-seq) data, replication timing measurements, and annotations of functional elements such as enhancers and promoters.
CpGenie [Zeng et al 2017]	https://github.com/gifford-lab/CpGenie/	Produces allele-specific DNA methylation prediction with single-nucleotide sensitivity that enables accurate prediction of methylation quantitative trait loci (meQTL). Contributes to the prediction of functional non-coding variants, including Expression

		Quantitative Trait Loci (eQTL) and disease-associated mutations.
DeepCpG [Angermueller et al 2017]	https://github.com/cangermueller/deepcpg	Identifies known and *de novo* sequence motifs that are predictive for DNA methylation levels or methylation variability, and to estimate the effect of single-nucleotide mutations.
iDeep [Pan et al 2017] and iDeepS [Pan et al 2017]	http://www.csbio.sjtu.edu.cn/bioinf/iDeep/ https://github.com/xypan1232/iDeepS	Predicts RBP binding sites by multimodal learning from multi-resource data, e.g. sequence, structure, domain specific features, and formats. Allows one to automatically capture the interpretable binding motifs for RBPs.
FactorNet [Quang et al 2017]	https://github.com/uci-cbcl/FactorNet	Predicts TFBS by leveraging a variety of features, including genomic sequences, genome annotations, gene expression, and single-nucleotide resolution sequential signals, such as DNase I cleavage data.
Basenji [Kelley et al 2017]	https://github.com/calico/basenji	Predicts cell type-specific epigenetic and transcriptional profiles in large mammalian genomes from DNA sequence alone. Identifies promoters and distal regulatory elements and synthesizes their content to make effective gene expression predictions. Model predictions for the influence of genomic variants on gene expression that align well to causal variants underlying eQTLs in human populations; and can be useful for generating mechanistic hypotheses to enable GWAS loci fine mapping.
Concise [Avsec et al 2017]	https://github.com/gagneurlab/concise	Predicts RBP binding sites using a spline transformation-based neural network module to model distances from regulatory sequences to genomic landmarks.
DeepATAC [Hiranuma et al 2017]	https://github.com/hiranumn/deepatac	Predicts binding locations from both DNA sequence and chromatin accessibility as measured by ATAC-seq, outperforming current approaches including DeepSEA.

Figure 5-2. Machine Learning Algorithms for Gauging Variant Effects

Examples of open-source deep learning software applications for the discovery of epigenomic regulatory interactions and variant annotation

Deep learning applications for detection of regulatory elements within the non-coding genome are beginning to emerge **[Angermueller et al 2016, Ching et al 2018, Park et al 2015]**. Most of existing applications are based on CNN architecture is trained either from *k*-mers **[Min et al 2017, Cao et al 2017]**, or directly on genomic sequence data. For example, DeepSEA **[Zhou et al 2015]** is one of the first deep learning-based algorithmic frameworks for predicting the chromatin effects of sequence alterations with single nucleotide sensitivity. In addition, it is trained on diverse sets of chromatin profiles from ENCODE and Roadmap Epigenomics Consortium projects **[Roadmap Epigenomics Consortium 2015, ENCODE Project Consortium 2012].** DeepSEA can accurately predict the epigenetic state of a sequence, including transcription factors binding, DNase I sensitivities, and histone marks in multiple cell types. In addition, it can further utilize its capabilities to predict the chromatin effects of sequence variants and prioritize regulatory variants. In another example, the DeepBind algorithm was implemented based on a deep convolutional neural network to calculate the ability of nucleotide sequences to bind transcription factors and RNA-binding proteins in order to characterize the effects of single point mutations on binding properties in various diseases **[Alipanahi et al 2015]**. More recently, the Basset CNN model was used to predict DNA accessibility within non-coding regions **[Kelley et al 2016]**; and is intended to predict allele-bias in DNA accessibility, which is indicative of causal variants**.** DNase-Seq data from 164 cell types that had been mapped by ENCODE and the Roadmap Epigenomics Consortium was used to create Basset. The Basset CNN learned both protein-DNA binding motif information, as well as the underlying regulatory knowledge that determines cell-specific DNA accessibility. In the analysis of GWAS SNPs that were determined to be casual autoimmune variants, Basset demonstrated that it could discriminate causal from non-casual SNPs in high Linkage Disequilibrium (LD). In contrast to inference of regulatory elements using annotation

based on pre-defined feature sets, models such as DeepSEA and Basset do not take handcrafted, preprocessed features. Instead, they adaptively learn them from raw sequence data during the training phase. This, combined with high expressive power, allows deep learning to outperform traditional machine learning models. More accurate prediction of non-coding variants and their functional annotations with deep learning methods promises to enable better understanding of pharmacoepigenomic variation and more accurate prediction of drug response and adverse events (AEs).

Other recent applications of deep learning models to prediction of regulatory elements and their interactions with the state-of-the-art performance include enhancer prediction **[Kim et al 2016, Xu et al 2016, Liu et al 2016]**; classification of gene expression using histone modification data as input **[Singh et al 2016]**; prediction of DNA methylation states from DNA sequence and incomplete methylation profiles in single cells **[Angermueller et al 2017]**; prediction of enhancer-promoter interactions from genomic sequence **[Singh et al 2016]**; prediction of DNA-binding residues in proteins **[Jiyun et al 2016]**; global transcription start prediction **[Eser et al 2016]**; and improved prediction of the impact of non-coding variants on DNA methylation **[Zeng et al 2017, Eser et al 2016]**. In 2016, Google and Verily Life Sciences published a pre-print describing "DeepVariant" – a deep learning-based universal SNP and small indel variant caller that won the "highest performance" award for the SNPs Administration-sponsored variant calling "Truth Challenge" in May 2016 **[Poplin et al 2018]**. Recently, an updated, open-source version of DeepVariant has been further evaluated on a diverse set of additional tests by DNAnexus **[Carroll et al 2017]**. These tests showed that application of a general deep learning framework exceeded the accuracy of traditional methods for SNP and indel calling that has been developed over the last

decade. Deep neural networks also demonstrated the ability to outperform conventional machine learning techniques in SNP–SNP interaction prediction **[Uppu et al 2016, Uppu et al 2016]**.

Since the publication of GKM-SVM **[Ghandi et al 2016]** and its variant DeltaSVM **[Ghandi et al 2014]** in 2014, such applications have proliferated. The use of DeltaSVM to gauge variant effect on chromatin accessibility and therefore enhancer function was used successfully in the walproate variant analysis, and exploration of the PCVs from the warfarin PIP experiment showed that the predictive power of this metric was orthogonal to the PIP. While such applications have proliferated and address a variety of different factors, the type which are most likely to be useful in the context of a PIP-style pipeline would be those directly relating to enhancer function or chromatin state, and which are trained on tissue specific data. This would probably be Basset **[Kelley et al 2016]** or DeepSEA **[Zhou et al 2015]**, or similar methods.

Variants that modified TFBS for TFs with binding activity at those loci, and with predicted allele bias, would have very robust evidence of allele dependence in the regulatory function of the host loci. In addition, evidence of concordant directionality between TF binding and enhancer function predictions will be an important form of evidence.

Target Genes

The target gene module in the current PIP, which evaluates target genes only by sequence proximity, is inadequate. Both the "variants" and valproate analyses showed the value of finding

target genes with Hi-C data, and recent work on target gene analysis with molecular QTLs and Hi-C data, along with machine learning, has been extremely fruitful.

An evolved PIP would locate target genes of regulatory variants with four methods:

1) Databases of CRISPR-validated **[Lopes et al 2016, Gasperini et al 2017, Klein et al 2018]** enhancer targets.

2) A comprehensive genome-wide and tissue-specific collection of molecular QTLs, comprising expression QTLs, DNA methylation QTLs, DNase accessibility QTLs, histone acetylation QTLs, and other modalities of molecular QTLs. Such a library should address major eQTL mapping experiments from GTEx and the national biobanks. A QTL relationship between a SNP and the expression of a gene or the epigenomic status of the gene body is indicative of an enhancer relationship.

3) Hi-C based methods for proximal target identification, such as TargetFinder **[Whalen et al 2016]** and the target finding capabilities of Juicer **[Durand et al 2016]**.

4) H-GREEN or similar methods for distal target identification.

The number, type, and strength of target identifications with these methods will be an important gauge of the strength of a regulatory interaction.

In addition, it may be advisable to consider using the squared genome wide collection of tissue specific enhancer target interactions with matrix densification machine learning methods to densify sparse measures of target genes. Of the methods above, both mQTL mapping and distal Hi-C mapping suffer from significant data sparsity, and densification methods may add value.

Pathway Mapping

The pathway mapping functionality in the current PIP relies on the human all-tissue grow and connect functions of Ingenuity Pathway Analysis **[Kramer et al 2013]**, consulted manually on the basis of the output genes. It should be possible to improve on this functionality in several ways:

Firstly, it is important that an evolved PIP's pathway mapping function run locally, run in an automated manner, and use versioned data. IPA licenses are available which offer API access to IPA commands, along with local storage of the database and/or versioned online access to historical quarterly updates. If the evolved PIP is to use IPA or another commercial or open-source pathway mapping tool, such features are essential.

More fundamentally, however, IPA, in evaluating genes only without attention to their origin or their PIP-determined relationships, is in some sense operating in a reductive manner. It can detect whether the collection of genes identified in a PIP experiment have known relationships with the phenotype under investigation or with each other, but it cannot assess the internal relationships among the set of genes and variants as identified by the PIP. In addition to the external, literature

and experiment based relationships, it should strike us as important to pathway relationships if, for example:

The same TFBS is altered and/or the same TF is present at multiple loci for a phenotype
The same gene is regulated by multiple variants for a phenotype
The same variant for a phenotype regulated multiple related genes
The same genes and variants are concordantly identified for a cluster of related phenotypes

While a full exposition of the types of relationships we should seek and evaluate would be essentially a reinvention of the entire field of pathway mapping, and thus outside the scope of this discussion, it will suffice to note that different pathway mapping options may exist, that new functionality recently added to IPA around regulatory variants may be helpful, and that particular needed functionality can be added.

Scoring

The scoring algorithm of the current PIP is minimal in nature, comprising an explicit scoring algorithm resulting a binary determination of status for each PCV for each of the components of the five box model, and a simple intersection at the end of scoring to determine a set of intermediate candidates. This method was adopted not because it is optimal but because it is simple.

In the evolved PIP, each component will generate a vector of numeric and qualitative scores of various types, which we wish to integrate into a holistic picture, not necessarily of which variants

"pass" a final binary, but of which variants from the set of PCVs are most likely to be predictive, which among them are likely to be mechanistically related, and which genes are likely to play in the mechanism.

Among possible scoring systems, systems in which "points" are awarded for various categoric and numeric elements and summed up to determine a "score" are appealing for their conceptual simplicity, but unlikely to be suited for this complex task, because of the interrelated and cooperative nature of the types of evidence which make up a regulatory variant determination. For example, a variant with no evidence of variant dependence is probably null, regardless of the potency of an enhancer element in which it is located or the target genes of that enhancer. Despite our experiments with "points" based scoring systems, they are unlikely to be fruitful.

The universe of possible explicit scoring algorithms, incorporating "points" and thresholds and sliding scales, is vast. And complex explicit algorithms have achieved great success in many applications in biology, bioinformatics, and medicine. It is possible to envision an ensemble of explicit methods capturing a great deal of complexity and yielding "good" scoring. Nevertheless, it will be useful to consider another possibility: the use of machine learning methods to address this question.

Machine Learning for PIP Scoring: Depth vs Context

Recent applications of deep learning in biomedicine have already demonstrated their superior performance compared to other machine learning approaches in a number of biomedical problems

[Ching et al 2018], including those in image analysis **[Litjent et al 2017, Iglovikov et al 2017, Rakhlin et al 2018, Shvets et al 2018]**, genomics **[Angermueller et al 2016, Mamoshina et al 2016**]; as well as drug discovery and repurposing **[Goh et al 2017, Ramsundar et al 2017]**. This great success of deep learning models in many tasks is thought to be enabled by the explosive growth of volume of raw data along with significant progress in computing, including the use of powerful graphical processing units (GPUs) that are specifically well-suited for the optimization of deep learning models.

Conventional machine learning algorithms are typically limited in their ability to process raw data **[Lecun et al 2015]**. Their performance heavily depends on the extraction of relevant representations or features that requires careful engineering and considerable domain expertise. In the past, biomedical datasets have typically been limited by sample size, and since often many more features could be measured, the performance of conventional machine learning algorithms degraded when useful information was buried in an excess of extracted features. This posed a challenge for the determination and extraction of the optimal feature set for the problem under examination. Two related and widely-used solutions are used to overcome this limitation: 1) dimensionality reduction methods that shrink the feature space to the set of most informative components **[Li R et al 2017]**; and 2) feature selection methods that identify a relatively small number of features that can accurately predict an outcome **[Vidyasagar et al 2015]**. While many of these general-purpose methods already exist, they are not necessarily optimized for pharmacogenomic biomarker discovery. This and other related pharmacological research applications require careful experimental design and choice of validation techniques. Overall, limitations of conventional machine learning methods include the need for extensive human

guidance, painstaking feature handcrafting, careful data pre-processing, and the above-mentioned dimensionality reduction to achieve top performance.

In contrast, deep learning methods model data by learning high-level representations with multi-layer computational models such as artificial neural networks (ANNs) **[Lecun et al 2015]**. While classic feed-forward ANNs might serve as drop-in replacement for other machine learning models and require input pre-processing and feature extraction; deep learning architectures, such as convolutional neural networks (CNNs), allow the algorithm to automatically learn features from raw and noisy data. Deep neural networks rely on algorithms that optimize feature engineering processes to provide the classifier with relevant information that maximizes its performance with respect to the final task. Such deep learning models can be thought of as automated "feature learning" or "feature detection," which facilitates learning of hierarchical, increasingly abstract representations of high-dimensional heterogeneous data **[Lecun et al 2015]**, also known as "representation learning." Some common deep learning methods include deep feed-forward artificial neural networks (ANNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), stacked auto-encoders (SAEs), deep belief networks (DBNs), and deep reinforcement learning techniques **[Lecun et al 2017, Leung et al 2016, Angermueller et al 2016]**. In biomedicine, these models are capable of unguided extraction of highly complex, nonlinear dependencies from raw data such as raw sequence data **[Ching et al 2018]**.

For this reason, since the data a PIP style pipeline accumulates about each SNP and gene is highly structured, it is likely that a conventional machine learning approach would be suited for this task without the need for a deep learning architecture.

Machine Learning for PIP scoring: Supervised Classification, Overfitting, and Data Availability

While most successful deep learning applications relied on large labeled data, many biological and clinical datasets until recently were limited by amount of available labeled samples compared with the big data analytics applications such as natural image processing and NLP. Over time, as the number of samples increases and the number of relevant high-quality labeled datasets expands, the wealth of pertinent pharmacogenomics data that can be used for analytics will begin to resemble a big data challenge on par with contemporary applications in other domains. The rich variety of heterogeneous data types in pharmacogenomics can improve the utilization of highly flexible deep networks that can deal with sparse, high-dimensional, multi-modal data. The real power of deep learning in a domain such as pharmacogenomics will be realized when it is combined with a prior domain knowledge **[Ching et al 2018]**, such as gene networks or pathways; a relevant example being used for the prediction of the pharmacological properties of drugs using transcriptomic data combined with pathway information **[Aliper et al 2016]**. Multimodal, multi-task, and transfer learning are often used to alleviate data limitations to some extent. Transfer learning approaches include training a deep network on a large existing dataset, and then using this pre-initialized model to learn from a smaller dataset, which typically leads to improved performance **[Ching et al 2018]**. When training data is not (fully) labeled, various semi-supervised techniques can be employed **[Ching et al 2018, Iglovikov et al 2017, Beaulieu-Jones et al 2016]**. Data quality is another important concern in deep learning applications. Although deep learning models can be trained directly on raw data, low quality datasets may require additional pre-processing and cleaning.

Publicly sharing the pre-processing code (e.g., Basset **[Kelley et al 2016]**) and cleaned data (e.g., MoleculeNet **[Wu et al 2018]**) is important to expedite further research and practical applications.

A trained machine learning model may represent some attributes of the dataset that do not accurately reflect their underlying relationships. This problem may be magnified as the size of the feature or parameter set is increased relative to the size of the input sample set. Such models exhibit poor predictive performance, as they over-represent minor variations in the training data. Overfitting is an issue of trade-off between generalization and approximation of the training data in a model. A model can underfit high-dimensional, heterogeneous dataset with complex hidden patterns if the model's representational power is low, which is often the case, for example, for linear models. Although overfitting is a common issue in machine learning, it is more likely to affect complex models, especially when there are not enough samples in the training set, learning is performed for too long, or where training examples are rare, causing the learner to follow the patterns in training data too closely. In the case of deep learning, overfitting is often a threat due to the high representational power of a model, which leads to the ability to "memorize" the whole training set very efficiently. Thus, deep learning methods require careful choice of model architecture and hyper-parameters. Although there are specific methods to prevent overfitting **[Lecun et al 2015, Angermueller et al 2017]**, in general, the trained model should exhibit robust performance on test data using relevant properties of the trained data. For more detailed description of overfitting and model selection, see **[Lever et al 2016]**.

