UNIVERSITÀ DEGLI STUDI DI BRESCIA

PH.D. IN

Genetica Molecolare Biotecnologie e Medicina Sperimentale

(Genetica Molecolare Applicata alle Scienze Mediche)

Settore Scientifico e disciplinare

MED/03

Ciclo

XXXIII

TITOLO DELLA DISSERTAZIONE

Towards Precision Psychiatry: A data-driven strategy for prioritizing antidepressant drug prescription based on predicted gene expression and drug-induced expression profiles

NOME DEL DOTTORATO DELL'ANNO FINALE

Muhammad Shoaib

Firma

NOME DEL SUPERVISORE

Prof. Massimo Gennarelli

Firma

NOME DEL COORDINATORE DEL PHD

Prof. Eugenio Monti

Contents

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor Prof. Massimo Gennarelli for giving me the opportunity to work in his research group and for his continuous support, encouragement, and mentorship throughout my Ph.D. I am also grateful to my senior colleagues, Chiara Magri, Edoardo Giocupuzzi, and Alessandra Minelli for advising me with their knowledge and expertise.

Secondly, my warmest thanks to my co-supervisor Prof. Cathryn Lewis for welcoming me to her research team at King's College, London. The ten months visiting period helped me to acquire a new set of skills in the field of statistical genetics. I am also grateful to postdocs Oliver Pain and Chiara Fabbri for their great supervision and their input to the research work presented in this thesis.

My sincere thanks to the external reviewers Francesco Mazzarotto from the University of Florence and Carlo Maj from the University of Bonn for carefully evaluating my Ph.D. thesis.

Finally, special thanks to my parents, my wife, and son for their unconditional support and love. They have always been an immense source of motivation and encouragement during my Ph.D. journey.

Statement of Authorship

Throughout this thesis, the scientific convention of using the pronoun ‘we’ has been adopted when presenting the research methodology and results. However, all work in Ph.D. thesis was performed and written by Muhammad Shoaib, with the following exception:

In Chapter 5, the genotyping data of STAR*D studies were collected, pre-processed and quality controlled by others prior to analysis. All statistical analysis was performed by Muhammad Shoaib and the first draft of the paper was written by Muhammad Shoaib. Subsequently, it was circulated among co-authors and underwent peer review before publication, which led to the editing of the manuscript and inclusion of additional analysis.

Moreover, supervisors and co-authors of the paper included in this thesis offered valuable

advice on the analyses, results interpretation, and editing of the manuscript.

Riassunto

Nella pratica clinica, la terapia con antidepressivi è un approccio per tentativi che richiede tempo per essere messo a punto ed in molti casi questo processo è sconfortante per i pazienti. Da qui nasce l'esigenza di sviluppare strumenti che permettano di indirizzare meglio il clinico nella scelta dei migliori farmaci da utilizzare. In questa tesi sono proposti degli approcci *in silico* per classificare gli antidepressivi in base alla loro ipotetica probabilità di efficacia e sono sviluppati modelli di machine learning per prevedere la risposta antidepressiva in individui affetti da disturbo depressivo maggiore.

Partendo dai risultati di uno studio di associazione genome-wide nella coorte STAR*D (N=1163), abbiamo inizialmente imputato la "signature" del profilo trascrittomico dei pazienti che rispondevano alla terapia con citalopram. Successivamente, utilizzando le correlazioni di Spearman, Pearson e il test di Kolmogorov Smirnov, abbiamo correlato il profilo trascrizionale dei pazienti che rispondevano alla terapia con 21 profili di espressione genica indotti da antidepressivi in cinque linee cellulari umane disponibili nel database delle mappe di connettività (Cmap). Infine, abbiamo ordinato gli antidepressivi in modo decrescente in base alla media degli indici di correlazione ottenuti con i tre diversi metodi e abbiamo calcolato la probabilità di ottenere casualmente tale posizione in classifica mediante permutazione. I farmaci con un grado di correlazione positivo più elevato erano quelli con una probabilità di efficacia maggiore.

In MCF7 (linea cellulare di cancro al seno), il farmaco con il rango medio più elevato è risultato essere l'escitalopram (p = 0,0014). Nelle linee cellulari A375 (melanoma umano) e PC3 (cancro alla prostata), escitalopram e citalopram sono risultati essere i più significativi (p = 0,0310 e 0,0276, rispettivamente). Nelle linee cellulari HA1E (rene) e HT29 (cancro del colon), invece, i profili trascrizionali del citalopram e dell'escitalopram non predicevano la risposta al citalopram.

La correlazione significativa tra i profili di espressione dei pazienti che rispondono al citalopram e i profili d'espressione indotti da citalopram e (es)citalopram in tre linee cellulari suggerisce che il nostro approccio può essere utile e, con futuri miglioramenti, può essere applicabile a livello individuale per personalizzare la prescrizione del trattamento.

Inoltre, abbiamo implementato modelli di regressione logistica e di regressione netta elastica per prevedere la risposta antidepressiva individuale in base alla correlazione tra i profili di espressione individuali imputati nei pazienti della coorte STAR*D e i profili d'espressione dei farmaci disponibili in Cmap. Il metodo di regressione logistica ha identificato negli antidepressivi triciclici i migliori predittori della risposta al citalopram. Il modello di regressione netta lineare invece è riuscito a identificare una correlazione significativa solo in una linea cellulare.