Preventing overfitting also requires a very careful design of the model evaluation scheme, including usage of cross-validation techniques, normalization by subjects, etc. Validation metrics

may include mean absolute error or root-mean-square error (sample standard deviation of the differences between predicted and observed outcomes) for regression; accuracy; precision (also known as positive predictive value (PPV) – the fraction of retrieved instances that are relevant); recall (sensitivity - the fraction of relevant instances that are retrieved); precision-recall curve (PRC) and area under the PRC (AUPR); receiver operating characteristic (ROC) and area under the ROC curve (AUC); and mean average precision (MAP), for ranking **[Lever et al 2016, Saito et al 2015]**. Although some of these may seem intuitive, correct determination requires great care, and is often fraught with sources of error that are not easily understood, except in the context of the problem under study. For example, while the AUC plot is a common visual method for classification performance evaluation, it is not the most informative when classes are represented largely by a different number of samples in the dataset **[Saito et al 2015]**, which is a common situation in pharmacogenomics. One test of the quality of the trained machine learning model is its ability to faithfully generalize into varying test sets that constitute different manifestations of the same problem.

In the context of the PIP, the consequences of this are clear. There does not currently exist a set of adequately characterized positive controls for pharmacogenomics regulatory variants, of a type, scale, or breadth, to allow the construction of a monolithic supervised machine learning algorithm for scoring PIP variants. Indeed it is not entirely clear from where, with any plausible level of effort, such a training set would come. This complicates matters considerably.

However, subcomponents of the overall model exist wherein training sets for machine learning are available, including the portions described above where machine learning algorithms are used to

generate scores in individual modules of the PIP featureset. These include the use of hidden Markov models to generate chromatin states, the use of neural networks to predict variant effects on DNase accessibility and chromatin state, motif discovery by clustering, matrix densification for regulatory variant target discovery, etc. Artificial intelligence algorithms also have been demonstrated for variant imputation.

It may be that databases of validated tissue specific enhancer variants may function as a training set along with omics atlas data, so that the entire ERV workflow (regulatory function, variant dependence, and target genes) may be carried out with a supervised-training machine learning algorithm to predict the presence or absence of a "validated-appearing" enhancer variant.

In addition to this, the relationship between metrics of PCV status, like target gene status, association p-values, etc, and status as a plausible causative variant for the phenotype, among variants which are called as potent regulatory variants, may be addressable with Bayesian statistical approaches constructed synthetically from statistical models in population genetics. Similar approaches may be applied to network membership in the pathway mapping portion of an evolved workflow.

Thus, although it appears intractable to attempt to replace PIP scoring with a monolithic machine learning algorithm using any current tools, it will be possible to gain more insight from more data by the replacement of progressively larger portions of the intermediate scoring (the more defined portions) with machine learning algorithms, even as the overall scoring system remains an ensemble constructed explicitly.

Using the Output of PIP-Style Pipelines to Develop Genetic Tests

The output of the PIP is a set of variants, genes, and pathways with a putative causal role in a phenotype of interest. Although such results may be useful for purposes of mechanistic research in the biology of a phenotype, or the search for druggable targets and repurposing opportunities. However, in the context of pharmacogenomics the object of most PIP analyses will be to develop genetic tests (and more broadly, clinical predictive models) for eventual clinical deployment. While an authoritative discussion of this subject is beyond the scope of this thesis, it will serve to discuss this topic briefly in light of the above.

Typically, first generation pharmacogenomics tests were designed by a simple linear combination of variant effect sizes, or with a linear regression model. Later first-generation tests such as the Genesight **[Health Quality Ontario 2017]** psychotropic panel were constructed manually as an explicit combinatorial decision tree algorithm on the basis of effect sizes, paired variant effects, and manual expertise and adjustment. Then, these explicitly constructed tests could be validated against genotyped cohorts with response data before being tested in randomized controlled trials.

The PIP and PIP-style pipelines can be used to design tests in this mode. But doing so would deprive the designer of much of the benefit of the PIP and of modern data sources. With significant pleiotropy between pharmacogenomic loci, the reduction of features to a tractable number, and the availability of genotyped cohorts and machine learning methods, it is anticipated that the ideal

method for designing a pharmacogenomics test on the basis of PIP results would incorporate these advances.

In this case, a genotyped cohort with genomic information on all the output loci of a PIP experiment, along with clinical variables deemed important in test design, and outcome information, would be gathered, preferably retrospectively from a GWAS or biobank. Then, a machine learning predictor would be trained to predict the outcome variable with this cohort, with cross validation. Since this is a relatively low dimensional space with supervisory information, and because the intention would be to clinically deploy the finished predictor, a relatively simple machine learning method like an SVM or Random Forest would probably suffice. Then, features could be reduced using an iterative marginal information analysis approach to arrive at a set of informative loci and clinical features to be used. A predictive model trained on these features would function as a finished genetic test and could be validated and deployed.

This approach is described schematically in **Figure 5-3**.

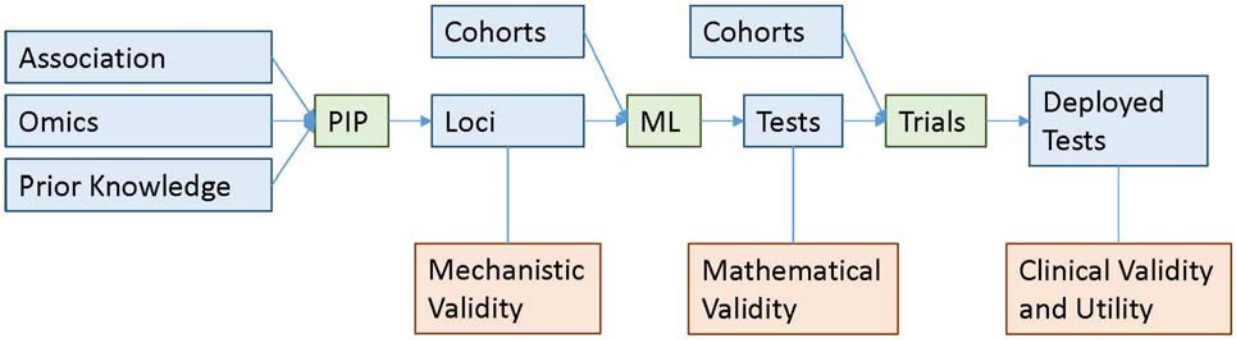


Figure 5-3: Using PIP-Style Pipelines to Design Genetic Tests

Schematic of an overall method for using PIP-style pipelines to design genetic-clinical tests. After loci are generated by a PIP experiment, they are used along with tractable and predictive clinical features in a cohort dataset to generate a predictive machine learning model, which is tuned with ablation analysis to find features with high marginal information. The minimal set of such features is used to generate a final predictive model which functions as a test, and can be validated in prospective and retrospective clinical trials.

Although historic regulatory approaches have demanded prospective clinical trials before clinical deployment of genetic tests and genetic prediction modules for clinical decision support, scientifically there is little conceptual distinction between a retrospective trial on a cohort which was not used in the construction of the test, and a prospective trial. As the scientific consensus around these issues congeals and makes its way into the culture of regulatory thinking, it is likely that the greater speed and lower expense of this approach, along with the clinical utility of making genetic testing more widespread, will carry the day in favor of this kind of development.

But in any event, even if prospective clinical trials were required in order to validate such predictive models, the fact that they were initiated on the basis of PIP-discovered variants with mechanistic validity should advantage such classifiers in the regulatory environment, compared with agnostic machine learning approaches based on whole genomes.

Much has been made of regulatory barriers and cultural hesitance to use genetic information in some medical specialties as explanatory elements for the slow progress of pharmacogenomics testing deployment in many clinical specialties. Historically, however, the single biggest factor preventing pharmacogenomics testing from reaching the clinic in any given case has not been regulatory or cultural barriers but the clinical utility of the underlying prediction. In instances in oncology and neuropsychiatry where such tests have added value, they have typically met with at least enough regulatory permission and cultural tolerance to be applied. It may be anticipated that if, in the fullness of time, tests for new phenotypes do add such utility, neither governments nor conservative clinicians will stand in their way indefinitely. And as the number of domains where such tests add value grows, they may become the object of much enthusiasm.

Using PIP-style Pipelines to Construct a Phenome-wide Pharmacoepigenomic Atlas

A PIP-style pipeline analysis, intended to discover variants with a causative role in a particular phenotype, takes place using a large amount of omics data which comes from the same database for each experiment, and is selected on the basis of preexisting knowledge about the phenotype under investigation. In the current version of the PIP, this principally comprises the MVF and MTF, containing information about the key mechanistic genes and associated variants, and the relevant tissues. In the case of the evolved featureset discussed in this chapter, other key information will come to the fore, including the most relevant Hi-C datasets, the degree of relatedness of clustered phenotypes, and the tissues for each phenotype. But regardless, a five box model pipeline requires a certain set of specific information about a phenotype system in order to run.

Many of the underlying elements of such a pipeline have begun to parallelize across all domains. For example, GWAS have now been conducted on thousands of phenotypes **[MacArthur et al 2017]**, and genotypes biobanks have enabled PheWAS **[Denny et al 2010]** methods to conduct parallel GWAS on thousands of phenotypes with one large cohort. Epigenome atlases now address an increasingly large cross section of the human body, with increasing granularity. Biomedical ontology databases now attempt to contain the knowledge on relatedness of phenotype categories and the contributions of tissues to phenotypes within a structured, computable vocabulary. And research in the design of EHR parsers is now proceeding **[Denny et al 2013, Beaulieu-Jones et al**

2016], gesturing to a future world wherein individual health records can be grouped into phenotypic classes on a reliable and automated basis for an ever larger span of phenotypes.

The maturity of these trends will culminate in a world wherein, instead of performing a targeted PIP-style analysis on a particular phenotype of interest and using it to design a test, the converged datasets and methods described here will be used to perform parallel analyses and parallel test design for thousands of phenotypes. The predictors for these phenotypes could, for genotyped patients, then be contributed directly to an EHR system and used for clinical decision support. Furthermore, the genotypes, clinical records, and outcomes of patients within an EHR system could be used to refine the predictors for such a system on an ongoing basis.

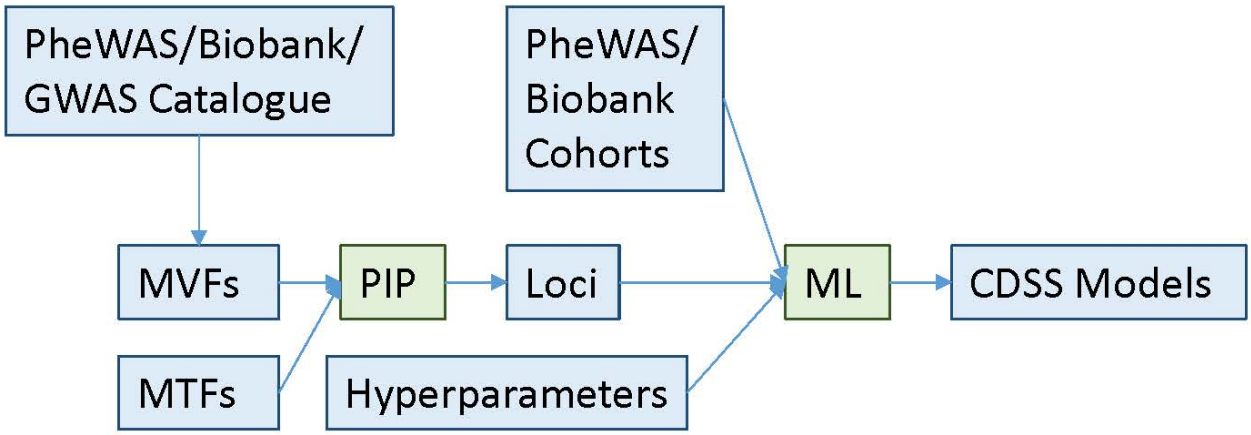


Figure 5-4: Conceptual Process for Constructing a Pharmacophenomic Atlas

Schematic representation of a conceptual process for parallelizing a PIP-style pipeline across thousands of phenotypes to create a genome-wide, phenome-wide pharmacophenome atlas. With either automation of parallelized manual work, MVFs and MTFs are created for thousands of phenotypes on the basis of PheWAS, biobanks, and/or the GWAS catalog, and thousands of PIP experiments run. Then, automated test generation as described in Figure 5-3 is undertaken in parallel for all the phenotypes based on separate cohorts from a biobank and/or PheWAS. Finally, this set of thousands of predictive models for a comprehensive collection of pharmacological phenotypes may be used in every CDSS and research context wherein it may add value.

The concluding section of this thesis will describe a future vision for the means by which this process could take place, which is also pictured schematically in **Figure 5-4**.

The first step in such an analysis would be the construction of a phenotype set for investigation. Each PheWAS analysis already constructs such a set, typically manually.

The construction of tissue files on an automated basis requires a degree of natural language processing. Ontologic mapping of phenotypes to relevant diseases, along with crawling of structured literature databases like the EBI GWAS catalog **[Macarthur et al 2017]**, and natural language processing, could be used to create variant input files on an automated basis, by locating the relevant phenotypes in natural language and extracting the locus and population information. In addition to this, the output of a PheWAS on the same set of phenotypes could be used to construct the variant files.

Similarly, ontologies mapping drugs and diseases to their sites of action (e.g. SNOMED **[Millar 2016]**), cell lines to their cytologic properties and tissues of origin (e.g. Cell Line Ontology **[Sarntivijai et al 2014]**), organs and tissues in a hierarchical tree (e.g. NeuroFMA **[Turner et al 2010]**, HOMER **[Zhang et al 2011]**, BRENDA **[Gremse et al 2011]**), and synonymous tissue names to each other, may allow the creation of tissue files on an automated basis. This would create a description of all the relevant tissues for each phenotype.

With this done, it would be possible to run a PIP-style pipeline in parallel on many drug-disease systems to create a genome wide, phenome wide atlas of significant pharmacogenomic loci and

gene networks. In particular, such experiments could be conducted in parallel on the basis of association data from EHR records for large populations **[Denny et al 2013]**. This atlas could include all the drug disease systems, disease risks, and other pharmacological phenotypes for which the underlying genetic associations and tissue specific omics are available.

Theoretically, if such automated methods as are described above proved impractical, and if sufficient resources were available, the manual curation of a library of tissue and variant files could be undertaken for a large library of phenotypes. The creation of input files for the current version of the PIP has already been reduced to a matter of days for two investigators, and with a suitable graphical interface for curation, could be reduced further, perhaps to the point where manual curation by a small dedicated staff with access to a variety of medical specialists became a tractable alternative. Certainly this has been the predominant approach in PheWAS design.

With a separate linked biobank not used in the original analysis, such a "PIP-WAS" atlas could be used to design predictive models for every phenotype under investigation. This would require the phenotypic extraction, key clinical variables, and clinical variable extraction for each phenotype from the second biobank to be automated as well. In addition, it would require the test generation and marginal information ablation analysis to be automated, a requirement which would require a lot of hyperparameter tuning. Nevertheless I am confident it is possible with current methods.

This comprehensive pharmacogenomics atlas would represent the industrialization of pharmacogenomics. It would enable, among other things, the scalable parallel design of tests for many diseases and drugs on the basis of genotype-linked EHR data, and the development of

microarrays or sequencing panels containing the genetic information for many tests, or all, in one package. Such information could then be available in the EHR for a variety of purposes, and could even be present on, e.g., a military "dog tag" in the form of a QR code or other linked identifier for use in emergency and traumatic care settings.

Nor, indeed, would the construction of such an atlas need to be a discrete versioned event. If genotyped EHR data were added to the system, predictive models could refresh on an ongoing basis with new information, and new phenotypes could be added to the catalogue as their relevance was established.

Coda: Pharmacogenomics in 2030

What would medicine look like, in a world wherein such an atlas had been constructed, and wherein it had been incorporated into the EHR systems in use in the clinic, and in which patient genotypes were routinely available? Every time a patient required general health guidance, clinicians would have access to predictive information suggesting the diseases and syndromes which might present the greatest risk. For every prescription decision, the pharmacogenomics of the various drugs available such such an indication could be displayed, along with the particular adverse events most concerning for each one, for the individual patient. And for public health concerns with rare indications, the application of an optimized classifier to the covered population of a health system could prospectively identify a cohort who would benefit from prophylactic guidance or treatment.

Only twelve years ago, before the advent of GWAS, genetic tests were designed manually for each phenotype on the basis of painstaking, locus-specific mechanistic work, almost exclusively using coding variants, and then loci were assayed individually before being reported, often with a delay of weeks. While the clinical side of this picture has seen only muted change in the time since, rapid and fundamental advances on the research side, including the PIP and H-GREEN, have made it possible to see forward to a future, twelve years from now, when the parallel design of genetic tests for thousands of phenotypes, incorporating tissue-specific regulatory variants, machine learning, and large cohort datasets, has made the display of genetic predictive information on relevant phenotypes a routine and automatic part of life in the health care clinic.

References

1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 491(7422), 56-65 (2012).

The 1000 Genomes Project Consortium. A global reference for human genetic vatiation. Nature, October 2015, 526:68-74

Aberg K, Adkins DE, Liu Y et al. Genome-wide association study of antipsychotic-induced QTc interval prolongation. *Pharmacogenomics.* 12:165–72. (2012)

Adkins DE, Åberg K, McClay JL et al. Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs. *Molecular psychiatry*, 16(3), 321-332. (2011)

Adli M, Hollinde DL, Stamm T et al. Response to lithium augmentation in depression is associated with the glycogen synthase kinase 3-beta− 50T/C single nucleotide polymorphism. Biol. Psychiatry. 62(11), 1295-1302 (2007).

Aguilar-Arnal L, Sassone-Corsi P. Chromatin landscape and circadian dynamics: Spatial and temporal organization of clock transcription. *PNAS.* DOI10.1073/pnas.1411264111 (2014).

Ahmed AI, Gajavelli S, Spurlock MS, Chieng LO, Bullock MR. Stem cells for therapy in TBI. Journal of the Royal Army Medical Corps. 2016 Apr 1;162(2):98-102.

Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33(8), 831-838 (2015).

Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. Mol Pharm 13(7), 2524-2530 (2016).

Allyn-Feuer A, Ade A, Luzum JA, Higgina GA, Athey BD. The pharmacoepigenomics informatics pipeline defines a pathway of novel and known warfarin pharmacogenomics variants. Pharmacogenomics, April 2018.

Alpert H, Koch C et al. Obstacle numbers of graphs. Discrete and Computational Geometry 44 (1), 223–244, 2010.

Amin N, Byrne E, Johnson J et al.: Genome-wide association analysis of coffee drinking suggests association with CYP1A1/CYP1A2 and NRCAM. *Molecular Psychiatry* 17:1116–1129. (2012)

Amos CI. Successful design and conduct of genome-wide association studies. Human Molecular Genetics 2007. https://doi.org/10.1093/hmg/ddm161

Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol 12(7), (2016).

Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol 18(1), 67 (2017).

Arthur R. Medicine Is Getting More Precise…For White People. FiveThirtyEight, Science & Health, August 2, 2017.

Athanasiadis G, Buil A et al. A genome-wide association study of the Protein C anticoagulant pathway. PLoS One, December 2011, 6(12):e29168.

Athey BD, Smith MF et al. The diameters of frozen-hydrated chromatin fibers increase with DNA linker length: evidence in support of variable diameter models for chromatin. J Cell Biol. 1990 Sep;111(3):795-806.

Avsec Z, Barekatain M, Cheng J, Gagneur J. Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks. bioRxiv doi:10.1101/165183, (2017).

Bader LA, Elewa H. The Impact of Genetic and non-Genetic Factors on Warfarin Dose Prediction in MENA Region: A Systematic Review. PLoS One, December 2016, doi: 10.1371/journal.pone.0168732

Badyaev AV. Epigenetic resolution of the 'curse of complexity' in adaptive evolution of complex traits. J. Physiol. 592 (11), 2251-2260 (2104).

Baik I, Cho NH, Kim SH, Han BG, Shin C. Genome-wide association studies identify genetic loci related to alcohol consumption in Korean men. *Am J Clin Nutr.* 93:809–816. (2011)

Bailey DG, Dresser GK. Interactions between grapefruit juice and cardiovascular drugs. Am J Cardiovasc Drugs. 2004;4(5):281-97.

Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Research* 21:381-395. (2011)

Banovich NE, Lan X et al. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genetics, September 2014. 10(9):e1004663.

Bare LA, Arellano AR et al. Genetic variants of F11, statin use and venous thrombosis. Journal of Thrombosis and Haematostasis, June 2011, 9(6):1249-1273.

Bargal S, Kight F et al. Implications of polymorphisms in BCKDK and GATA-4 genetic regions on stable warfarin dose in African Americans. Pharmacotherapy, December 2016, 36(12):e267.

Barnes GD, Lucas E et al. National trends in ambulatory oral anticoagulant use. Am J Med, December 2015, 128(12):1300-5.

Baumgartner A, Weier JF, Weier HUG. Chromosome-specific DNA Repeat Probes. J. Histochem. Cytochem. 2006

Baxter JS, Leavy OC et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. Nature Communicationsvolume 9, Article number: 1028 (2018)

Beaulieu-Jones BK, Greene CS, Pooled Resource Open-Access ALSCTC. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform 64, 168-178 (2016).

Beagrie RA, Scialdione A et al. Complex multi-enhancer contacts captured by genome architecture mapping. Nature volume 543, pages 519–524 (23 March 2017)

Benedetti F, Serretti A, Pontiggia A et al. Long-term response to lithium salts in bipolar illness is influenced by the glycogen synthase kinase 3-β− 50 T/C SNP. Neurosci. Letters. 376 (1) 51-55 (2005).

Benedetti F, Bollettini I, Barberi I et al. Lithium and GSK3-beta promoter gene variants influence white matter microstructure in bipolar disorder. Neuropsychopharmacol. 38, 313-27 (2013).

Benjamini Y, Hochberg B. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc*. 57:1 289-300. (1995)

Benner C, Heinz S et al. HOMER (v4.10, 5-16-2018): Software for motif discovery and next generation sequencing analysis. 2018. http://homer.ucsd.edu/homer/

Bel JT, Spector TD. A twin approach to unraveling epigenetics. Trends Genet. 27(3), 116-125. (2011).

Ben-Elazar S, Chor B, Yakhini Z. Extending partial haplotypes to full genome haplotypes using chromosome conformation capture data. Bioinformatics, 32, 2016, i559–i566 doi: 10.1093/bioinformatics/btw453

Berezin V, Walmod PS. Cell Adhesion Molecules: Implications in Neurological Diseases. Vol. 8. *Springer Science & Business Media*, (2013)

Berto S, Perdomo-Sabogal A, Gerighausen D, Qin J, Nowick K. A consensus network of gene regulatory factors in the human frontal lobe. Frontiers in Genetics. 2016;7.

Betzig E, Patterson GH et al. Imaging intracellular fluorescent proteins at nanometer resolution. Science. 2006 Sep 15;313(5793):1642-5.

Bharadwaj R, Peter CJ, Jiang Y, Roussos P, Vogel-Ciernia A, Shen EY, Mitchell AC, Mao W, Whittle C, Dincer A, Jakovcevski M. Conserved higher-order chromatin regulates NMDA receptor gene expression and cognition. Neuron. 2014 Dec 3;84(5):997-1008.

Black JL, O'Kane DJ, Mrazek DA. The impact of CYP allelic variation on antidepressant metabolism: a review. Expert Opin. Drug Metab. Toxicol. February 2007. 3(1):21-31

Blanchard G, Roquain E. Adaptive False Discovery Rate Control under Independence and Dependence. Journal of Machine Learning Research, 2009.

Blum K, Smolen A, Downs BW et al. Genetic Addiction Risk Score (GARS): Testing For Polygenetic Predisposition and Risk to Reward Deficiency Syndrome (RDS). *INTECH Open Access Publisher*, (2011)

Boney B., Cavalli G., Organization and function of the 3D genome, Nature Rev. Genet. 17(11), (2016), 661-678.

Booth AD, Colin AJT. On the efficiency of a new method of dictionary construction. Information and Control Volume 3, Issue 4, December 1960, Pages 327-334. https://doi.org/10.1016/S0019-9958(60)90901-3

Borel C, Ferreira, PG, Santoni F et al. Biased allelic expression in human primary fibroblast single cells. *Amer. J. Human Genet.* 96, 1–11 (2015).

Boyle AP, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790-1797 (2012).

Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell, 2017, DOI: https://doi.org/10.1016/j.cell.2017.05.038

Brandler WM. Common variants in left/right asymmetry genes and pathways are associated with relative hand skill. PLoS Genetics. 9(9):e1003751 (2013);

Brown KM, Tracy DK. Lithium: the pharmacodynamic actions of the amazing ion. Ther. Adv. Psychopharmacol. 3(3), 163-176 doi: 10.1177/2045125312471963. (2013).

Brownstein MJ. A brief history of opiates, opioid peptides, and opioid receptors. *Proc Natl Acad Sci* USA 90 (12): 5391–5393 doi:10.1073/pnas.90.12.5391 (1993)

Buenrostro J, Wu B, Chang H, Greenleaf W. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015; 109: 21.29.1–21.29.9. doi: 10.1002/0471142727.mb2129s109

Buil A, Tregouet DA et al. C4BPB/C4BPA is a new susceptibility locus for venous thrombosos with unknown protein S-independent mechanism: results from genome-wide association and gene expression analyses followed by case-control studies. Blood, June 2010, 115(23):4644-50.

Bunney B.G., Li J. Z. et al, Circadian dysregulation of clock genes: clues to rapid treatments in major depressive disorder, Mol. Psychiatry. 20(1), (2015),48-55.

Bush WS, Oetjens MT et al. Unravelling the human genome-phenome relationship using phenome-wide association studies, Nature Rev. Genet. 17(3), (2016), 129-145.

Byrne EM, Gehrman PR, Medland SE. A genome-wide association study of sleep habits and insomnia. Amer. J. Med. Genet. 162B (5), 439-451 (2013).

Califano A, Butte AJ, Friend S et al. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nature Genet.* 44 (8), 841-847 (2012).

Campos-de-Sousa S, Guindalini C, Tondo L, Munro J, Osborne S, Floris G, Pedrazzoli M, Tufik S, Breen G, Collier D. Nuclear receptor rev-erb-α circadian gene variants and lithium carbonate prophylaxis in bipolar affective disorder. Journal of biological rhythms. 2010 Apr 1;25(2):132-7.

Can A, Schulze TG, Gould TD. Molecular actions and clinical pharmacogenetics of lithium therapy. Pharmacol Biochem Behav. 123, 3-16 doi: 10.1016/j.pbb.2014.02.004 (2014).

Cao Z, Zhang S. gkm-DNN: efficient prediction using gapped k-mer features and deep neural networks. bioRxiv doi:10.1101/170761, (2017).

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. Bioinformatics, 25(2), 288-289 (2009).

Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank. October 2015, 16:286.

Carroll A, Thangaraj N. Evaluating DeepVariant: A New Deep Learning Variant Caller from the Google Brain Team. (2017). https://blog.dnanexus.com/2017-12-05-evaluating-deepvariant-googles-machine-learning-variant-caller/

Cederbaum AI. Alcohol Metabolism. *Clin Liver Dis.* 16:667–685. (2012)

Cerami EG, Gross BE, Demir E et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* 39(suppl 1), D685-D690 (2011).

Cha PC, Mushiroda T et al. Genome-wide association study identifies genetic determinants of warfarin responsiveness for Japanese. Hum Mol Genet. December 2010. 19(23):4735-44

Chakraborty A, Ay F. Identification of copy number variations and translocations in cancer cells from Hi-C data. Bioinformatics. 2017 Oct 18. doi: 10.1093/bioinformatics/btx664.

Chang CC, Chow CC et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. arXiv, October 2014, 1410.4803v1

Chang KC, Sasano T, Wang YC, Huang SKS: Nitric oxide synthase 1 adaptor protein, an emerging new genetic marker for QT prolongation and sudden cardiac death. *Acta Cardiol Sin* 29:217–225. (2013)

Chang M, Soderberg MM et al. CYP2C19*17 affects R-warfarin plasma clearance and warfarin INR/dose ratio in patients on stable warfarin maintenance therapy. Eur J Clin Pharmacol. 2015 Apr;71(4):433-9.

Chen B, Gilbert LA et al. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. Cell. 2013 Dec 19; 155(7): 1479–1491. doi: 10.1016/j.cell.2013.12.001

Chen B, Huang B. Imaging genomic elements in living cells using CRISPR/Cas9. Methods Enzymol. 2014;546:337-54. doi: 10.1016/B978-0-12-801185-0.00016-7.

Chen CH, Lee CS, Lee MT et al. Variant GADL1 and response to lithium therapy in bipolar I disorder. New Engl. J. Med. 370(2), 119-128 (2014).

Chen DT, Jiang X, Akula N et al. Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. Mol. Psychiatry. 18(2):195-205 (2013).

Chen H, Chen J et al. Functional Organization of the human 4D Nucleome. PNAS 2015. 112(26):8002-8007

Chen H, Comment N et al. Chromosome conformation of human fibroblasts grown in 3-dimensional spheroids. Nucleus. 2015;6(1):55-65. doi: 10.1080/19491034.2014.1003745.

Chen H, Seaman L et al. Chromosome conformation and gene expression patterns differ profoundly in human fibroblasts grown in spheroids versus monolayers. Nucleus, July 2017. 8(4):383-391

Chen J, Shao L et al. A Pharmacogenetics-based Warfarin Maintenance Dosing Algorithm from Northern Chinese Patients. PLoS One, August 2014, doi:10.1371/journal.pone.0105250

Chena H, Chena J, Muir LA et al. Functional organization of the human 4D nucleome. . *PNAS.* 112(26), 8002-8007 (2015)

Cheng MC, Lu CL, Luu SU et al. Genetic and functional analysis of the DLG4 gene encoding the post-synaptic density protein 95 in schizophrenia. PloS one, 5(12), e15107 (2010).

Chiesa A, Crisafulli C, Porcelli S et al. Influence of GRIA1, GRIA2 and GRIA4 polymorphisms on diagnosis and response to treatment in patients with major depressive disorder. European Arch. Psychiatry Clin. Neurosci. 262, 305-311 (2012).

Ching T, Himmelstein DS, Beaulieu-Jones BK et al. Opportunities And Obstacles For Deep Learning In Biology And Medicine. bioRxiv doi:10.1101/142760, (2018).

Chiu CT, Chuang DM. Molecular actions and therapeutic potential of lithium in preclinical and clinical studies of CNS disorders. Pharmacol Ther. 128(2), 281-304 doi: 10.1016/j.pharmthera.2010.07.006 (2010).

Chiu LD, Su L et al. Use of a white light supercontinuum laser for confocal interference-reflection microscopy. J Microsc. 2012 May; 246(2): 153–159. doi: 10.1111/j.1365-2818.2012.03603.x

Chrousos G.P. The hypothalamic–pituitary–adrenal axis and immune-mediated inflammation. NEJM. 332(20), (1995), 1351-1363.

Chu T, Zhou H, Wang T, Lu L, Li F, Liu B, Kong X, Feng S. In vitro characteristics of valproic acid and all-trans-retinoic acid and their combined use in promoting neuronal differentiation while suppressing astrocytic differentiation in neural stem cells. Brain Research. 2015 Jan 30;1596:31-47.

Clarke M, Razmjou S et al. Ketamine modulates hippocampal neurogenesis and pro-inflammatory cytokines but not stressor induced neurochemical changes. Neuropharmacology. 2017 Jan;112(Pt A):210-220. doi: 10.1016/j.neuropharm.2016.04.021.

Claussnitzer M, Dankel SN, Klocke B et al. Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. Cell. 156(1), 343-358 (2014).

Cloney R. Gene expression: Dynamic enhancer-promoter interactions for transcriptional bursting. Nature Rev. Genet. 17 (437), (2016), doi:10.1038/nrg.2016.81.

Cock PJA, Fields CJ et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010 Apr; 38(6): 1767–1771. doi: 10.1093/nar/gkp1137

Coetzee SG, Rhie SK, Berman BP, Coetzee GA, Noushmehr H. FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res.* 40(18):e139 (2012).

Cohen J. A power primer. *Psychol. Bulletin.* 112, 155-159 (1992)
Collins FS, McKusick MD. Implications of the Human Genome Project for Medical Science. JAMA, 2001. 285(5):540-544

Cohen NM, Olivares-Chauvet P et al. SHAMAN: bin-free randomization, normalization and screening of Hi-C matrices. BioRxiv 2017. http://dx.doi.org/10.1101/187203

Condit C, Templeton A et al. Attitudinal barriers to delivery of race-targeted pharmacogenomics among informed lay persons. Genetics in Medicine, 2003, 5:385-392.

Cooper GM, Johnson JA et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood, August 2008, 112(4):1022-7

Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet*. 12, 628–640 (2011).

Cooper RS, Tayo B, Zhu X. Genome-wide association studies: implications for multiethnic samples. Hum Mol Genet, October 2008, 17(R2):R151-R155.

Cornelis MC, Monda KL, Yu K et al. Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption. *PLoS genetics*, 7(4), e1002033. (2011)

Cortijo S, Wardenaar R, Colomé-Tatché M et al. Mapping the epigenetic basis of complex traits. Science. 343 (6175): 1145-1148 (2014).

Crawford RR, Higdon AN, Casey DB, Good DE, Mungrue IN. Multiple lithium-dependent Brugada syndrome unmasking events in a bipolar patient. Clinical Case Reports. 3, 14–18 (2015).

Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Rev. Genet.* 2(4), 292-301(2001).

Cremer T, Cremer M. Chromosome Territories. *Cold Spring Harbor Perspect. Biol.* 2(3):a003889. (2010)

Cremer T, Cremer M et al. The 4D Nucleome: Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments. FEBS Letters, October 2015. 589(20A):2931-43

Crews D, Gillette R, Scarpino SV et al. Epigenetic transgenerational inheritance of altered stress responses. *PNAS.* 109 (23), 9143–9148 (2012).

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences. 2010 Dec 14;107(50):21931-6.

Crisafulli C, Chiesa A, De Ronchi D et al. Influence of GRIA1, GRIA2 and GRIA4 polymorphisms on diagnosis and response to antipsychotic treatment in patients with schizophrenia. Neurosci Lett. 6, 170-174 doi: 10.1016/j.neulet.2011.10.074 (2012).

Crocq M. Historical and cultural aspects of man's relationship with addictive drugs. *Dialogues Clin Neurosci*. 9(4) pp. 355-361, PMC3202501 (2007)

Croft D, Mundo AF, Haw R et al. The Reactome pathway knowledgebase. Nucleic Acids Res. (2013).

Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet.* 381(9875), 1371-1379 (2013).

Cruceanu C, Ambalavanan A, Spiegelman D et al. Family-based exome-sequencing approach identifies rare susceptibility variants for lithium-responsive bipolar disorder. Genome. 56, 634-640 (2013).

Dai C, Li W et al. Mining 3D genome structure populations identifies major factors governing the stability of regulatory communities. Nature Commun. 7 (11549), (2016), doi:10.1038/ncomms11549.