Abstract

In clinical practice, antidepressant prescription is a trial-and-error approach, which is time-consuming and discomforting for patients. This study investigated an in-silico approach for ranking antidepressants based on their hypothetical likelihood of efficacy and machine learning models to predict antidepressant response in individuals suffering from major depressive disorder.

We predicted the transcriptional profile of citalopram remitters by performing an in-silico transcriptome-wide association study on STAR*D genome-wide association study data (N=1163). The transcriptional profile of remitters was compared with 21 antidepressant-induced gene expression profiles in five human cell lines available in the connectivity map database. Spearman correlation, Pearson correlation, and the Kolmogorov Smirnov test were used to determine the similarity between antidepressant-induced profiles and remitter profiles and, subsequently, the average rank of antidepressants across the three methods and a p-value for each rank were calculated using a permutation procedure. The drugs with the top ranks were those having a high positive correlation with the expression profiles of remitters and that may have higher chances of efficacy in the tested patients.

In MCF7 (breast cancer cell line), escitalopram had the highest average rank, with an average rank higher than expected by chance (p=0.0014). In A375 (human melanoma) and PC3 (prostate cancer) cell lines, escitalopram and citalopram emerged as the highest ranked antidepressants, (p=0.0310 and 0.0276, respectively). In HA1E (kidney) and HT29 (colon cancer) cell types, citalopram and escitalopram did not fall among the top antidepressants.

The correlation between citalopram remitters' and (es)citalopram-induced expression profiles in three cell lines suggests that our approach may be useful and with future improvements, it can be applicable at the individual level to tailor treatment prescription.

Furthermore, we implemented logistic regression and elastic net regression models to predict antidepressant response based on the correlation between inferred expression profiles of individuals with major depressive disorder from STAR*D and in-vitro drug profiles from the connectivity map. The logistic regression method suggested tricyclic antidepressants as the most significant predictors associated with the response phenotype. Moreover, when we applied the elastic net regression model to five cell lines, the model performed well only in one cell line (HA1E) to predict drug response in STAR*D participants.

General part

1. Introduction

1.1 Major Depressive Disorder (MDD)

Major Depressive Disorder (MDD), also widely known as depression is a primary health issue and the third leading cause of disability in adolescents and young adults, while being the second leading cause of disability in middle-aged adults on a global scale (James et al. 2018). According to the World Health Organization, more than 264 million people are living with depression worldwide. According to the diagnostic and statistical manual of mental disorder, 5th edition (DSM-5) criteria, the diagnosis of MDD requires observation of at least five or more symptoms in an individual for two weeks (American Psychiatric Associations 2013) (*Table 1.1*).

Table 1.1. Nine core symptoms of MDD according to DSM-5 criteria

1. Depressed Mood*
2. Markedly diminished interest or pleasure in all or almost all activities*
3. Significant weight loss or weight gain, increase or decrease in appetite
4. Insomnia or hypersomnia
5. Psychomotor agitation or retardation
6. Fatigue
7. Feeling of worthlessness or guilt
8. Lack of concentration or indecisiveness
9. Recurrent thoughts of death and suicide

*One of these symptoms must be present for diagnosis.

Based on family, twin, and adoption studies, it has been observed that genetic factors play an important role in MDD. Forty to fifty percent heritability was reported by twin studies whereas family studies suggested a two to three-fold increment in lifetime risk of developing MDD among first-degree relatives (Lohoff 2010). In 2017, the Psychiatric Genomic Consortium (PGC) identified 44 loci associated with MDD after conducting a genome-wide association study (GWAS) of 130,664 MDD cases and 330,470 controls (Wray et al. 2018). In 2019, the meta-analysis of three large GWASs identified 102 independent variants (246,363 cases and 561,190 controls) associated with MDD. Briefly, from GWASs emerged that MDD is a multifactorial and polygenic disorder, where multiple sets of susceptible genes interact with each other and with the environment, predisposing individuals to the development of the illness. MDD is often comorbid with other health conditions such as cardiac disease, diabetes, and obesity (Whooley and Wong 2003), suggesting that MDD could share genetic and environmental factors with other disorders.