Dallaspezia S, Poletti S, Lorenzi C, Pirovano A, Colombo C, Benedetti F. Influence of an interaction between lithium salts and a functional polymorphism in SLC1A2 on the history of illness in bipolar disorder. *Mol Diagn Ther.* 16:303–309. (2012)

Daneshjou R, Cavallari LH et al. Population-specific single-nucleotide polymorphisms confers increased risk of venous thromboembolism in African Americans. Mol Genet Genomic Med, June 2016, 4(5):513-20.

Danielson PB. The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. Curr Drug Metab, December 2002, 3(6):561-97.

Darbar D, Yang T, Churchwell K, Wilde AA, Roden DM. Unmasking of brugada syndrome by lithium. Circulation. 112, 1527–1531 (2005).

Dash PK, Orsi SA, Zhang M, Grill RJ, Pati S, Zhao J, Moore AN. Valproate administered after traumatic brain injury provides neuroprotection and improves cognitive function in rats. PloS one. 2010 Jun 30;5(6):e11383.

De Wit E, Vos ES et al. CTCF Binding Polarity Determines Chromatin Looping. Mol Cell. 2015 Nov 19;60(4):676-84. doi: 10.1016/j.molcel.2015.09.023.

Degner JF, Pai AA et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012 Feb 5;482(7385):390-4. doi: 10.1038/nature10808.

Dekker J, Rippe K et al. Capturing chromosome conformation. Science 2002. Science. 2002 Feb 15;295(5558):1306-11.

Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Rev. Genet. 14, 390-403 (2013).

Dekker SE, Bambakidis T, Sillesen M, Liu B, Johnson CN, Jin G, Li Y, Alam HB. Effect of pharmacologic resuscitation on the brain gene expression profiles in a swine model of traumatic brain injury and hemorrhage. Journal of Trauma and Acute Care Surgery. 2014 Dec 1;77(6):906-12.

Delaneau O, Ongen H et al. A complete tool set for molecular QTL discovery and analysis. Nature Communications volume 8, Article number: 15452 (2017)

Deng Q, Ramskold D, Reinius B. et al. Single-cell RNA-seq reveals dynamic, random mono-allelic gene expression in mammalian cells. *Science*. 343, 193–196 (2014).

Deng W, Lee J, Wang H et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell. 149, 1233–1244 (2012).

Dennis G Jr, Sherman BT, Hosack DA et al. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* 4(5) P3 (2003).

Denny JC, Ritchie MD et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations, Bioinformatics. 26(9), (2010), 1205-1210.

Denny JC, Bastarache L et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association data. Nature Biotechnology, December 2013, 31(12):1102-10.

Desch KC, Ozel AB et al. Linkage analysis identifies a locus for plasma von willebrand factor undetected by genome-wide association. PNAS, January 2013, 110(2):588-93.

Diekhof EK, Falkai P, Gruber O. Functional neuroimaging of reward processing and decision-making: a review of aberrant motivational and affective processing in addiction and mood disorders. Brain research reviews. 2008 Nov 30;59(1):164-84.

Dietz DM, LaPlant Q, Watts EL et al. Paternal transmission of stress-induced pathologies. *Biol. Psychiatry*. 70, 408 – 414 (2011).

Dima D, Jogia J, Collier D, Vassos E, Burdick KE, Frangou S. Independent modulation of engagement and connectivity of the facial network during affect processing by CACNA1C and ANK3 risk genes for bipolar disorder. JAMA Psychiatry. 70(12), 1303-1311. (2013).

Dixon JR, Selvaraj S, Yue F et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380 (2012).

Djurovic, S, Gustafssonc O, Mattingsdala M et al: A genome-wide association study of bipolar disorder in Norwegian individuals, followed by replication in Icelandic sample. *Journal Affective Disorders.* 126: 312-316. (2010)

Domschke K, Reif A: Behavioral genetics of affective and anxiety disorders. In *Behavioral Neurogenetics*. Edited by Cryan JF, Reif A. Springer Berlin Heidelberg; 2012:463–502. [Greyer MA, Ellenbroek BA, Marsden CA, Barnes RE (Series Editors): *Current Topics in Behavioral Neurosciences,* vol 12.]

Dostie J, Richmond TA et al. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006. 16: 1299-1309.

DrugBank database, version, 2 (2015).

Du J, Creson TK, Wu LJ et al. The role of hippocampal GluR1 and GluR2 receptors in manic-like behavior. J Neurosci. 28, 68-79 (2008).

Duell EJ, Sala N, Travier N et al: Genetic variation in alcohol dehydrogenase (ADH1A, ADH1B, ADH1C, ADH7) and aldehyde dehydrogenase (ALDH2), alcohol consumption and gastric cancer risk in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort. *Carcinogenesis.* 33:361–367. (2012)

Durand NC, Shamim MS et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016 Jul; 3(1): 95–98. doi: 10.1016/j.cels.2016.07.002

Eastwood SL, McDonald B, Burnet PW, Beckwith JP, Kerwin RW, Harrison PJ. Decreased expression of mRNAs encoding non-NMDA glutamate receptors GluR1 and GluR2 in medial temporal lobe neurons in schizophrenia. Mol. Brain Res. 29 (2) 211-23. (1995).

Edelstein LC, Luna EJ et al. Human genome-wide association and mouse knockout approaches identify platelet supervillin as an inhibitor of thrombus formation under shear stress. Circulation, June 2012, 125(22):2762-71.

Edwards SL, Beesley J, French JD, Dunning M: Beyond GWAS: Illuminating the dark road from association to function. *Am J Hum Genet* 93:779–797. (2013)

The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, June 2007, 447:799-816

The ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. *Nature.* 489:57–74 (2012).

Enoch MA, Rosser AA, Zhou Z, Mash DC, Yuan Q, Goldman D. Expression of glutamatergic genes in healthy humans across 16 brain regions; altered expression in the hippocampus after chronic exposure to alcohol or cocaine. *Genes, Brain and Behavior*. 13(8), 758-76 (2014)

Eriksson N, Wallentin L et al. Genetic determinants of warfarin maintenance dose and time in therapeutic range: a RE-LY genomics substudy. Pharmacogenomics, August 2016, 17(13):1425-1439.

Ernst E, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotech.* doi:10.1038/nbt.3157 (2015)

Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnol.* 28:817–825 (2010).

Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods.* 9,215–216 (2012).

Eser U, Churchman LS. FIDDLE: An integrative deep learning framework for functional genomic data inference. bioRxiv doi:10.1101/081380, (2016).

Eskiw CH, Rapp A et al. RNA polymerase II activity is located on the surface of protein-rich transcription factories. J Cell Sci. 2008 Jun 15;121(Pt 12):1999-2007. doi: 10.1242/jcs.027250.

Evans DA, Clarke CA. Pharmacogenetics. Br Med Bull. 1961 Sep;17:234-40.

Fabbri C, Crisafulli C et al. Progress and prospects in pharmacogenetics of antidepressant drugs. Expert Opin. Drug Metab. Toxicol. October 2016, 12(10):1157-68

Fadista J, Manning AK et al. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet. 2016 Aug; 24(8): 1202–1205. doi: 10.1038/ejhg.2015.269

Fagny M, Paulson JN et al. A network-based approach to eQTL interpretation and SNP functional characterization. BioRxiv, November 2016, doi:10.1101/086587

Farh K K-H et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 518, 337-343 (2015). doi:10.1038/nature13835

Feldman MW, Ramachandran S. Missing compared to what? Revisiting heritability, genes and culture. Philos Trans R Soc Lond B Biol Sci. 2018 Apr 5; 373(1743): 20170064.

Ferreira MA, O'Donovan MC, Meng YA et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet. 40(9), 1056-1058 (2008).

Flanagan JM, Popendikyte V et al. Intra- and Interindividual Epigenetic Variation in Human Germ Cells. Am. J. Hum. Genet. 2006. https://doi.org/10.1086/504729

Flicek P, Amode MR, Barrell Det al. Ensembl 2014. *Nucleic Acids Res.* gkt1196 (2013).

Flister MJ, Tsaih SW, O'Meara CC et al. Identifying multiple causative genes at a single GWAS locus. Genome Res. 23, 1996–2002 (2013).

Flockhart DA, O'Kane D et al. Pharmacogenetic testing of CYP2C9 and VKORC1 alleles for warfarin. Genetics in Medicine, 2008, 10:139-150

Florez JC. Pharmacogenomics in type 2 diabetes: precision medicine or discovery tool? Diabetologia, May 2017, 60(5):800-807

Fraga MF, Ballestar E, Paz MF et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc. Natnl. Acad. Sci. USA. 102(30), 10604-10609. (2005).

Frank RA, McRae AF, Pocklington AJ et al. Clustered coding variants in the glutamate receptor complexes of individuals with schizophrenia and bipolar disorder. PLoS One, 6(4), e19011 (2011).

Franklin TB, Mansuy IM. Epigenetic inheritance in mammals: Evidence for the impact of adverse environmental effects. Neurobiology of disease, 2010, https://doi.org/10.1016/j.nbd.2009.11.012

Fu W, O'Connor TD, Jun G et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature.* 493, 216–220 (2013).

Fudemberg G, Imakaev M et al. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 2016 May 31; 15(9): 2038–2049. doi: 10.1016/j.celrep.2016.04.085

Fuyaka T, Lim B. et al. Enhancer control of transcriptional bursting. Cell. 166 (2), (2016), 358–368.

Gage BF, Bass AR et al. Effect of genotype-guided warfarin dosing on clinical events and anticoagulation control among patients undergoing hip or knee arthroplasty: the GIFT randomized clinical trial. JAMA, September 2017, 318(12):1115-1124.

Gage SH, Smith GD et al. G = E: What GWAS Can Tell Us about the Environment. PLoS Genetics 2016, https://doi.org/10.1371/journal.pgen.1005765

Ganguly NK, Bano R, Seth SD. Human Genome Project: Pharmacogenomics and drug development. Indian J. Exp. Biol. October 2001. 39(10):955-61

Gao, X, Becker, LC, Becker, DM et al. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet. Epidemiol.* 34, 100-105 (2010).

Gasperini M, Findlay GM et al. CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. Am. J. Hum. Genet. 2017. https://doi.org/10.1016/j.ajhg.2017.06.010

Gelernter J, Kranzler HR, Sherva R et al. Genome-Wide Association Study of Nicotine Dependence in American Populations: Identification of Novel Risk Loci in Both African-Americans and European-Americans. *Biological psychiatry* (2014)

Germain M, Saut N et al. Genetics of venous thrombosis: insights from a new genome wide association study. PLoS One, September 2011. 6(9):e25581.

Germain M, Chasman DI et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. Am J Hum Genet, April 2015, 96(4):532-42.

Ghamari A, van de Corput MPC et al. In vivo live imaging of RNA polymerase II transcription factories in primary cells. Genes & Dev. 2013. 27: 767-777 doi: 10.1101/gad.216200.113

Ghandi M, Lee D et al. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLoS Comput Biol 2014. 10: e1003711.

Ghandi M, Mohammad-Noori M et al. gkmSVM: an R package for gapped-kmer SVM. Bioinformatics 2016. 32: 2205-2207.

Ghodke-Puranik Y, Thorn CF, Lamba JK, Leeder JS, Song W, Birnbaum AK, Altman RB, Klein TE. Valproic acid pathway: pharmacokinetics and pharmacodynamics. Pharmacogenetics and Genomics. 2013 Apr;23(4):236.

Giacomini KM, Yee SW et al. Genome-Wide association studies of drug response and toxicity: an opportunity for genome medicine. Nature Reviews Drug Discovery, 2017. 16:70

Gibbs RA et al. The international HapMap project. *Nature*. 426(6968), 789-796 (2003).
Gilad Y, Rifkin SA et al. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in Genetics, 2008, DOI: https://doi.org/10.1016/j.tig.2008.06.001

Gibson G. Rare and Common Variants: Twenty arguments. Nature Reviews in Genetics, 2011. doi: 10.1038/nrg3118

Giudicessi JR, Kullo IL, Ackerman MJ. Precision Cardiovascular Medicine: State of Genetic Testing. Mayo Clinic Proceedings, April 2017, 92(4):642-662

Goel N., Stunkard A. J. et al. Circadian rhythm profiles in women with night eating syndrome. J. Biol. Rhythms 24(1), (2009), 85-94.

Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. J Comput Chem 38(16), 1291-1307 (2017).

Goldstein JA, Kelly SM, LoPresti PP et al. SMAD signaling drives heart and muscle dysfunction in a Drosophila model of muscular dystrophy. Hum Mol Genet. 2011 Mar 1;20(5):894-904. doi: 10.1093/hmg/ddq528.

Gong Y, Lazaris C et al. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. Nature Communications volume 9, Article number: 542 (2018)

Gonzalez-Sandoval A, Gasser SM. On TADs and LADs: Spatial Control Over Gene Expression. Trends in Genetics. 2016 Aug 31;32(8):485-95.

Gopalakrishnan S, Hor P, Ichida JK. New approaches for direct conversion of patient fibroblasts into neural cells. Brain Research. 2015 Oct 16.

Göttlicher M, Minucci S, Zhu P, Krämer OH, Schimpf A, Giavara S, Sleeman JP, Coco FL, Nervi C, Pelicci PG, Heinzel T. Valproic acid defines a novel class of HDAC inhibitors inducing differentiation of transformed cells. The EMBO journal. 2001 Dec 17;20(24):6969-78.

Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.* gks1066 (2012).

Green EK, Grozeva D, Forty L et al. Association at SYNE1 in both bipolar disorder and recurrent major depression. Mol. Psychiatry. 18(5), 614-617 (2013).

Greliche N, Germain M et al. A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis. BMC Med Genet, March 2013, 14:36.

Gremse M, Chang A et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res, January 2011, 39(Database issue):D507-D513.

Griminger P. Vitamin K Antagonists: The First 50 Years. The Journal of Nutrition, Volume 117, Issue 7, 1 July 1987, Pages 1325–1329

Grof P, Duffy A, Cavazzoni P et al. Is response to prophylactic lithium a familial trait? J. Clin. Psychiatry. 63:942–947 (2002).

Gross DS, Garrard WT. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* 57, 159–197 (1988).

Grundberg E, Small KS et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nature Genetics, September 2012, 44:1084-1089.

Guengerich FP. Role of Cytochrome P450 Enzymes in Drug-Drug Interactions. Advances in Pharmacology Volume 43, 1997, Pages 7-35

Guenther CA, Tasic B, Luo L, Bedell MA, Kingsley DM. A molecular basis for classic blond hair color in Europeans. *Nature Genet.* 46, 748–52 (2014)

Guida M, Iudice A, Bonanni E, Giorgi FS. Effects of antiepileptic drugs on interictal epileptiform discharges in focal epilepsies: an update on current evidence. Expert review of neurotherapeutics. 2015 Aug 3;15(8):947-59.Guo Z, Zhang L, Wu Z, Chen Y, Wang F, Chen G. In vivo direct reprogramming of reactive glial cells into functional neurons after brain injury and in an Alzheimer's disease model. Cell Stem Cell. 2014 Feb 6;14(2):188-202.

Guo Y, Xu Q et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell. 2015 Aug 13; 162(4): 900–910. doi: 10.1016/j.cell.2015.07.038

Gurvich N, Tsygankova OM, Meinkoth JL, Klein PS. Histone deacetylase is a target of valproic acid-mediated cellular differentiation. Cancer Research. 2004 Feb 1;64(3):1079-86.

Gusev A, Bhatia G et al. Quantifying Missing Heritability at Known GWAS Loci. PLoS Genetics, 2013.

Hadjab S, Franck MC, Wang Y, Sterzenbach U, Sharma A, Ernfors P, Lallemend F. A local source of FGF initiates development of the unmyelinated lineage of sensory neurons. The Journal of Neuroscience. 2013 Nov 6;33(45):17656-66.

Hafner AS, Penn AC, Grillo-Bosch D et al. Lengthening of the stargazin cytoplasmic tail increases synaptic transmission by promoting interaction to deeper domains of PSD-95. Neuron. 86, 2, 475–489 doi:10.1016/j.neuron.2015.03.013 (2015).

Halaweish I, Bambakidis T, Chang Z, Wei H, Liu B, Li Y, Bonthrone T, Srinivasan A, Bonham T, Chtraklin K, Alam HB. Addition of low-dose valproic acid to saline resuscitation provides neuroprotection and improves long-term outcomes in a large animal model of combined traumatic brain injury and hemorrhagic shock. Journal of Trauma and Acute Care Surgery. 2015 Dec 1;79(6):911-9.

Halsall JA, Turan N, Wiersma M, Turner BM. Cells adapt to the epigenomic disruption caused by histone deacetylase inhibitors through a coordinated, chromatin-mediated transcriptional response. Epigenetics & Chromatin. 2015 Sep 16;8(1):1.

Hall-Flavin DK, Winner JG, Allen JD et al. Utility of integrated pharmacogenomic testing to support the treatment of major depressive disorder in a psychiatric outpatient setting. Pharmacogenetics Genomics. 23(10), 535-548 (2013).

Hampton LH, Daubresse, M, Chang H-Y et al. Emergency department visits by adults for psychiatric medication adverse events. *JAMA Psychiatry*. 71, 1006-1014 (2014).

Hannon E, Spiers H et al. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. Nature Neuroscience volume 19, pages 48–54 (2016)

Hardison RC. Discovering enhancers directly by activity. Nat Methods. 2014 May; 11(5): 491–492. doi: 10.1038/nmeth.2933

Harter K, Levine M, Henderson SO. Anticoagulation Drug Therapy: A Review. West J Emergency Med. January 2015, 16(1):11-17.

Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature.* 489, 391-399 (2012)

He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. Genome Res. 2: 1015-1025 (2012).

He Y, Gorkin DU et al. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. PNAS, January 2017, 114(9):E1633-E1640.

He X, Li Y, Liu Z, Yue X, Zhao P, Hu J, Wu G, Mao B, Sun D, Zhang H, Song X. The association between CCL2 polymorphisms and drug-resistant epilepsy in Chinese children. Epileptic Disorders. 2013 Sep 1;15(3):272-7.

Health Quality Ontario. Pharmacogenomic Testing for Psychotropic Medication Selection: A Systematic Review of the Assurex GeneSight Psychotropic Test. Ont Health Technol Assess Ser. 2017; 17(4): 1–39.

Hebbring SJ, Rastegar-Mojarad M et al, Application of clinical text data for phenome-wide association studies (PheWASs), Bioinformatics 31(12), (2015), 1981-1987.

Heit JA, Armasu SM et al. A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. J Thromb Haemost, August 2012. 10(8):1521-31

Heit JA, Armasu SM et al. Identification of unique venous thromboembolism-susceptibility variants in African-Americans. Thromb Haemost, April 2017, 117(4):758-768.

Hellman A, Chess A. Extensive sequence-influenced DNA methylation polymorphism in the human genome. Epigenetics Chromatin. 3, 11 (2010).

Heyn H, Moran S, Hernando-Herraez I et al. DNA methylation contributes to natural human variation. Genome Res. 23(9), 1363-1372 (2013).

Higgins GA, Allyn-Feuer A, Barbour E, Athey BD. A glutamatergic network mediates lithium response in bipolar disorder as defined by epigenome pathway analysis. Pharmacogenomics 2015.

Higgins GA, Allyn-Feuer A, Athey BD. Epigenomic mapping and effect sizes of noncoding variants associated with psychotropic drug response. Pharmacogenomics 2015

Higgins GA, Allyn-Feuer A, Handelman S, Sadee W, Athey BD. The epigenome, 4D nucleome and next-generation neuropsychiatric pharmacogenomics. Pharmacogenomics 2015

Higgins GA, Allyn-Feuer A, Georgoff P, Nikolian V, Alam HB, Athey BD. Mining the Topography and dynamics of the 4D Nucleome to identify novel CAN drug pathways

Higgins GA, Georgoff P, Nikolian V, Allyn-Feuer A, Pauls B, Higgins R, Athey BD, Alam HE. Network Reconstruction Reveals that Valproic Acid Activates Neurogenic Transcriptional Programs in Adult Brain Following Traumatic Injury. Pharmaceutical Research, 2017.

Hiranuma N, Lundberg S, Lee S-I. DeepATAC: A deep-learning method to predict regulatory factor binding activity from ATAC-seq signals. bioRxiv doi:10.1101/172767, (2017).

Hoffman SMG, Keeney DS. Fine-scale mapping of CYP gene clusters: An example from human CYP4 family. Methods in Enzymology, 2002, 357:36-44.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*. 9, 473–476 (2012).

Hoft NR, Stitzel JA, Hutchison KE, Ehringe MA. CHRNB2 promoter region: association with subjective effects to nicotine and gene expression differences. *Genes, Brain and Behavior* 10: 176-185. (2011)

Hon G, Wang W, Ren B. Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput. Biol. 5, e1000566 (2009).

Hou L, Heilbronner U, Degenhardt F, Adli M, Akiyama K, Akula N, Ardau R, Arias B, Backlund L, Banzato CE, Benabarre A. Genetic variants associated with response to lithium treatment in bipolar disorder: a genome-wide association study. The Lancet. 2016 Jan 22.

Houlihan LM, Davies G et al. Common variants of large effect in F12, KNG1, and HRG are associated with activated partial thromboplastin time. Am J Hum Genet, April 2010. 86(4):626-31.

Hsieh TH, Weiner A et al. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell. 2015 Jul 2;162(1):108-19. doi: 10.1016/j.cell.2015.05.048.

Hu M, Deng K et al. Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. Quant Biol. 2013 Jun; 1(2): 156–174. doi: 10.1007/s40484-013-0016-0

Huang H, Liu CM et al. Ketamine Affects the Neurogenesis of the Hippocampal Dentate Gyrus in 7-Day-Old Rats. Neurotox Res. 2016 Aug;30(2):185-98. doi: 10.1007/s12640-016-9615-7.

Huang J, Li K et al. Dissecting super-enhancer hierarchy based on chromatin interactions. Nature Communicationsvolume 9, Article number: 943 (2018)

Huang Y, Pastor WA et al. The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. PLoS One, 2010. https://doi.org/10.1371/journal.pone.0008888

Hug CB, Grimaldi AG, Kruse K, and Vaquerizas JM (2017) Chromatin architecture emerges during zygotic genome activation independent of transcription. Cell 169: 216-228.

Hughes TA, Pombo A et al. On the structure of replication and transcription factories. J Cell Sci Suppl. 1995;19:59-65.

Huizinga TWJ, Pisetsky DS, Kimberly RP. Associations, Populations, and the Truth. Arthritis and Rheumatology, July 2004, 50(7):2066-2071.

Hung CC, Ho JL, Chang WL, Tai JJ, Hsieh TJ, Hsieh YW, Liou HH. Association of genetic variants in six candidate genes with valproic acid therapy optimization. Pharmacogenomics. 2011 Aug 15;12(8):1107-17.

Huroy S. Mechanisms of EphB2 mediated opiate-dependent tolerance and learning. MS thesis. University of Toronto, Department of Pharmaceutical Sciences; (2012)

Hyman SE. Revitalizing psychiatric therapeutics. *Neuropsychopharmacol. Rev.* 39, 220 – 229 (2014).

Iglovikov V, Rakhlin A, Kalinin A, Shvets A. Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks. arXiv preprint arXiv:1712.05053, (2017).
International HapMap Constortium. A second generation human haplotype map of over 3.1 million SNPs. Nature, 449:851-862. 2007.

Jagannathan L, Swaminathan K, Kumar SM, Kumar GR, Dey A. Bio-informatics based analysis of genes implicated in alcohol mediated liver injury. *Gene.* 494(1), 130-139. (2012)

Jamshidi Y, Nolte IM, Dalageorgou C et al. Common variation in the NOS1AP gene is associated with drug-induced QT prolongation and ventricular arrhythmia. *Journal Amer Coll Cardiol* 60: 841-850. (2012)

Jia, P, Wang L, Meltzer HY, Zhaoa Z: Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. *Schizo. Res*. 122: 38-42. (2010)

Jia Z, Agopyan N, Miu P, Xiong Z, Henderson J, Gerlai R, Taverna FA, Velumian A, MacDonald J, Carlen P, Abramow-Newerly W, Roder J. Enhanced LTP in mice deficient in the AMPA receptor GluR2. Neuron. 17:945–956 (1996)

Jie F et al. Canonical transient receptor potential 4 and its small molecule modulators. *China Life Sci.* 58 (1), 39–47 (2015) doi: 10.1007/s11427-014-4772-5

Jin F, Li Y, Dixon JR et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 503, 290-294 (2013).

Jiyun Z, Qin L, Ruifeng X, Lin G, Hongpeng W. CNNsite: Prediction of DNA-binding residues in proteins using Convolutional Neural Network with sequence features. Presented at: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2016.

John, S. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet. 43, 264–268 (2012).

John SE, Antony D et al. Genetic variants associated with warfarin response in Kuwaiti population. Pharmacogenomics, June 2017, 18(8): 757-764

Johnson JA, Cavallari LH. Warfarin Pharmacogenetics. Trends in Cardiovascular Medicine, revised January 2016. 25(1):33-41.

Johnson RC, Nelson GW et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics. 2010; 11: 724. doi: 10.1186/1471-2164-11-724

Jones MJ, Fejes AP, Kobor MS. DNA methylation, genotype and gene expression: who is driving and who is along for the ride? Genome Biology.14 (126), 1-3 (2013).

Jowhar Z, Gudla PR et al. HiCTMap: Detection and analysis of chromosome territory structure and position by high-throughput imaging. Methods. 2018 Jun 1;142:30-38. doi: 10.1016/j.ymeth.2018.01.013.

Kalinin AA, Allyn-Feuer AA et al. 3D Cell Nuclear Morphology: Microscopy Imaging Dataset and Voxel-Based Morphometry Classification Results. BioRxiv, 2017. https://doi.org/10.1101/208207

Kalinin AA, Higgins GA, Reamaroon N, Soroushmehr S, Allyn-Feuer A, Dinov ID, Najarian K, Athey BD. Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics, May 2018

Kaplun A, Hogan JD, Schacherer F, Peter AP, Krishna S, Braun BR, Nambudiry R, Nitu MG, Mallelwar R, Albayrak A. PGMD: a comprehensive manually curated pharmacogenomic database. The Pharmacogenomics Journal. 2016 Apr 1;16(2):124-8.

Kapoor A, Sekar RB, Hansen NF et al. QT Interval-International GWAS Consortium: An enhancer polymorphism at the cardiomyocyte intercalated disc protein NOS1AP locus is a major regulator of the QT interval. *Amer Journal Human Genet* 94: 854-869. (2014)

Kanehisa M, Goto S, Furumichi M et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*. 38(S1) D355-D360 (2010).

Kang HJ, Kawasawa YI, Cheng F et al. Spatio-temporal transcriptome of the human brain. *Nature*. 478, 483-489 (2011)

Karolchik D, Barber GP, Casper J et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res*. 42, D764–D770 (2014).

Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F et al Extensive variation in chromatin states across humans. Science. 8, 342 (6159):750-752 (2013).

Kato N, Takeuchi F, Tabara Y. Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nat Genet.* 43, pp. 531-538. Doi: 10.1038/ng.834 (2011)

Kaufman J, Edward Z. Do abused children become abusive parents? Amer. J. Orthopsychiatry. 57(2), 186 (1987).

Kelkhoff D, Downing T, Li S. Mechanotransduction to Epigenetic Remodeling. In Molecular and Cellular Mechanobiology 2016 (pp. 163-173). Springer New York.

Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res 26(7), 990-999 (2016).

Kelley DR, Reshef YA. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. bioRxiv doi:10.1101/161851, (2017).

Kellis M, Wold B, Snyder MP et al. Defining functional DNA elements in the human genome. *PNAS.* 111:6131–6138 (2014).

Kertes DA, Kalsi G, Prescott CA et al. Neurotransmitter and neuromodulator genes associated with a history of depressive symptoms in individuals with alcohol dependence. *Alcoholism: Clinical and Experimental Research*. 35(3), 496-505 (2011)

Kestler HA, Wawra C, Kracher B, Kühl M. Network modeling of signal transduction: establishing the global view. Bioessays. 2008 Nov 1;30(11-12):1110-25.

Kilpinen H, Waszak SM, Gschwind AR et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science. 342(6159), 744-747 (2013).

Kim B, Kim CY, Lee MJ, Joo YH. Preliminary evidence on the association between XBP1-116C/G polymorphism and response to prophylactic treatment with valproate in bipolar disorders. Psychiatry research. 2009 Aug 15;168(3):209-12.

Kim HJ, Leeds P, Chuang DM. The HDAC inhibitor, sodium butyrate, stimulates neurogenesis in the ischemic brain. Journal of Neurochemistry. 2009 Aug 1;110(4):1226-40.

Kim SG, Harwani M, Grama A, Chaterji S. EP-DNN: A Deep Neural Network-Based Global Enhancer Prediction Algorithm. Sci Rep 6, 38433 (2016).

Kim SY, Kang JY et al. Metabolism of R- and S-warfarin by CYP2C19 into four hydroxywarfarins. Drug Metab Lett. 2012 Sep 1;6(3):157-64.

Kimmel SE. Warfarin therapy: in need of improvement after all these years. Expert Opin. Pharmacother., April 2010, 9(5): 677-686

Kimmel SE, French B et al. A Pharmacogenetic versus a clinical algorithm for warfarin dosing. NEJM, December 2013, 369:2283-2293

Kimmel SE. Warfarin pharmacogenomics: current best evidence. J. Thromb. Haemo. 2015 13(S1):S266-S271.

Kino T., Chrousos G.P. Circadian CLOCK-mediated regulation of target-tissue sensitivity to glucocorticoids: implications for cardiometabolic diseases. In Pediatric Adrenal Diseases, vol. 20, pp. 116-126. Karger Publishers, 2010.

Klein JC, Chen W et al. Identifying Novel Enhancer Elements with CRISPR-Based Screens. ACS Chemical Biology 2018. DOI: 10.1021/acschembio.7b00778

Kleinjan DA, van Heyningen V. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am. J. Hum. Genet. 76, 8–32 (2005).

Kolder IC, Mizusawa Y, Postema PG et al. Family-based genome-wide association analysis for the identification of genetic modifiers of heart rate and electrocardiographic indices of conduction and repolarization in a large Dutch family with a mutation in SCN5A. *Genetic modifiers in familial cardiac rhythm disorders*, 105. (2012)

Konturek P.C., Brzozowski T. et al. Gut clock: implication of circadian rhythms in the gastrointestinal tract. J. Physiol Pharmacol. 62(2), (2011), 139-150.

Korbel J, Lee C. Genome assembly and haplotyping with Hi-C. Nature Biotechnology volume 31, pages 1099–1101 (2013).

Krämer A, Green J, Pollard J Jr, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis (IPA). *Bioinformatics.* btt703. (2013).

Krefting J, Andrade-Navarro MA, Ibn-Salem J. Evolutionary stability of topologically associating domains is associated with conserved gene regulation. BioRxiv, 2017, http://dx.doi.org/10.1101/231431

Krijger PHL and de Laat W (2017) Can We Just Say: Transcription Second? Cell 169: 184-185.

Krug A, Witt SH, Backes H. A genome-wide supported variant in CACNA1C influences hippocampal activation during episodic memory encoding and retrieval. European Arch. Psychiatry Clin. Neurosci. 264(2), 103-110 (2014).

Kudzi W, Dodoo ANO, Mills JJ. Characterisation of CYP2C8, CYP2C9 and CYP2C19 polymorphisms in a Ghanaian population. BMC Med Genet. 2009; 10: 124.

Kuhn, T. S. The Structure of Scientific Revolutions. University of Chicago press. (1962; 2012).

Kumar V, Westra HJ, Karjalainen J et al. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PLoS Genet. 9(1), e1003201 (2013).

Kumasaka N, Aoki M, Okada Y et al. Haplotypes with copy number and single nucleotide polymorphisms in CYP2A6 locus are associated with smoking quantity in a Japanese population. *PLoS One* 7:e44507. (2012)

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb 19;518(7539):317-30.

Kuttikat A, Shenker N. Pharmacological modulation of central nociception in the management of chronic musculoskeletal pain. Pain Management. 1(6), 549-556 (2011).

Labrie V, Pai S, Petronis A. Epigenetics of major psychosis: progress, problems and perspectives. Trends Genet. 28 (9), 427–435 (2012).

Ladouceur, CD Almeida JR, Birmaher B et al. Subcortical gray matter volume abnormalities in healthy bipolar offspring: Potential neuroanatomical risk marker for bipolar disorder? J. Amer. Acad. Child. Adolesc. Psychiatry. 47:532-539 (2008).

LaFramboise T. (1 July 2009). "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances". Nucleic Acids Research. 37 (13): 4181–4193. doi:10.1093/nar/gkp552.

Lando D, Stevens TJ et al. Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: An evaluation of single-cell Hi-C protocols. Nucleus. 2018 Jan 1;9(1):190-201. doi: 10.1080/19491034.2018.1438799.

Lange LA, Willer CJ, Rich SS. Recent developments in genome and exome-wide analysis of plasma lipids. *Curr. Opinion in Lipidology,* 26(2) pp 96-102, doi: 10.1097/MOL. 0000000000000159 (2015)

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008 Dec 29;9(1):1.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods volume 9, pages 357–359 (2012).

Laosombat V, Sattayasevana B et al. Glucose-6-phosphate dehydrogenase variants associated with favism in Thai children. Int J Hematol. 2006 Feb;83(2):139-43.

Larrance DT, Twentyman CT. Maternal attributions and child abuse. J. Abnormal Psychol. 92 (4), 449 (1983).

Leclercq I, Desager JP, Horsmans Y. Inhibition of chlorzoxazone metabolism, a clinical probe for CYP2E1, by a single ingestion of watercress. Clin Pharmacol Ther. 1998 Aug;64(2):144-9.

Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 521(7553), 436-444 (2015).

Lee C-C et al. The phospholipid-binding protein SESTD1 negatively regulates dendritic spine density by interfering with Rac1-Trio8 signaling pathway. *Sci. Reports.* 5:13250 (2015) doi: 10.1038/srep13250

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. A method to predict the impact of regulatory variants from DNA sequence. Nature Genetics. 2015 Aug 1;47(8):955-61.

Lee D. LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics. 2016 Jul 15;32(14):2196-8. doi: 10.1093/bioinformatics/btw142.

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res*. 21(7), 1109-1121 (2011).

Lee KE, Chang BC et al. Effects of CYP4F2 gene polymorphisms on warfarin clearance and sensitivity in Korean patients with mechanical cardiac valves. Ther Drug Monit. 2012 Jun;34(3):275-82.

Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR: Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 28: 2540-2542. (2012)

Lee W, Han K et al. Use of FISH to detect chromosomal translocations and deletions. Analysis of chromosome rearrangement in synovial sarcoma cells from paraffin-embedded specimens. Am J Pathol. 1993 Jul; 143(1): 15–19.

Leung, D., Jung, I., Rajagopal et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*. 518(7539), 350-354. (2015).

Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. P Ieee 104(1), 176-197 (2016).