Several therapeutic options are available for treating MDD, including psychological treatments such as behavioural activation, cognitive behavioural therapy (CBT), interpersonal psychotherapy (IPT), and pharmacotherapy. Antidepressants (ADs) are usually prescribed for treating moderate and severe MDD cases. There are five major classes of antidepressants: 1) Selective Serotonin reuptake inhibitors (SSRIs), 2) Serotonin and norepinephrine reuptake inhibitors (SNRIs), 3) Noradrenergic and specific serotonergic antidepressants (NASSAs), 4) Tricyclic antidepressants (TCAs), (5) Monoamine oxidase inhibitors (MAOIs). These different classes of drugs work by preventing the reabsorption of neurotransmitters in the brain. SSRIs inhibit the reuptake of serotonin whereas SNRIs inhibit the reuptake of both serotonin and norepinephrine. Further, TCAs work by modulating three neurotransmitter molecules in nerve cells that are serotonin, norepinephrine, and acetylcholine. The MAOIs function by blocking the effect of monoamine oxidase enzyme, increasing the availability of neurotransmitters for mood regulation. Moreover, NASSAs act by antagonizing alpha-adrenergic receptors and certain serotonin receptors which results in the enhancement of adrenergic and serotonergic neurotransmission in the brain (Fasipe 2018). Antidepressant choice in MDD is based on prescription guidelines and prior clinical experience, but the lack of reproducible predictors of AD response makes it a ‘trial and error’ approach which can take up to several weeks or months and a number of treatment changes before symptom remission is achieved. More than 60% of patients fail to achieve remission after being treated with the first AD. Several studies have demonstrated that AD response is a trait with a genetic component, indeed AD response frequently clusters in families (O’Reilly, Bogue, and Singh 1994) (Franchini et al. 1998) and common genetic variants were estimated to explain 42% of the variance in AD response (Tansey et al. 2013a). Due to the heterogeneous and polygenic attribute of AD response, researchers are employing Big Data, GWAS, and multi-markers approaches to study the AD response trait among MDD patients (Musker and Wong 2019). The lack of reproducible biomarkers predicting AD response and limited knowledge of clinical improvement are primary challenges of depression treatment (Labermaier, Masana, and Müller 2013).

1.2 The Pharmacogenetics of Antidepressants

Based on the above rationale, pharmacogenetics represents a key contributor to the implementation of precision medicine.

The term pharmacogenetics has been in use since 1959. Pharmacogenetics was first referred to as the relationship between phenotypic variation in metabolism and response to certain drugs. Then in the 1980s, thanks to the scientific advances in human genetics, the genetic basis of this phenotypic variation become clearer and pharmacogenetics becomes the study of how genetic variants affect a person's response to drugs. At the end of the 1990s, with the advent of the Human Genome Project, the term pharmacogenomics started to be used in addition to pharmacogenetics. Both terms are now used interchangeably in the literature (PharmGKB, 2017a).

In pharmacogenetics (PGx), genomic information is used to study the drug response among individuals and to develop effective, safe medications and define doses that will be tailored to a person's genetic makeup (Kisor, Hoefer, and Decker 2019). Single nucleotide polymorphisms (SNPs), deletions, insertions, and short tandem repeats are different types of genetic variations that might have a role in drug response and can be utilized as predictive markers to evaluate treatment response in a patient (Kisor, Hoefer, and Decker 2019).

As far as pharmacogenomics of AD is concerned, many studies focalized on genes implicated in the pharmacokinetics or pharmacodynamics of AD, and more recent hypothesis-free approaches have identified novel candidates for AD response. However, results are not always concordant, and the modest effects observed have confirmed the polygenicity of the trait and the involvement of multiple genetic variants of small effect.

In the following sections, an overview of the main pharmacogenomic studies of AD drug response will be provided.

1.2.1 Candidate gene studies

To understand the PGx of AD response, candidate gene studies pointed out several genes that may influence drug response. This approach has focused mainly on two classes of genes. The first class includes those genes that encode proteins involved in the Pharmacokinetics of drugs. The mechanisms, such as drug absorption, metabolism, distribution, and elimination which have an impact on the delivery of drug to the target site, are regulated by pharmacokinetic genes. The cytochrome P450 (CYP) gene family is a category of enzymes with a substantial role in the oxidation and reduction of endogenous and xenobiotic substances. This gene family includes *CYP2D6*, *CYP2C19*, *CYP2C9*, *CYP3A4*, and *CYP1A2* genes and they are important in the metabolism of various ADs (Gaedigk et al. 2018). The isoenzymes responsible for AD metabolism and considered determinant in AD clinical outcome are *CYP2D6* and *CYP2C19*. The genes coding for these enzymes are highly polymorphic and the different alleles encode for an enzyme with normal, partially or totally defective activity or increased activity. Based on the allelic combinations of *CYP2D6* and *CYP2C19* genes and their effect on enzymatic activity, individuals are categorized into poor metabolizers (PMs), intermediate metabolizers (IMs), extensive metabolizers (EMs), and ultra-rapid metabolizers (UMs) (Nassan et al. 2016). Individuals carrying two defective alleles are PMs whereas UMs carry two alleles with an increased activity or gene duplications. EMs individuals instead have two functional alleles; therefore, they have normal enzymatic functions and drug metabolism. IMs have one defective allele, hence they may have slower drug metabolism (Corponi 2019). Prior studies suggested a relationship between *CYP2D6* variants and the concentration of antidepressants in the blood plasma. According to the reported evidence, individuals carrying PMs variant are at higher risk of toxic reactions, while UMs require a higher drug dosage to achieve a therapeutic level of drug concentration in the blood plasma (Hicks et al. 2017). Despite various studies that have advanced our understanding of *CYP* genes and their role in ADs metabolism, we still do not have strong evidence linking these genes to the clinical outcome for commonly used ADs. For instance, for selective SSRIs, there was not a strong correlation between *CYP* genes and the clinical outcome, which suggests metabolizer status dependent therapies do not have a significant clinical impact (Fabbri et al. 2018). The second class of genes considered by candidate genes studies are pharmacodynamics target genes which are directly affected by the drugs. As we have a limited understanding of ADs pharmacodynamics, the receptor and target binding sites of ADs and their complete mechanism of actions are unknown. The selection of relevant pharmacodynamics genes thus poses a challenge. Since it has been hypothesized that the monoaminergic system is involved in the pathophysiology of MDD, genes encoding monoamine neurotransmitters have been investigated for the ADs response (Fabbri, Di Girolamo, and Serretti 2013). One notable example in this context is the *SLC6A4* gene which has extensively been studied (Licinio and Wong 2011). The allelic differences of *SLC6A4* modulate the expression of serotonin transporter protein which results in varied serotonin uptake. Studies based on meta-analysis suggest that the long L allele of serotonin transporter-linked promoter region (5-HTTLPR) predicts better SSRI response in the Caucasian population while it was found to have an opposite effect in the Asian population (Fabbri, Di Girolamo, and Serretti 2013). Additionally, variants of other pharmacodynamic genes (*HTR1A*, *HTR2A*, *COMT*, *GNB3*, *CNR1*, *NPY*, *MAOA*, *FKBP5*, and *BDNF*) has also been previously investigated (A. Serretti and Artioli 2004). Because of the involvement of the hypothalamic-pituitary-adrenal (HPA) axis in MDD pathophysiology, *FKBP5* and *NR3C1* are important candidates for the heterogeneous behaviour of AD response. Both genes play a significant role in glucocorticoid pathway, hence, they are promising pharmacodynamic candidate genes. Variants of the *FKBP5* gene have been found to be associated with differential therapeutic response. MDD patients who were homozygous for the T allele of the rs1360780 SNP in the *FKBP5* gene responded faster to SSRIs, TCAs, and mirtazapine compared with the carriers of C alleles. However, these results are still preliminary and