Lever J, Krzywinski M, Altman N. Points of Significance: Model selection and overfitting. Nature Methods 13(9), 703-704 (2016).

Levin FR, Hennessy G. Bipolar disorder and substance abuse. Biological Psychiatry. 2004 Nov 15;56(10):738-48.

Li C1, Dong X1, Fan H et al. The 3DGD: a database of genome 3D structure. *Bioinformatics*. 30, 1640-1642 (2014).

Li G, Ruan X, Auerbach RK et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 148(1-2), 84–98 (2012).

Li G, Cai L et al. Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics 2014 15 (Suppl 12):S11 https://doi.org/10.1186/1471-2164-15-S12-S11

Li H, Liu F, Ren C, Bo X, Shu W. Genome-wide identification and characterization of HOT regions in the human genome. bioRxiv. 2016 Jan 1:036152.

Li MJ, Wang LY, Xia Z, Sham PC, Wang J. GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nuc. Acids Res.* 41:W150-W158 (2013).

Li R, Kim D, Ritchie MD. Methods to analyze big data in pharmacogenomics research. Pharmacogenomics 18(8), 807-820 (2017).

Libbrecht MW, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, Noble WS. Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. *Genome Res.* 25, 1–14 (2015).

Lieberman-Aiden E, van Berkum NL, Williams L et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. *326*(5950), 289-293. (2009).

Liefooghe A, Touzet H, Varre JS. Large scale matching for position weight matrices. CPM (Combinatorial Pattern Matching), volume 4009 of Lecture Notes in Computer Science, p. 401-412, 2006.

Lim CS, Baldessarini RJ, Vieta E, Yucel M, Bora E, Sim K. Longitudinal neuroimaging and neuropsychological changes in bipolar disorder patients: review of the evidence. Neurosci. Biobehav. Rev. 37(3), 418-435 (2013).

Limdi NA, Veenstra DL. Warfarin Pharmacogenetics. Pharmacotherapy. 2008 Sep; 28(9): 1084–1097.

Lin YF, Huang MC, Liu HC. Glycogen synthase kinase 3β gene polymorphisms may be associated with bipolar I disorder and the therapeutic response to lithium. J. Affective Dis. 147.1 401-406 (2013).

Lind M, Boman K et al. Von Willebrand factor predicts major bleeding and mortality during oral anticoagulant treatment. J Intern Med, March 2012, 271(3):239-46.

Lind PA, Zhu G, Montgomery GW et al. Genome-wide association study of a quantitative disordered gambling trait. *Addiction Biol.* 18(3), 511-522. (2013)

Lip GY, Lowe GD et al. Effects of warfarin therapy on plasma fibrinogen, von Willebrand factor, and fibrin D-dimer in left ventricular dysfunction secondary to coronary artery disease with and without aneurysms. Am J Cardiol, September 1995, 76(7):453-8.

Litjens G, Kooi T, Bejnordi BE et al. A survey on deep learning in medical image analysis. arXiv preprint arXiv:1702.05747, (2017).

Liu F, Li H, Ren C, Bo X, Shu W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. Sci Rep 6, 28517 (2016).

Liu XS, Chopp M, Kassis H, Jia LF, Hozeska-Solgot A, Zhang RL, Chen C, Cui YS, Zhang ZG. Valproic acid increases white matter repair and neurogenesis after stroke. Neuroscience. 2012 Sep 18;220:313-21.

Liu Y, Blackwood DH, Caesar S et al. Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. Mol. Psychiatry. 16(1), 2-4 (2011).

Liu Z, Guo X, Jiang Y, Zhang H. NCK2 is significantly associated with opiates addiction in African-origin men. *Scientific World Journal.* 2013:748979. (2013)

Lonsdale J, Thomas J, Salvatore M et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genet.* 45:580–585 (2013).

Lopes R, Korkmaz G, Agami R. Applying CRISPR–Cas9 tools to identify and characterize transcriptional enhancers. Nat. Rev. Mol. Cell. Biol. 2016.

Lopez S, Buil A et al. A genome-wide association study in the genetic analysis of idiopathic thrombophilia project suggests sex-specific regulation of mitochontrial DNA levels. Mitochondrion, September 2014, 18:34-40.

Lu M, Lewis CM, Traylor M. Pharmacogenetic testing through the direct-to-consumer genetic testing company 23andMe. BMC Medical Genomics, June 2017, 10:47.

Lu, Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database.* baq036 (2011).

Luger K, Dechassa ML, Tremethick DJ. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell. Biol.*, 13:436-447. (2012)

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, Santos-Simarro F. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015 May 21;161(5):1012-25.

Lv L, Sun Y, Han X, Xu CC, Tang YP, Dong Q. Valproic acid improves outcome after rodent spinal cord injury: potential roles of histone deacetylase inhibition. Brain Research. 2011 Jun 17;1396:60-8.

Malatkova P, Sokolova S et al. Carbonyl reduction of warfarin: Identification and characterization of human warfarin reductases. Biochem Pharmacol, Jun 2016, 109:83-90.

Malhi GS, Tanious M, Das P, Coulston CM, Berk M. Potential mechanisms of action of lithium in bipolar disorder. CNS drugs, 27(2), 135-153 (2013).

Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. Mol Pharmaceut 13(5), 1445-1454 (2016).

Manchia M, Adli M, Akula N et al. Assessment of response to lithium maintenance treatment in bipolar disorder: A Consortium on Lithium Genetics (ConLiGen) report. PloS One. 8(6), e65636 (2013).

Makino C, Fujii Y, Kikuta R et al. Positive association of the AMPA receptor subunit GluR4 gene (GRIA4) haplotype with schizophrenia: linkage disequilibrium mapping using SNPs evenly distributed across the gene region. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 116(1), 17-22 (2003).

Manolio TA, Collins FS et al. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8; 461(7265): 747–753. doi: 10.1038/nature08494

Manolio TA. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, 363,166–176 (2010).

Manuel MN, Mi D, Mason JO, Price DJ. Regulation of cerebral cortical neurogenesis by the Pax6 transcription factor. Frontiers in cellular neuroscience. 2015 Mar 10;9:70.

Marchion DC, Bicaku E, Daud AI, Sullivan DM, Munster PN. Valproic acid alters chromatin structure by regulation of chromatin modulation proteins. Cancer Research. 2005 May 1;65(9):3815-22.

Marsh S, King CR et al. Population variation in VKORC1 haplotype structure. Journal of Thrombosis and Haematostasis, January 2006,

Martin P, McGovern A et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nature Communications, November 2015, 6:10069

Maurano MT, Humbert R, Rynes E et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science.* 337:1190–1195. (2012)

MacArthur J, Bowler E et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research, January 2017, 45(Database Issue):D896-D901.

McCarthy MJ, Leckband SG, Kelsoe JR. Pharmacogenetics of lithium response in bipolar disorder. Pharmacogenomics. 11, 1439–1465 (2010).

McCarthy MJ, Nievergelt CM, Shekhtman T, Kripke DF, Welsh DK, Kelsoe JR. Functional genetic variation in the REV-ERBA pathway and lithium response in the treatment of bipolar disorder. Genes Brain Behav. 10 (8) 852-61 (2011).

McCarthy M.J, Nievergelt C. M. et al. A survey of genomic studies supports association of circadian clock genes with bipolar disorder spectrum illnesses and lithium response. PLoS One. 7(2), (2012), e32091.

McCarthy MJ, Le Roux MJ, Wei H, Beesley S, Kelsoe JR, Welsh DK. Calcium channel genes associated with bipolar disorder modulate lithium's amplification of circadian rhythms. Neuropharmacology. 29. 101:439-48. (2016) doi:10.1016/j.neuropharm.2015.10.017

McClay JL, Shabalin AA et al. High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. Genome Biology, December 2015. 16:291.

McDaniell R, Lee BK, Song L et al. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328:235–239 (2010).

McDonald MG, Rieder MJ et al. CYP4F2 is a vitamin K1 oxidase: An explanation for altered warfarin dose in carriers of the V433M variant. Mol Pharmacol. 2009 Jun;75(6):1337-46.

McGeachie MJ, Clemmer GL, Lasky-Su J et al. Joint GWAS Analysis: Comparing similar GWAS at different genomic resolutions identifies novel pathway associations with six complex diseases. *Genomics Data*. 2, 202–211 (2014).

McVicker G, van de Geijn B, Degner JF et al. Identification of genetic variants that affect histone modifications in human cells. Science. 342, 747-749 (2013).

Mead AN, Stephens DN. Involvement of AMPA receptor GluR2 subunits in stimulus-reward learning: evidence from glutamate receptor gria2 knock-out mice. J. Neuroscience. 2003 Oct 22;23(29):9500-7.

Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41(D1), D377-D386 (2013).

Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic Acids Research. 2016 Jan 4;44(D1):D336-42.

Mehta D, Klengel T, Conneely KN et al. Childhood maltreatment is associated with distinct genomic and epigenetic profiles in posttraumatic stress disorder. *Proceedings of the National Academy of Sciences*, 110(20), 8302-8307. (2013)

Merikangas, KR, Jin R, He JP et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch. Gen. Psychiatry. 68 (3), 241-251 (2011).

Merkenschlager M, Odom DT. CTCF and cohesin: Linking gene regulatory elements with their targets. *Cell.* 152, 1285-1297 (2013).

Meyer UA, Zanger UM, Schwab M. Omics and drug response. *Annu. Rev. Pharmacol. Toxicol.* 53, 475–502 (2013).

Miettinen TP, Björklund M. NQO2 is a reactive oxygen species generating off-target for acetaminophen. *Mol. Pharm.* 11:4395–4404. (2014)

Mifsud B, Tavares-Cadete F et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015 Jun;47(6):598-606. doi: 10.1038/ng.3286

Mihalas AB, Hevner RF. Control of Neuronal Development by T-Box Genes in the Brain. Current Topics in Developmental Biology. 2016 Sep 1.

Millar J. The Need for a Global Language – SNOMED CT Introduction. Stud Health Technol Inform, 2016, 225:683-5.

Min X, Zeng WW, Chen N, Chen T, Jiang R. Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. Bioinformatics 33(14), I92-I101 (2017).

Mishra A, Hawkins RD. Three-dimensional genome architecture and emerging technologies: looping in disease. Genome Med. 2017; 9: 87. doi: 10.1186/s13073-017-0477-2

Mitchell KJ. The genetics of neurodevelopmental disease. *Current opinion in neurobiology*, 21(1), 197-203. (2012)

Mitjans M, Arias B et al. Exploring Genetic Variability at PI, GSK3, HPA, and Glutamatergic Pathways in Lithium Response: Association With IMPA2, INPP1, and GSK3B Genes. J. Clin. Psychopharm., October 2015. 35(5):600-604

Moaddeb J, Haga SB. Pharmacogenetic testing: current evidence of clinical utility. Ther. Adv. Drug Saf. 4(4):155-169 (2013).

Modig F, Patel M, Magnusson M, Fransson PA. Study II: Mechanoreceptive sensation is of increased importance for human postural control under alcohol intoxication. *Gait & posture,* 35(3), 419-427. (2012)

Mokry M, Middendorp S, Wiegerinck CL et al. Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. *Gastroenterology*. 146, 1040–1047 (2014).

Mrazek, D. Psychiatric Pharmacogenomics. Oxford University Press, 2010.

Mullins N, Perroud N, Uher R et al. Genetic relationships between suicide attempts, suicidal ideation and major psychiatric disorders: A genome-wide association and polygenic scoring study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 165(5), 428-437. (2014)

Mumbach MR, Rubin AJ et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nature Methods volume 13, pages 919–922 (2016)

Murphy E & McMahon FJ. Pharmacogenetics of antidepressants, mood stabilizers, and antipsychotics in diverse human populations. Dis. Med. (2013). 16(87), 113-122.

Murphy FC, Sahakian BJ. Neuropsychology of bipolar disorder. The British Journal of Psychiatry. 2001 Jun 1;178(41):s120-7.

Musunuru K, Strong A, Frank-Kamenetsky M et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature.* 466:714–719. (2010)

Nagano T, Lubling Y et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature volume 502, pages 59–64 (03 October 2013)

Narayanan R, Tuoc TC. Roles of chromatin remodeling BAF complex in neural differentiation and reprogramming. Cell Tissue research. 2014 Jun 1;356(3):575-84.

Nelson DR, Goldstone J, Stegeman JJ. The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s. Philos. Trans R Soc Lond B Biol Sci. February 2013, 368(1612):20120474.

Nelson EC, Lynskey MT, Heath AC et al. Association of OPRD1 polymorphisms with heroin dependence in a large case-control series. *Addiction Biol*. 19(1), 111-121 (2014).

Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. Cell. 2012 Sep 14;150(6):1274-86.

The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neurosci.* 18, 199–209 (2015).

Nichols MH, Corces VG. A CTCF Code for 3D Genome Architecture. Cell. 2015 Aug 13; 162(4): 703–705. doi: 10.1016/j.cell.2015.07.053

Ninkovic J, Steiner-Mezzadri A, Jawerka M, Akinci U, Masserdotti G, Petricca S, Fischer J, Von Holst A, Beckers J, Lie CD, Petrik D. The BAF complex interacts with Pax6 in adult neural progenitors to establish a neurogenic cross-regulatory transcriptional network. Cell Stem Cell. 2013 Oct 3;13(4):403-18.

Nishizaki SS, Boyle AP. Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. Trends Genet. Janaury 2017 33(1):34-45

Noga JT, Vladar K, Torrey EF. A volumetric magnetic resonance imaging study of monozygotic twins discordant for bipolar disorder. Psychiatry Res. 106, 25-34 (2004).

Noordermeer D, de Wit E, Klous P et al. Variegated gene expression caused by cell-specific long-range DNA interactions. Nat. Cell. Biol. 13(6), 944-954 (2011).

Nora EP, Dekker J, Heard E. Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? Bioessays. 2013 Sep; 35(9): 818–828. doi: 10.1002/bies.201300040

Northcott PA, Lee C et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature. 2014 Jul 24; 511(7510): 428–434. doi: 10.1038/nature13379

Notwell JH, Heavner WE, Darbandi SF, Katzman S, McKenna WL, Ortiz-Londono CF, Tastad D, Eckler MJ, Rubenstein JL, McConnell SK, Chen B. TBR1 regulates autism risk genes in the developing neocortex. Genome Research. 2016 Aug 1;26(8):1013-22.

O'Leary NA, Wright MW et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-45

Oedegaard KJ, Alda M, Anand A, Andreassen OA, Balaraman Y, Berrettini WH, Bhattacharjee A, Brennand KJ, Burdick KE, Calabrese JR, Calkin CV. The Pharmacogenomics of Bipolar Disorder study (PGBD): identification of genes for lithium response in a prospective sample. BMC psychiatry. 2016 May 5;16(1):1.

Ogu CC, Maxa JL. Drug interactions due to cytochrome P450. Proc (Bayl Univ Med Cent). 2000 Oct; 13(4): 421–423.

Olova N, Krueger F et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biology 2018. https://doi.org/10.1186/s13059-018-1408-2

Onengut-Gumuscu S, Chen W-G, Burren O et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for co-localization of causal variants with lymphoid gene enhancers. *Nature Genet.* (2015). doi:10.1038/ng.3245

Ortega VE, Meyers DA. Pharmacogenetics: Implications of Race and Ethnicity on Defining Genetic Profiles for Personalized Medicine. J Allergy Clin Immunol, January 2014, 133(1):16-26.

Osborne CS, Chakalova L, Brown KE et al. Active genes dynamically co-localize to shared sites of ongoing transcription. *Nature Genet.* 36, 1065–1071 (2004).

Ostrousky O, Meged S, Loewenthal R et al. NQO2 gene is associated with clozapine-induced agranulocytosis. *Tissue Antigens*. 62:483–491. (2003)

Ou X, Crane DE, MacIntosh BJ, Young LT, Arnold P, Ameis S, Goldstein BI. CACNA1C rs1006737 genotype and bipolar disorder: Focus on intermediate phenotypes and cardiovascular comorbidity. Neuroscience & Biobehavioral Reviews. 2015 Aug 31;55:198-210.

Oudot-Mellakh T, Cohen W et al. Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein c anticoagulant pathway: the MARTHA project. Br. J. Haematol. April 2012. 157(2):230-9.

Ouwenga R, Lake AM et al. Transcriptomic Analysis of Ribosome-Bound mRNA in Cortical Neurites In Vivo. J Neurosci. 2017 Sep 6;37(36):8688-8705. doi: 10.1523/JNEUROSCI.3044-16.2017.

Palacios R, Gazave E, Goñi J et al. Allele-specific gene expression is widespread across the genome and biological processes. *PLoS One* 4:e4150 (2009).

Pan X, Shen HB. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. BMC Bioinformatics 18(1), 136 (2017).

Pan X, Rijnbeek P, Yan J, Shen H-B. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. bioRxiv doi:10.1101/146175, (2017).

Pandit S, Zhou Y, Shiue L et al. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol. Cell.* 50(2), 223-235 (2013).

Papantonis A, Cook PR. Transcription factories: genome organization and gene regulation. Chem Rev. 2013 Nov 13;113(11):8683-705. doi: 10.1021/cr300513p.

Park J, Gail MH, Weinberg CR et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *PNAS*, 108(44) pp 18026-18031. Doi: 10.1073/pnas.1114759108 (2011)

Park JH, Wacholder S, Gail MH et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet.* June, 42 pp 570-575. (2010) Doi: 10.1038/ng.610

Park Y, Kellis M. Deep learning for regulatory genomics. Nat Biotechnol 33(8), 825-826 (2015).