need to be replicated in other samples (A. Serretti, Drago, and Liebman 2009)(Chiara Fabbri et al, 2014). Another interesting candidate is G (guanine nucleotide-binding) protein beta 3 subunit gene (*GNB3*). Variants of this gene were found to be related with better AD response in multiple investigations. (A. Serretti, Drago, and Liebman 2009). Another candidate gene is the corticotrophin releasing hormone (*CRH*) receptor 1 gene. Researchers detected fluoxetine therapeutic response associated with the variants of this gene. Also, neurotrophic factors are considered as the promising candidates for pharmacogenomic investigations (A. Serretti, Drago, and Liebman 2009).
Despite valuable contributions made by candidate gene studies, the findings reported are still inconsistent and could not be further replicated. Since, the availability of high-throughput technologies has allowed researchers to conveniently perform a genome-wide analyses. Therefore, PGx studies of AD response have now shifted from the candidate gene to GWAS.

1.2.2 Role of GWAS in elucidating the architecture of antidepressant treatment response.

GWAS is a hypothesis-free technique to identify SNPs associated with the phenotypes/traits of interest without any prior knowledge of causal variants.

GWAS have contributed significantly to elucidating the etiology of complex polygenic psychiatric diseases. After attaining sufficient sample size and power, GWAS can be used as a powerful tool in the identification of genetic variants associated with a particular trait or a disease. In contrast to candidate gene studies, GWAS is a more pertinent approach to disentangle the complexity of polygenic conditions, such as AD response, where the mechanisms of action are not fully elucidated (Breen et al. 2016). This technique is ideal for studying the genetic component of non-mendelian conditions that are likely determined by a mixture of environmental and genetic determinants, mostly common and with small effect sizes.

Data from three large trials have been often used in GWASs to detect genomic regions associated with AD response: the Sequence treatment alternative to relieve depression (STAR*D) study (n = 1948) (Garriock et al. 2010), the Genome-based Therapeutic Drugs for Depression (GENDEP) project (n = 706) (Uher et al. 2010), and the Munich Antidepressant Response Signature (MARS) project (n = 339) (Ising et al. 2009). When analyzed in isolation, none of these cohorts led to the identification of genome wide significance associations except for the GENDEP analysis of patient subset treated with Nortriptyline with a finding which achieved GWAS significance threshold as mentioned in the table 1.1. Even two large meta-GWASs of the above data were unable to identify genome-wide significant variants. In the first study, data from GENDEP, MARS, and the STAR*D were meta-analyzed for a total of 2256 MDD cases (Uher et al. 2013). The second study was performed on 2897 MDD cases which included the data from NEWMEDS (Novel Methods Leading to New Medications in Depression and Schizophrenia) and STAR*D (Tansey et al. 2012). When the additional analysis in the first meta-analysis was restrained to citalopram and escitalopram, an intergenic variant rs10783282 in the chromosome region (5q.15.1) was identified. The largest GWAS on AD to date was performed using the clinical and genomic information of the 23 and me cohort. In this GWAS, Li et al compared 1311 treatment-resistant to 7795 responder patients and found no variants reaching the genome-wide significance threshold ($P > 5 \times 10^{-8}$). For another GWAS cohort of the same study of bupropion responders (*n*=2675) vs non-responders (*n*=1861), they found one variant rs1908557 in region (4q22.1) associated with bupropion response (Li et al. 2016) as mentioned in table 1.1. Recently, Wigmore and colleagues tested the association between genetic variants and AD resistance using prescription data in the population and family-based GENDEP cohort, and, however, failed to identify SNPs reaching genome-wide significance. The most significant SNP identified was an intergenic variant located at 10p26.13 (lead SNP rs188352979, $P = 3.25 \times 10^{-7}$, OR = 2.87, CI = 2.47–3.28) (Wigmore et al. 2020). The main GWAS on AD response are summarized in Table 1.1.

Table 1.1 Summary of published GWAS of antidepressant response.

Author/ date	**Study name**	**Sample size**	**AD**	**Top SNP**	**P-value**
Ising et al. (2009)	MARS	N=339	Various ADs	rs6989467	7.6×10^{-7}
Uher et al. (2010)	GENDEP	N=706	Notriptyline	rs2500535	3.6×10^{-8}
Garriock et al. (2010)	STAR*D	N=1491	Citalopram	rs6966038	1.6×10^{-7}
Tansey et al. (2012)	NEWMEDS includes GENDEP sample	N =1790	SRIs	rs10783282	1.1×10^{-6}
Li et al. (2016)	23 and me	N= 4536	Bupropion	rs1908557	2.6×10^{-8}
Wigmore et al. (2020)	GENDEP, GS:SFHS	N=4213	Various ADs	rs188352979	3.2×10^{-7}

While the candidate gene and GWAS approach have provided useful insights for studying various phenotypes, these techniques have not yet established replicated genetic variants relevant to the AD treatment response with clinical significance. Probably, one possible explanation is that these methods, focusing on single SNPs, and did not model the polygenic nature of AD response. For this reason, scientists are exploring other analytical methods allowing the integration of multiple SNPs (Lin et al. 2018).

1.2.3 The polygenicity of AD response: the multi-marker approaches

In a polygenic disorder, a single variant is not informative for assessing the risk of disease. Hence individual's genetic loading for a trait or a disease can be calculated using a statistical technique termed as 'polygenic risk score' (PRS). The PRS is a single value estimate of an individual's genetic liability to a phenotype and this can be calculated as a sum of risk alleles carried by an individual, weighted by the corresponding allele effect sizes derived from GWAS summary statistics (Lewis and Vassos 2020).

The PRS analysis was tested for AD treatment response trait in GENDEP (n=736) and STAR*D cohort (n=1409), but no significant associations were found (García-González et al. 2017). Similarly, another study used PRS for MDD and neuroticism as predictors of antidepressant response within 3 treatment sub-cohorts from GENDEP and 2 sub-cohorts from the Pharmacogenomics Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS). This study couldn't significantly predict antidepressant response using PRS of MDD and neuroticism. However, the investigators reported that higher genetic loading for both phenotypes was associated with less favourable drug response (Ward et al. 2018). Moreover, previous studies have reported C-reactive protein (CRP) as a marker of inflammation and its association with antidepressant response. Zwicker et al analysed data from GENDEP studies and calculated PRS for CRP level-based genome-wide results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. The researchers reported that a higher PRS for CRP protein was associated with a better response to escitalopram and worse response to amitriptyline (Zwicker et al. 2018)

1.2.4 Translating GWAS findings into effective therapeutics: moving from SNPs to transcriptomic profiles

Previous studies have investigated possible approaches of translating GWAS findings into effective therapeutic options. Researchers studied whether, for example, top GWAS variants can serve as drug targets (Sanseau et al*, 2012*). There are a number of limitations of considering top GWAS variants only, as most of the time they lie in the non-coding region of the genome and do not encode for drug-targeted proteins. Moreover, there are chances of missing multitarget drugs. Based on PRS analysis, many complex traits are influenced by SNPs with small effect sizes, and prior studies have ignored them and considered only the top significant genes for evaluating drug target proteins. Keeping in mind these limitations, imputing gene expression signatures from GWAS summary statistics may be a refined approach and can be used for comparing traits influenced by transcriptomic changes with drug-induced gene expression patterns. Analysis of *in-vitro* transcriptional profiles of drugs (from reference databases) and disease signatures (from GWAS studies) is already an established approach in the domain of drug repositioning. For instance, Sirota and colleagues found that cimetidine showed an opposite expression pattern to that associated with lung adenocarcinoma. Cimetidine causes genes that were highly expressed in lung adenocarcinoma to be lowly expressed and vice versa. Researchers experimentally validated this drug as a potential treatment (Sirota et al. 2011). Similarly, topiramate was found as a possible treatment for inflammatory bowel disease and this hypothesis was validated in an animal model (Dudley et al. 2011). So and his colleagues have proposed a drug repurposing strategy for various psychiatric disorders based on the GWAS summary statistics and imputed gene expression profiles corresponding to psychiatric traits. They found a number of repositioning candidates for psychiatric conditions while many of them were also supported by clinical and pre-clinical evidences (So et al. 2017).

2. Statistical methods behind pharmacogenomics studies: GWAS, PRS, and TWAS

2.1 Genome-Wide Association Studies (GWAS)

One of the major focus of human genetics is to identify genetic factors responsible for common and rare Mendelian diseases. Besides elucidating the complexity of common and rare diseases, GWAS has successful applications in the domain of pharmacogenetics and personalized medicine (Cooper et al. 2008). The purpose of GWAS is to scan genetic variants across genomes of many people to find those allelic variants or genotypes, which are associated with a disease or phenotype, that is those variants that are observed more frequently than expected by chance in subjects with the phenotype under study. After identifying causal variants, researchers can use this information to develop better methodologies to detect, treat, and prevent the disease. With the completion of the Human genome and International Hapmap projects, these sources of information can be used as important research tools for finding the genetic causes of diseases (Bush and Moore 2012). One of the first major successes of GWAS was the identification of the *Complement factor H* gene as a risk factor towards the onset of age-related macular generation (Haines and Hauser MA 2005).