Parker SCJ, Stitzel ML, Taylor DL et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci USA 110:17921–17926. (2013)

Parra EJ, Botton MR et al. Genome-wide association study of warfarin maintenance dose in a Brazilian sample. Pharmacogenomics, August 2015. 16(11):1253-63.

Parsons MJ, Lester KJ, Barclay NL, Nolan PM, Eley TC, Gregory AM. Replication of genome-wide association studies (GWAS) loci for sleep in the British G1219 cohort. Amer. J. Med. Genet. 162B (5), 431-438 (2013).

Passmore MJ, Garnham J, Duffy A et al. Phenotypic spectra of bipolar disorder in responders to lithium versus lamotrigine. Bipolar Disord. 5, 110–114 (2003).

Patillon B, Luisi P et al. Positive Selection in the Chromosome 16 VKORC1 Genomic Region Has Contributed to the Variability of Anticoagulant Response in Humans. PLoS One, December 2012, doi:10.1371/journal.pone.0053049

Peled, E. Abused women who abuse their children: A critical review of the literature. Aggress. Violent Behavior. 16(4), 325-330 (2011).

Perlis RH, Smoller JW, Ferreira MA et al. A genomewide association study of response to lithium for prevention of recurrence in bipolar disorder. Amer. J. Psychiatry.166, 718–725 (2009).

Perera MA, Cavallari LH et al. Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study. Lancet, August 2013, 382(9894):790-6.

Perucca E. Pharmacological and therapeutic properties of valproate. CNS Drugs. 2002 Oct 1;16(10):695-714.

Peters JE, Lyons PA et al, Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. PLoS Genetics, 2016, https://doi.org/10.1371/journal.pgen.1005908

Petronis A, Gottesman II, Kan P, Kennedy JL, Basile VS, Paterson AD, Popendikyte V. Monozygotic twins exhibit numerous epigenetic differences: clues to twin discordance? Schizophrenia Bull. 29(1), 169-178 (2003).

Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature. 465, 721–727 (2010).

Phillips ML, Schwartz HA. A critical appraisal of neuroimaging studies of bipolar disorder: Toward a new conceptualization of underlying neural circuitry and a road map for future research. Amer. J. Psychiatry, advanced online publication, June (2014).

Pirmohamed M. Pharmacogenetics and Pharmacogenomics. Br J Clin Pharmacol. 2001 Oct; 52(4): 345–347. doi: 10.1046/j.0306-5251.2001.01498.x

Pirmohamed M, Burnside G et al. A Randomized Trial of Genotype-Guided Dosing of Warfarin. NEJM, December 2013. 369:2294-2303.

Plomin R, Deary IJ. Genetics and intelligence differences: five special findings. Mol Psychiatry 2015. doi: 10.1038/mp.2014.105

Plomin R, von Stumm S. The new genetics of intelligence. Nature Reviews in Genetics 2018.

Pope B.D., Ryba T. et al, Topologically associating domains are stable units of replication-timing regulation, Nature. 515 (7527), (2014),402-405.

Pomerantz, J. Google Scholar and 100 percent availability of information. Info. *Tech. Lib.* 25(2), 52-56 (2013).

Poplin R, Newburger D, Dijamco J et al. Creating a universal SNP and small indel variant caller with deep neural networks. bioRxiv doi:10.1101/092890, (2018).

Poulos TL, Finzel BC, Howard AJ. High-resolution crystal structure of cytochrome P450cam. J Mol Biol. 1987 Jun 5;195(3):687-700.

Provencal N, Binder EB. The effects of early life stress on the epigenome: From the womb to adulthood and even before. *Exp. Neurology,* Vol. 286, pp 10-20 (2014).

Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat. Genet. 43(10), 977-983 (2011).

Purcell S, Neale B et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics. 2007, 81.

Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res 44(11), e107 (2016).

Quang D, Xie X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. bioRxiv doi:10.1101/151274, (2017).

Qin Q, Feng JX. Imputation for transcription factor binding predictions based on deep learning. Plos Computational Biology 13(2), (2017).

Quinodoz SA, Ollikainen N et al. Higher-order inter-chromosomal hubs shape 3-dimensional genome organization in the nucleus. BioRxiv. doi: https://doi.org/10.1101/219683

Rackham OJ, Firas J, Fang H, Oates ME, Holmes ML, Knaupp AS, Suzuki H, Nefzger CM, Daub CO, Shin JW, Petretto E. A predictive computational framework for direct reprogramming between human cell types. Nature Genetics. 2016 Jan 18.

Rao SSP, Huntley MH, Durand NC et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 159, 1665–1680 (2014).

Raj Arjun A., van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 135 (2), (2008), 216-226.

Rajapakse I, Groudine M. On emerging nuclear order. J. Cell Biology, 2011. DOI: 10.1083/jcb.201010129

Rakhlin A, Shvets A, Iglovikov V, Kalinin AA. Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis. arXiv preprint arXiv:1802.00752, (2018).

Ramani V, Shendure J, Duan Z. Understanding Spatial Genome Organization: Methods and Insights. Genomics, Proteomics & bioinformatics. 2016 Feb 29;14(1):7-20.

Ramani V, Deng X et al. Massively multiplex single-cell Hi-C. Nature Methods volume 14, pages 263–266 (2017)

Ramsundar B, Liu B, Wu Z et al. Is Multitask Deep Learning Practical for Pharma? Journal of Chemical Information and Modeling doi:10.1021/acs.jcim.7b00146, (2017).

Razvodovsky Y, Borodinsky A, Pascual-Mora M et al. Basic research. *Alcohol and Alcoholism*, 48(suppl 1), i41-i46. (2013)

Regev A, Teichmann S, Lander ES et al. The Human Cell Atlas. BioRxiv. doi:10.1101/121202

Rietschel M, Mattheisen M, Frank J et al. Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression. Biol. Psychiatry. 68(6) 578-585 (2010).

Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature.* 518, 317–330 (2015) doi: 10.1038/nature14248

Robson MI, de las Heras JI et al. Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. Genome Res. 2017. 27: 1126-1138 doi: 10.1101/gr.212308.116

Rocanin-Arjo A, Cohen W et al. A meta-analysis of genome wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential. Blood, January 2014. 123(5):777-85.

Rosenbloom KR, Dreszer TR, Pheasant M et al. ENCODE whole-genome data in the UCSC genome browser. *Nucleic Acids Res*. 38, D620-D625 (2010).

Rossiter MC. INCIDENCE AND CONSEQUENCES OF INHERITED ENVIRONMENTAL EFFECTS. Annual Review of Ecology and Systematics, November 1996.

Roussos P, Mitchell AC, Voloudakis G et al. A role for noncoding variation in schizophrenia. *Cell Reports*. 9, 1417–29 (2014).

Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010 Dec 24;330(6012):1787-97.

Rueckert EH, Barker D, Ruderfer D et al. Cis-acting regulation of brain-specific ANK3 gene expression by a genetic variant associated with bipolar disorder. Mol. Psychiatry. 18(8), 922-999 (2013).

Ruiz-Velasco M, Zaugg JB. Structure meets function: How chromatin organisation conveys functionality. Current Opinion in Systems Biology Volume 1, February 2017, Pages 129-136 https://doi.org/10.1016/j.coisb.2017.01.003

Rust MJ, Bates M, Zhuang X. Stochastic optical reconstruction microscopy (STORM) provides sub-diffraction-limit image resolution. Nat Methods. 2006 Oct; 3(10): 793–795. doi: 10.1038/nmeth929

Rybakowski JK. Lithium: sixty years thereafter. Neuropsychobiol. 62:5–7 (2010).

Rybakowski JK. Genetic influences on response to mood stabilizers in bipolar disorder. CNS Drugs. 27, 165–173 (2013).

Rybakowski JK. Response to lithium in bipolar disorder: Clinical and genetic findings. ACS Chem. Neurosci. (2014).

Rybakowski JK, Dmitrzak-Weglarz M, Dembinska-Krajewska D, Hauser J, Akiskal KK, Akiskal HH. Polymorphism of circadian clock genes and temperamental dimensions of the TEMPS-A in bipolar disorder. J. Affective Disorders, 159, 80–84 (2014).

Sabater-Lleal M, Martinez-Perez A et al. A genome-wide association study identifies KNG1 as a genetic determinant of plasma factor XI level and activated partial thromboplastin time. Arterioscler Thromb Vasc Biol, August 2012. 32(8):2008-16.

Sadee W, Hartmann K et al. Missing heritability of common diseases and treatments outside the protein-coding exome. Hum Genet. 2014 Oct; 133(10): 1199–1215. doi: 10.1007/s00439-014-1476-7

Sahar S, Zocchi L, Kinoshita C, Borrelli E, Sassone-Corsi P. Regulation of BMAL1 protein Stability and circadian function by GSK3b-mediated phosphorylation. PLoS ONE. 5, 1 e8561. (2010).

Saint-André V, Federation AJ, Lin CY, Abraham BJ, Reddy J, Lee TI, Bradner JE, Young RA. Models of human core transcriptional regulatory circuitries. Genome research. 2016 Mar 1;26(3):385-96.

Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10(3), e0118432 (2015).

Sanborn AL, Rao SSP et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. PNAS November 24, 2015. 112 (47) E6456-E6465. https://doi.org/10.1073/pnas.1518552112

Sangkuhl K, Klein TE, Altman RB. Clopidogrel pathway. Pharmacogenet Genomics, July 2011, 20(7):463-465.

Sarntivijai S, Lin Y et al. CLO: The cell line ontology. J Biomed Semantics, August 2014, 5:37.

Schalkwyk LC, Meaburn EL, Smith R et al. Allelic skewing of DNA methylation is widespread across the genome. *Am J. Hum Genet.* 86:196–212 (2010).

Schaub MA, Boyle AP, Kundaje A et al. Linking disease associations with regulatory information in the human genome. *Genome Res*. 22, 1748-1759 (2012).

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 43(10), 969-976 (2011).

Schmitt E, Schwarz-Finsterle J et al. COMBinatorial Oligo FISH: directed labeling of specific genome domains in differentially fixed cell material and live cells. Methods Mol Biol. 2010;659:185-202. doi: 10.1007/978-1-60761-789-1_13.

Schneider E, Pliusch G et al. Spatial, temporal and interindividual epigenetic variation of functionally important DNA methylation patterns. Nucleic Acids Res. 2010 Jul;38(12):3880-90. doi: 10.1093/nar/gkq126.

Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. Rare Allele Hypotheses for Complex Diseases. Curr Opin Genet Dev. 2009. doi: 10.1016/j.gde.2009.04.010

Schreiber J, Libbrecht M, Bilmes J, Noble W. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. bioRxiv doi:10.1101/103614, (2017).

Schulze TG, Detera-Wadleigh SD, Akula N et al. Two variants in Ankyrin 3 (ANK3) are independent genetic risk factors for bipolar disorder. Mol. Psychiatry. 14(5), 487-491 (2009).

Schulze T. The Consortium on Lithium Genetics (ConLiGen) genome-wide association studies of lithium response phenotypes in bipolar disorder [abstract book]. CINP Congress; Stockholm: 36. 3–7 (2012).

Schumacher-Schuh AF, Altmann V, Rieck M et al. Association of common genetic variants of HOMER1 gene with levodopa adverse effects in Parkinson's disease patients. Pharmacogenomics J. doi: 10.1038/tpj.2013.37. (2013).

Schurr J, Coras R, Rössler K, Pieper T, Kudernatsch M, Holthausen H, Winkler P, Woermann F, Bien CG, Polster T, Schulz R. Mild Malformation of Cortical Development with Oligodendroglial Hyperplasia in Frontal Lobe Epilepsy: A New Clinico-Pathological Entity. Brain Pathology. 2016 Feb 1.

Scordo MG, Pengo V et al. Influence of CYP2C9 and CYP2C19 genetic polymorphisms on warfarin maintenance dose and metabolic clearance. Clin Pharmacol Ther. 2002 Dec;72(6):702-10.

Seaman L, Meixner W et al. Periodicity of nuclear morphology in human fibroblasts. Nucleus. 2015;6(5):408-16. doi: 10.1080/19491034.2015.1095432.

Seaman L, Rajapakse R. 4D nucleome Analysis Toolbox: analysis of Hi-C data with abnormal karyotype and time series capabilities. Bioinformatics. 2018 Jan 1;34(1):104-106. doi: 10.1093/bioinformatics/btx484.

Seelan RS, Khalyfa A, Lakshmanan J, Casanova MF, Parthasarathy RN. Deciphering the lithium transcriptome: microarray profiling of lithium-modulated gene expression in human neuronal cells. Neuroscience. 151, 1184-1197 (2008).

Seitan VC, Faure AJ, Zhan Y et al. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. 223, 2066-2077 (2013).

Selvakumar B, Huganir, RL, Snyder SH. S-nitrosylation of stargazin regulates surface expression of AMPA-glutamate neurotransmitter receptors. Proc. Natl. Acad. Sci. USA. 106, 16440-16445 (2009).

Selvaraj S, Dixon JR, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotech.* 31(12), 1111-1118. (2013). doi:10.1038/nbt.2728

Shachar S, Pegaroro G, Misteli T. HIPMap: A High-Throughput Imaging Method for Mapping Spatial Gene Positions. Cold Spring Harb Symp Quant Biol. 2015;80:73-81. doi: 10.1101/sqb.2015.80.027417

Shannon P, Richards M. An Annotated Collection of Protein-DNA Binding Sequence Motifs. R Bioconductor 2018. DOI: 10.18129/B9.bioc.MotifDb

Shao L, Kner P et al. Super-resolution 3D microscopy of live whole cells using structured illumination. Nature Methods volume 8, pages 1044–1046 (2011)

Sherry ST, Ward MH, Kholodov M et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308-311 (2001).

Shvets A, Rakhlin A, Kalinin A, Iglovikov V. Automatic Instrument Segmentation in Robot-Assisted Surgery Using Deep Learning. bioRxiv doi:10.1101/275867, 275867 (2018).

Shi G, Boerwinkle E, Morrison AC, Gu CC, Chakravarti A, Rao DC: Mining gold dust under the genome wide significance level: A two-stage approach to analysis of GWAS. *Genet Epidemiol* 35:111–118. (2011)

Shi J, Marconett CN et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nature Communications, February 2014. 5:3365.

Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. 20, 883-889 (2010).

Silberberg G, Levit A, Collier D, Clair DS, Munro J, Kerwin RW, Tondo L, Floris G, Breen G, Navon R. Stargazin involvement with bipolar disorder and response to lithium treatment. Pharmacogenetics and genomics. 2008 May 1;18(5):403-12.

Simonis M, Klous P et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nature Genetics volume 38, pages 1348–1354 (2006)

Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics 32(17), i639-i648 (2016).

Singh S, Yang Y, Poczos B, Ma J. Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks. bioRxiv doi:10.1101/085241, (2016).

Sivakumaran S, Agakov F et al. Abundant Pleiotropy in Human Complex Diseases and Traits. American Journal of Human Genetics 2011. https://doi.org/10.1016/j.ajhg.2011.10.004

Skoda RC, Gonzalez FJ et al. Two mutant alleles of the human cytochrome P-450db1 gene (P450C2D1) associated with genetically deficient metabolism of debrisoquine and other drugs. Proc Natl Acad Sci U S A. 1988 Jul;85(14):5240-3.

Slatkin M. Epigenetic inheritance and the missing heritability problem. Genetics. 182(3), 845-850 (2009).

Smith KR, Kopeikina KJ, Fawcett-Patel JM et al. Psychiatric risk factor ANK3/Ankyrin-G nanodomains regulate the structure and function of glutamatergic synapses. Neuron. 84(2), 399-415 (2015).

Son EY, Crabtree GR. The role of BAF (mSWI/SNF) complexes in mammalian neural development. InAmerican Journal of Medical Genetics Part C: Seminars in Medical Genetics 2014 Sep 1 (Vol. 166, No. 3, pp. 333-349).

Song J et al. Genome-wide association study identifies SESTD1 as a novel risk gene for lithium-responsive bipolar disorder. *Mol. Psychiatry.* 1–8. (2015) doi:10.1038/mp.2015.165

Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010 Feb; 2010(2): pdb.prot5384. doi: 10.1101/pdb.prot5384

Soronen P, Ollila HM, Antila M et al. Replication of GWAS of bipolar disorder: association of SNPs near CDH7 with bipolar disorder and visual processing. Mol. Psychiatry. 15, 4–6 (2010).

Soumier A, Carter RM et al. New Hippocampal Neurons Mature Rapidly in Response to Ketamine But Are Not Required for Its Acute Antidepressant Effects on Neophagia in Rats. eNeuro. 2016 Mar-Apr; 3(2): ENEURO.0116-15.2016.

Spellmann I, Rujescu D, Musil R et al. Homer-1 polymorphisms are associated with psychopathology and response to treatment in schizophrenic patients. J Psychiatr Res. 45(2):234-241 (2011).

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, Heitner S. The UCSC Genome Browser database: 2016 update. Nucleic Acids Research. 2016 Jan 4;44(D1):D717-25.

Spurrell CH, Dickel DE, Visel A. The Ties That Bind: Mapping the Dynamic Enhancer-Promoter Interactome. Cell 167, November 17, 2016

Squassina A, Manchia M, Borg J et al. Evidence for association of an ACCN1 gene variant with response to lithium treatment in Sardinian patients with bipolar disorder. Pharmacogenomics. 12, 1559–1569 (2011).

Stambolic V, Ruel L, Woodgett JR. Lithium inhibits glycogen synthase kinase-3 activity and mimics wingless signalling in intact cells. Curr. Biol. 6, 1664-1668 (1996).