In this section, the key steps for conducting GWAS, and the statistical tools for data analysis will be presented. As mentioned earlier, the primary goal of GWAS is to identify SNPs that are responsible for phenotypic variations of complex human traits. Using chip array technology, hundreds of thousands of SNPs are typed and tested across a large number of individuals to find their correlation with the phenotype of interest (O'dushlaine et al. 2015). Broadly, GWAS are developed following four main sequential steps: quality control (QC) of genotyped data, imputation, association testing and interpretation of results.

2.1.1 QC of genotyped and sample data

One of the first QC steps is generally represented by the removal of those variant sites with low calling rate in the analysed dataset. Broadly speaking, some of the mostly used indicators that are representative of data quality are missing call rate (MCR), minor allele frequency (MAF), and Hardy Weinberg equilibrium (HWE) (Pongpanich *et al. 2010*). Large deviations from HWE could indicate genotyping errors and SNPs with greater missingness rates reflects bad genotype probe performance. Since many calling algorithms perform poorly with minor alleles, therefore, SNPs with lower MAF are more prone to genotyping errors. To map those SNPs which are missed by the sequencing method can be imputed by considering external sources such as HapMap and 1000 genome projects (1000

Genome project consortium 2015) (1000 Genome project consortium 2005). The imputations of unmeasured SNPs are carried out by considering LD structure and haplotype frequencies. Association testing for both genotyped and imputed SNPs are performed separately because of the uncertainty involved in the imputation process. However, they can also be performed together possibly by taking into account uncertainties (e.g., dosages). Other indicators include sex discrepancy, heterozygosity, and relatedness among samples. Sex discrepancy indicates sample mix-ups and needs to be addressed during the QC step. X chromosome homozygosity estimate should be greater than 0.8 in males, whereas, in females, it should be less than 0.2. Further, it is important to remove individuals with high or low heterozygosity rates as it indicates sample contamination and inbreeding. Another important QC step is to check relatedness among individuals and to calculate identity by descent (IBD) of all sample pairs in the analysis. Individuals with relatedness above a certain threshold need to be removed as it will affect the results of the association analysis (Marees et al. 2018). However, there are also statistical association test which are robust in terms of relatedness across individuals used in GWAS such as SAIGE (Scalable and Accurate Implementation of Generalized mixed model) (Zhou et al. 2018)

2.1.2 Imputation of the data

Imputation in genetics refers to the statistical inference of unobserved genotypes (Figure 2.1).

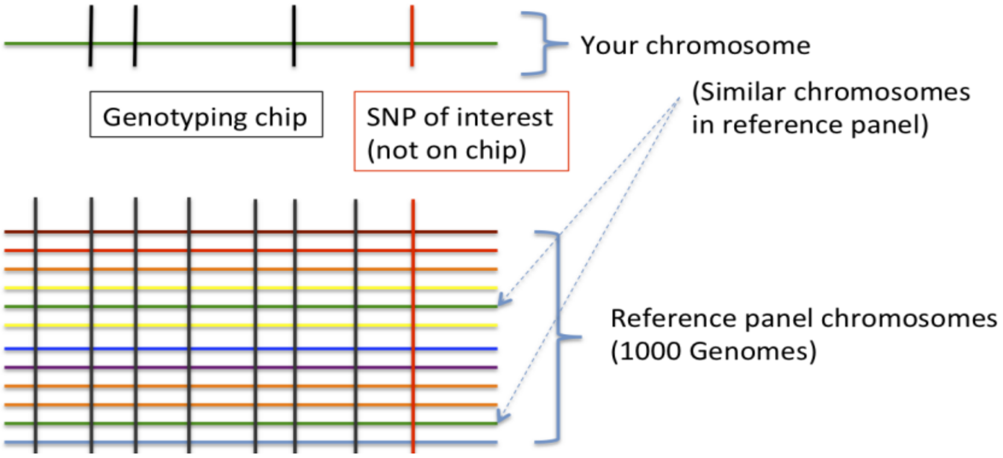

Your chromosome
Genotyping chip
SNP of interest (not on chip)
(Similar chromosomes in reference panel)
Reference panel chromosomes (1000 Genomes)

Figure 2.1 Imputation scheme. Imputation is achieved by using known haplotypes in a population, for instance from the HapMap or the 1000 Genomes Project in humans, thereby allowing to test initially-untyped genetic variants for association with a trait of interest. Genotype imputation hence helps tremendously in narrowing-down the location of probably causal variants in genome-wide association studies.