Stamatoyannopoulos J. Connecting the regulatory epigenome. Nature Genetics, April 2016, 48:479-480.

Stankiewicz AM, Swiergiel AH, Lisowski P. Epigenetics of stress adaptations in the brain. *Brain Res. Bull.*, Vol. 98, pp 76-92 (2013).

Staunstrup NH, Starnawska A et al. Genome-wide DNA methylation profiling with MeDIP-seq using archived dried blood spots. Clinical Epigenetics 2016. https://doi.org/10.1186/s13148-016-0242-1

Stein R. Race reemerges in debate over 'personalized medicine.' The Washington Post, Health & Science, July 31, 2011.

Stergiopoulos K, Brown DL. Genotype-Guided vs Clinical Dosing of Warfarin and its Analogues: Meta-analysis of Randomized Clinical Trials. JAMA Internal Medicine, August 2014. 174(8):1330-1338.

Stevens TJ, Lando D et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature volume 544, pages 59–64 (06 April 2017)

Stormo GD, Schneider TD et al. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Research. 1982 10 (9): 2997–3011. doi:10.1093/nar/10.9.2997

Strakowski SM, Adler CM, Almeida J et al. The functional neuroanatomy of bipolar disorder: a consensus model. Bipolar Disord. 14, 313–325 (2012).

Stringer S, Wray NR, Kahn RS, Derks EM. Underestimated Effect Sizes in GWAS: Fundamental Limitations of Single SNP Analysis for Dichotomous Phenotypes. *PLoS One*, doi: 10.1371/journal.pone.0027964 (2011)

Stunnenberg HG, International Human Epigenome Consortium et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell, November 2016, 167(5):1145-1149

Su T, Bao Z, Zhang QY, Smith TJ, Hong JY, Ding X. Human cytochrome P450 CYP2A13: Predominant expression in the respiratory tract and its high efficiency metabolic activation of a tobacco-specific carcinogen, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. *Cancer Res.* 60:5074–5079. (2000)

Sulem P, Gudbjartsson DF, Geller F et al. Sequence variants at CYP1A1-CYP1A2 and AHR associate with coffee consumption. *Hum Mol Genet*, 20:2071–2077. (2011)

Sullivan PF1, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Rev. Genet.* 13, 537-551 (2012).

Sunkin SM, et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nuc. Acids Res.* 41, D996-D1008 (2013).

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic acids research. 2014 Oct 28:gku1003.

Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics and Chromatin, December 2015, 8:57

Takeuchi F, McGinnis R et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genetics, March 2009, 5(3):e1000433.

Talpale M et al. A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. *Cell.* 17; 158(2), 434–448 (2014) doi: 10.1016/j.cell.2014.05.039

Tang W, Schwienbacher C et al. Genetic associations for advanced partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease. Am J Hum Genet, July 2012. 91(1):152-62.

Tang W, Teichert M et al. A genome-wide association study for venous thromboembolism: the extended cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium. Genet Epidemiol. July 2013, 37(5):512-521.

Taşan M, Musso G, Hao T et al. Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nature Methods*. doi:10.1038/nmeth.3215 (2014).

Teng L, He B, Wang J, Tan K. 4DGenome: a comprehensive database of chromatin interactions. Bioinformatics. 2015 Aug 1;31(15):2560-4.

Teichert M, Eijgelsheim M et al. A genome-wide association study of acenocoumarol maintenance dosage. Hum. Mol. Genet., October 2009. 18(19):3758-68.

Tesar, Paul J., Robert H. Miller, and Fadi J. Najm. Cell fate conversion of differentiated somatic cells into glial cells. U.S. Patent Application 13/280,562, filed April 26, 2012.

Thakurela S, Sahu SK, Garding A, Tiwari VK. Dynamics and function of distal regulatory elements during neurogenesis and neuroplasticity. Genome Research. 2015 Sep 1;25(9):1309-24.

Thibodeau A, Marquez EJ et al. Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin. Sci Rep. 2017; 7: 14466. doi: 10.1038/s41598-017-14389-7

Thier S, Lorenz D, Nothnagel M et al. Polymorphisms in the glial glutamate transporter SLC1A2 are associated with essential tremor. *Neurology*. 79(3), 243-248. (2012)

Timofeeva MN, McKay JD, Smith GD et al. Genetic polymorphisms in 15q25 and 19q13 loci, cotinine levels, and risk of lung cancer in EPIC. *Cancer Epidemiol Biomarkers Prev* 20:2250–2261. (2011)

Tjong H, Li W et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proceedings of the National Academy of Sciences. (2016), E1663–E1672, doi: 10.1073/pnas.1512577113.

Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*. 42:441–447. (2010)

Tomás M, Napolitano C, De Giuli L et al. Priori Polymorphisms in the NOS1AP gene modulate QT interval duration and risk of arrhythmias in the long QT syndrome. *Journal Amer Coll Cardiol* 55: 2745-2752. (2010)

Tomita H, Yokooji Y, Ishibashi T, Imanaka T, Atomi H. An archaeal glutamate decarboxylase homolog functions as an aspartate decarboxylase and is involved in β-alanine and coenzyme A biosynthesis. J. Bacteriol. 196(6), 1222-1230 (2014).

Tregouet DA, Heath S et al. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. Blood, May 2009. 113(21):5298-303.

Trinks J, Hulaniuk ML, Redal MA, Flichman D. Clinical utility of pharmacogenomics in the management of hepatitis C. Pharmgenomics Pers. Med. 7:339-47 (2014).

Turecki G, Grof P, Grof E et al. Mapping susceptibility genes for bipolar disorder: a pharmacogenetic approach based on excellent response to lithium. Mol. Psychiatry. 6:570–578 (2001).

Turner JA, Mejino JLV et al. Application of neuroanatomical ontologies for neuroimaging data annotation. Front. Neuroinformatics, June 2010, 4:10.

Undas A, Brzezinska-Kolarz B et al. Factor XIII Val34Leu polymorphism and gamma-chain cross-linking at the site of microvascular injury in healthy and coumadin-treated subjects. J Thromb Haemost., September 2005, 3(9):2015-2021

Uppu S, Krishna A. Improving Strategy for Discovering Interacting Genetic Variants in Association Studies. 9947, 461-469 (2016).

Uppu S, Krishna A, Gopalan RP. A Deep Learning Approach to Detect SNP Interactions. Journal of Software 11(10), 965-975 (2016).

United States Federal Drug Administration. Table of Pharmacogenomic Biomarkers in Drug Labeling. https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm Retrieved 2015.

United States Federal Drug Administration. Highlights of Prescribing Information: Coumadin (warfarin sodium) tablets, for oral use. Revised August 2017. Reference ID 4139177.

Urban TJ. Race, Ethnicity, Ancestry, and Pharmacogenetics. Mt. Sinai J. Medicine, March 2010, 77(2):133-9.

Valvezan AJ, Klein PS. GSK-3 and Wnt signaling in neurogenesis and bipolar disorder. Frontiers Mol. Neurosci. 5 doi: 10.3389/fnmol.2012.00001. (2012).

Van Steensel B, Belmont AS. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. Cell. 2017 May 18;169(5):780-791. doi: 10.1016/j.cell.2017.04.022.

Venter JC, Adams MD, Myers EW et al. The sequence of the human genome. *Science.* 291, 1304-1351(2001).

Verpelli C, Galimberti I, Gomez-Mancilla B, Sala C. Molecular basis for prospective pharmacological treatment strategies in intellectual disability syndromes. *Developmental neurobiology*, 74(2), 197-206. (2014)

Vesell ES, Page JG. Genetic control of drug levels in man: phenylbutazone. Science. 1968 Mar 29;159(3822):1479-80.

Vidyasagar M. Identifying Predictive Features in Drug Response Using Machine Learning: Opportunities and Challenges. Annu Rev Pharmacol 55, 15-34 (2015).

Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Südhof TC, Wernig M. Direct conversion of fibroblasts to functional neurons by defined factors. Nature. 2010 Feb 25;463(7284):1035-41.

Visscher PM, Wray NR et al. 10 Years of GWAS: Biology, Function, and Translation. AJHG, July 2017, 101(1):5-22.

Wadelius M, Chen LY et al. Association of warfarin dose with genes involved in its action and metabolism. Hum Genet, March 2007, 121(1):23-34.

Wanat MA. Novel oral anticoagulants: a review of new agents. Postgrad Med, July 2013, 125(4):103-14.

Wang C, Xu Y, Feng X et al. Linkage analysis and whole-exome sequencing exclude extra mutations responsible for the parkinsonian phenotype of spinocerebellar ataxia-2. *Neurobiol. Aging.* 36 545-e1 (2015)

Wang D, Chen H, Momary KM, Cavallar LH, Johnson JA, Sadée W. Regulatory polymorphism in vitamin K epoxide reductase complex subunit 1 (VKORC1) affects gene expression and warfarin dose requirement. *Blood.* 112: 1013-1021 (2008).

Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D, Birney E. Factorbook. org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Research. 2013 Jan 1;41(D1):D171-6.

Wang LS, Shang JJ et al. Influence of ORM1 polymorphisms on the maintenance stable warfarin dosage. Eur J Clin Pharmacol, May 2013, 69(5):1113-20.

Wang S, Su JH, Beliveau BJ, Bintu B, Moffitt JR, Wu CT, Zhuang X. Spatial organization of chromatin domains and compartments in single chromosomes. Science. 2016 Aug 5;353(6299):598-602.

Wang SB, Feng JY et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Scientific Reports volume 6, Article number: 19444 (2016)

Wang Z, Jacobs KB, Yeager M et al. Improved imputation of common and uncommon SNPs with a new reference set. Nat. Genet. 44, 6–7 (2012).

Ward LD, Kellis M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40 (D1), D930-D934 (2012).

Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators, and target genes for human complex traits and disease. Nucleic Acids Res. January 2016. 4;44(D1):D877-81

Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genet.* 46, 1160–1165 (2014).

Weinshilboum R, Wang L. Pharmacogenomics: bench to bedside. Nature Reviews in Drug Discovery, September 2004. 3:739-748

Weischenfeldt J, Dubash T et al. Pan-cancer analysis of somatic copy number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat Genet. 2017 Jan; 49(1): 65–74. doi: 10.1038/ng.3722

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature.* 447, 661–678 (2007).

Welter D1, MacArthur J, Morales J et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nuc. Acids Res*. 42, D1001–D1006 (2014).

Wen L, Li X et al. Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. Genome Biology 2016. https://doi.org/10.1186/gb-2014-15-3-r49

Wendt KS, Grosveld FG. Transcription in the context of the 3D nucleus. *Current Opin. Genet. Develop*. 25, 62–67 (2014).

Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016 May; 48(5): 488–496. doi: 10.1038/ng.3539

Whirl-Carrillo M, McDonagh EM et al. Pharmacogenomics Knowledge for Personalized Medicine. Clinical Pharmacology & Therapeutics (2012) 92(4): 414-417.

Wiley LK, Vanhouten JP et al. STRATEGIES FOR EQUITABLE PHARMACOGENOMIC-GUIDED WARFARIN DOSING AMONG EUROPEAN AND AFRICAN AMERICAN INDIVIDUALS IN A CLINICAL POPULATION. Pac Symp Biocomput, 2016, 22:545-556.

Willer CJ, Schmidt EM, Sengupta S et al. Discovery and Refinement of Loci Associated with Lipid Levels. *Nat Genet.* Nov. 45(11) doi: 10.1038/ng.2797 (2013)

Williams FM, Carter AM et al. Ischemic stroke is associated with the ABO locus: the EUROCLOT study. Ann Neurol, January 2013. 73(1):16-31.

Williams SP, Athey BD et al. Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length. Biophys J. 1986 Jan;49(1):233-48.

Wojcicki, Anne. Interview with Bloomberg News, April 7, 2017. https://www.bloomberg.com/news/videos/2017-04-07/23andme-wins-fda-approval-to-sell-genetic-tests-video Retrieved October 11, 2017.

Wong ES, Schmitt BM et al. Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. Nature Communicationsvolume 8, Article number: 1092 (2017)

Woodcock CL, Ghosh RP. Chromatin Higher-order Structure and Dynamics. *Cold Spring Harbor Perspect. Biol.* 2/5/a000596. (2010)

Wright D, Salehian O. Brugada-type electrocardiographic changes induced by long-term lithiumuse. Circulation. 122, e418–e419 (2010).

Wu Z, Ramsundar B, Feinberg Evan n et al. MoleculeNet: a benchmark for molecular machine learning. Chemical Science 9(2), 513-530 (2018).

Wutz G, Varnai C et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. EMBO Volume36, Issue24 15 December 2017

Wysokinski A & Kloszewska I. Mechanisms of increased appetite and weight gain induced by psychotropic medications. (2014). J. Advanced Clin. Pharmacol. 1.1, 12-33.

Xia K, Shabalin AA et al. seeQTL: a searchable database for human eQTLs. Bioinformatics, February 2012, 28(3):451-2.

Xiao W, Mindrinos MN, Seok J, Cuschieri J, Cuenca AG, Gao H, Hayden DL, Hennessy L, Moore EE, Minei JP, Bankey PE. A genomic storm in critically injured humans. The Journal of Experimental Medicine. 2011 Dec 19;208(13):2581-90.

Xin L, Liu YH et al. The Era of Multigene Panels Comes? The Clinical Utility of Oncotype DX and MammaPrint. World J Oncol. 2017 Apr; 8(2): 34–40. doi: 10.14740/wjon1019w

Xu M, Ning C, Ting C, Rui J. DeepEnhancer: Predicting enhancers by convolutional neural networks. Presented at: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2016.

Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L. Deep Learning for Drug-Induced Liver Injury. J Chem Inf Model 55(10), 2085-2093 (2015).

Yan, Q. (2015). Circadian Rhythms and Cellular Networks in Cancer. In Cellular Rhythms and Networks (pp. 61-70). Springer International Publishing.

Yang TP, Beazley C, Montgomery SB et al. Genevar: A database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics.* 26(19), 2474-2476 (2010).

Yao L, Tak YG, Berman BP, Farnham PJ. Functional annotation of colon cancer risk SNPs. *Nature Comm.* 5:5114 (2015) doi: 10.1038/ncomms6114

Yao Y, Robinson AM, Metz GAS et al. Ancestral exposure to stress epigenetically programs preterm birth risk and adverse maternal and newborn outcomes. *BMC Medicine* 12:121 (2014).

Yasuda SU, Zhang L, Huang SM. The Role of Ethnicity in Variability in Response to Drugs: Focus on Clinical Pharmacology Studies. Clinical Pharmacology and Therapeutics, July 2008, 84(3):417-423.

Yi J, Zhang L, Tang B, Han W, Zhou Y, Chen Z, Jia D, Jiang H. Sodium valproate alleviates neurodegeneration in SCA3/MJD via suppressing apoptosis and rescuing the hypoacetylation levels of histone H3 and H4. PLoS One. 2013 Jan 28;8(1):e54792.

Yong WS, Hsu FM, Chen PY. Profiling genome-wide DNA methylation. Epigenetics and Chromatin 2016. doi: 10.1186/s13072-016-0075-3

Zanger UM, Schwab M: Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther* 138:103–41. (2013)

Zanger UM, Klein K, Thomas M et al: Genetics, epigenetics and regulation of drug metabolizing cytochrome P450 enzymes. *Clin Pharmacol Ther* 95:258–261. (2014)

Zeng H, Gifford DK. Predicting the impact of non-coding variants on DNA methylation. Nucleic Acids Res 45(11), e99 (2017).

Zhang C, Li Z et al. A study of N-methyl-D-aspartate receptor gene (GRIN2B) variants as predictors of treatment-resistant major depression, Psychopharmacol. 231(4), (2014), 685-693.

Zhang F, Chen JY. HOMER: a human organ-specific molecular electronic repository. BMC Bioinformatics, October 2011, 12 Suppl 10:S4.

Zhang JW, Liu Y et al. Inhibition of human liver cytochrome P450 by star fruit juice. J Pharm Pharm Sci. 2007;10(4):496-503.

Zhang R, Lahens NF, Balance HI et al. A circadian gene expression atlas in mammals: Implications for biology and medicine. *PNAS.* 111, 16219-16224 (2014).

Zhang S, Cao Z. gkm-DNN: efficient prediction using gapped k-mer features and deep neural networks. bioRxiv, July 2017, doi: 10.1101/170761

Zhao C, Eisinger BE, Driessen TM, Gammie SC. Addiction and reward-related genes show altered expression in the postpartum nucleus accumbens. Frontiers in behavioral neuroscience 8 (2014).

Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12(10), 931-934 (2015).

Zhou K, Pearson ER. Insights from genome-wide association studies of drug response. *Annu. Rev. Pharmacol. Toxicol.* 53, 299–310 (2013).

Zhou X, Li1 D, Zhang B et al. Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nature* Biotech. doi:10.1038/nbt.3158 (2015)

Zhou X, Maricque B, Xie M et al. The human epigenome browser at Washington University. *Nature Methods*. 8, 989-990 (2011).

Zhou X, Lowdon RF et al. Exploring long range genome interactions using the WashU Epigenome Browser. Nature Methods, May 2013. 10(5):375-6

Zimmerman T, Rietdorf J, Pepperkok R. Spectral imaging and its applications in live cell microscopy. FEBS Letters Volume 546, Issue 1, 3 July 2003, Pages 87-92. https://doi.org/10.1016/S0014-5793(03)00521-0

Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. PNAS 2011. https://doi.org/10.1073/pnas.1119675109

Zuko A, Kleijer KT, Oguro-Ando A et al. Contactins in the neurobiology of autism. *European journal of pharmacology*, 719(1), 63-74. (2013)