The number of markers in the association studies can be increased by performing the imputation process that infers the missing genotypes and harmonize the datasets for metanalysis used in association testing. The untyped markers in the sample of interest are imputed using the LD structure among the markers which are evaluated in an external reference panel with a dense genetic map (Porcu et al. 2013). For most of the GWASs, the HapMap reference panel has previously been used but now it has been replaced by a 1000 genome reference set with almost 39.7M biallelic variants (1000 Genome project consortium 2015). There are different software available for the genotype imputation such as Plink, Beagle, MaCH/minimac, fastPHASE and IMPUTE/IMPUTE2 (Browning and Browning 2008) (B. Howie et al. 2012) (Howie, Donnelly, and Marchini 2009) (Yun et al. 2009) (Purcell et al. 2007). All of them are based on various algorithms and offers different limitations and accuracy. For balancing the computational cost of large reference panels, the developers of IMPUTE2 and MaCH software introduced 2 steps procedure for carrying out imputation. In the first step, the genotypes of GWAS individuals are phased and their most likely haplotypes are estimated. Subsequently, in the next step, the genotypes of the reference panel are imputed into the phased genotypes of the GWAS sample (Fuchsberger, Abecasis, and Hinds 2015).

2.1.3 Association testing

After the QC and the imputation steps, the statistical analysis of genetic data is performed by testing each SNP for its independent association with the trait of interest. The single-locus statistical tests used are different depending on the quantitative or binary nature of the analysed traits.

The most common tests used to analyse binary traits are the chi-squared test and Fisher exact test, whereas ANOVA and t-test are applied to test the association of single SNP to quantitative variables. When potential confounding variables need to be controlled for, such as age, gender, medication, or population stratification, the generalized linear model (GLM) can be applied.

GLM is a commonly used family of statistical methods to relate several continuous and/or categorical predictors to a single outcome variable. Analysis of variance (ANOVA), which is similar to the linear regression method, is implemented for the analysis of continuous variables (Wang et al. 2019).

a) Linear Regression (GLM)

The assumptions made by GLM are, the residuals are normally distributed, each group has the same trait variance, and the groups are independent.

Given an input vector $X^T = (X_1, X_2, \ldots\ldots\ldots\ldots\ldots X_P)$, a linear regression model can be expressed as:

$$\boldsymbol{f(x) = \beta_o + \sum_{j=1}^{p} X_j \beta_j + \varepsilon_i} \tag{1}$$

with the error terms $\varepsilon_i \sim N(0, \sigma 2)$.

The regression coefficients (β_j) are estimated by minimizing the residual sum of squares (RSS) which is the sum of the squared difference between actual y_i and predicted output variable *f(x):*

$$\boldsymbol{RSS(\beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2} \tag{2}$$

$$= \boldsymbol{\sum_{i=1}^{n} \left(y_i - \beta_o - \sum_{j=1}^{p} X_{ij} \beta_j\right)^2} \tag{3}$$

In a linear regression model, the β values are acquired using maximum likelihood estimation which gives the parameter values that maximize the likelihood of observing the outcome variable y. In practice, t statistics and type III f-tests are used to check the significance of parameter estimates.

When potential confounding variables such as age, sex, and medication need to be controlled for, the extended form of GLM can be written as:

$$\boldsymbol{g(\mu) = \sum_j \beta_j X_j + uG + \sum_k Y_k PC_k} \tag{4}$$

where μ= E(Y); the g() is the link function that performs a monotone transformation on the mean of the response variable, X_j is the jth covariate representing a clinical or environmental risk factor, β_j is the regression coefficient of X_j, G is the genotype of the test SNP with coefficient u, PC_k is the kth top principal component calculated from the genotype matrix, and Y_k is the effect of PC_k. If phenotype Y is a binary variable, then a logit link function can be applied and the GLM reduces to logistic regression (Wang, Cordell, and Van Steen 2019).

b. Logistic Regression

Binary or dichotomous trait data are analysed using contingency tables or logistic regression methods. Logistic regression is an extension of linear regression in which the linear model is transformed using a logistic function. This prediction method is suitable for categorical output variables and calculates the probability of binomial traits, hence performs classification (Peng, Lee, and Ingersoll 2002).

For a binary outcome with two output possibilities A and B, the log odds of class A, as opposed to class B, can be modelled by the following expression:

$$\log\left(\frac{P\left(y_i = A \middle| x_{i1}, x_{i2}, \dots, x_{ip}\right)}{P\left(y_i = B \middle| x_{i1}, x_{i2}, \dots, x_{ip}\right)}\right) = \beta_O + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \qquad (5)$$

As $P\left(y_i = B \middle| x_{i1}, x_{i2}, \dots, x_{ip}\right)$ is $1 - P\left(y_i = A \middle| x_{i1}, x_{i2}, \dots, x_{ip}\right)$, the probability of outcome categorized into class A is given by:

$$P\left(y_i = A \middle| x_{i1}, x_{i2}, \dots, x_{ip}\right) = \frac{e^{\beta_O + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_O + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}} \qquad (6)$$

The above equation shows that the model outputs probabilities in the range between 0 and 1. Due to the linear effects of predictor variables on the outcome function, logistic regression models can be categorized into the class of GLMs.

Results obtained after conducting GWAS are represented in a tabular format known as summary statistics. Summary statistics include for each analysed SNPs: information on chromosome position, SNP identifier, MAF, effect size (odds ratio/beta), standard error, and p-value. Summary statistics files are available for different phenotypes and are made public and stored in dedicated databases. The most famous repository of GWAS summary statistics is the GWAS catalogue maintained by the European Bioinformatics Institute (https://www.ebi.ac.uk/gwas/)

2.1.4 Interpretation of GWAS results

Association testing will generate a test statistic for each SNP, measuring its association with the trait being studied and a p-value reflective of statistical significance. Manhattan and quantile-quantile (QQ) plots (Fig. 2.2) are standard graphical tools for visualizing GWAS results (Jiang and Wang 2018). A Manhattan plot is a scatter plot that displays the level of significance of each SNP based on its chromosomal location, displayed on the x-axis. On the y axis, the negative log-base-10 of the p-value for each of the SNP being tested is presented. This way, stronger association signals are characterized by higher values on the graph and are visible at the top. Another widely used plot for the graphical representation of GWAS results is a QQ plot, in which the x-axis displays the expected distribution of association test statistics under the null hypothesis of no association across millions of SNPs. The expected p-values are compared with the observed p-values which are along the y-axis of the plot.

Practically, the majority of SNPs won't be associated with the trait analysed and, therefore, a large number of p-values will lie on the diagonal. A small deflection of observed p-values from the null hypothesis line at the end represents the true associations among thousands of SNPs not associated with the trait (Fig. 2.2 b) (Jiang and Wang 2018). Genomic inflation factor λ is used to quantify if the genetic signal is inflated or not. λ is defined as the median of the resulting chi-squared test statistics divided by the expected median of the chi-squared distribution. It can be calculated from z-scores, chi square statistics or p-values.

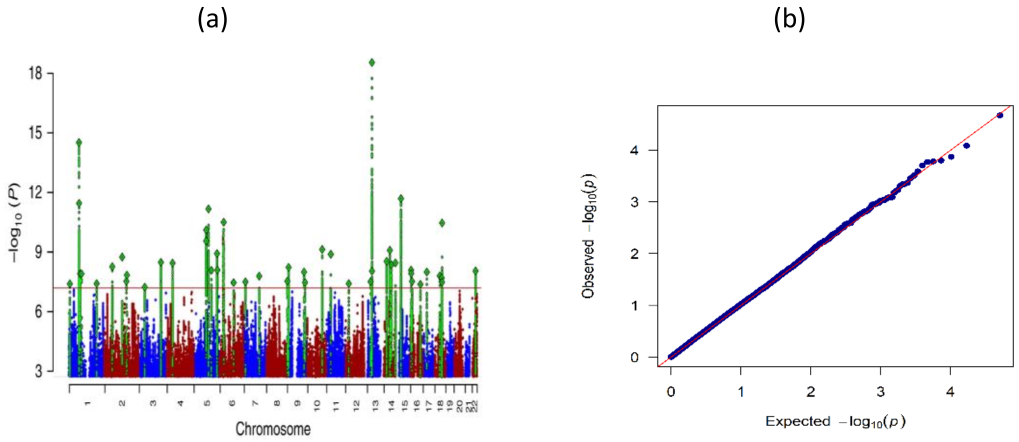


Fig 2.2. (a) Manhattan plot from the meta-analysis of 135,458 MDD cases and 399,401 controls (Wray et al. 2018). *(b) QQ-plot from GWAS comparing STAR*D first stage participants (citalopram remitters and non-remitters).*

<u>*a. The problem of multiple test corrections*</u>

The null hypothesis is rejected, and a statistical test is considered significant if the p-value is lower than the predefined threshold $\alpha = 0.05$. This implies that 5% of the time, the null hypothesis is rejected even if is true. This probability test is applicable to the single statistical test. As we usually consider millions of SNPs in GWAS that is why the aggregated likelihood of getting false positives is much higher (Jiang and Wang 2018).

One of the simplest methods to overcome the issue of multiple testing is the '*Bonferroni correction*' method which adjusts the value of α from 0.05 to α/k where k is the number of statistical tests conducted (Hochberg 1988). The Bonferroni correction at level 0.05 yields the significance threshold 5×10^{-8}, known as the 'genome-wide significance level'. This test is conservative as it assumes all SNPs are independent of each other and does not consider linkage disequilibrium among GWAS markers.

Alternative methods to solve the issue of multiple testing are 'False discovery rate (FDR)' and 'Permutation procedure'.

The FDR test developed by Benjamini and Hochberg determines the false positives among the tests declared as significant, hence, produces fewer false positives (Benjamani and Hochberg 2016). The permutation procedure is another approach for controlling false positives. This method is computationally intensive but straightforward. It generates an empirical null distribution of test statistics by shuffling the genotype-phenotype relationship of each individual in a dataset and this process is repeated for a predefined number of times to approximate false discovery rate (Martin, Westfall, and Young 1994).

2.2 Polygenic Risk Score Analysis

GWAS of a disease tests one SNP at a time, thus, explains a small fraction of disease heritability (Wray et al. 2014). In Common diseases with a polygenic architecture, many genetic variants with small effect sizes act together to increase the risk of a disorder. For this reason, statistical methods (such as polygenic score risk analysis) that analyse the joint effect of many SNPs have been developed. The goal of PRS is to consider thousands of SNPs that could not achieve the GWAS significance threshold but may have a possible role in the disease etiology and may account for the greater amount of heritability. The PRS method utilizes association statistics from discovery GWAS and, after SNPs pruning based on the LD structure and weighing them according to their effect sizes, PRS tests the combined predictive ability in an independent sample (Figure 2.3). P-value informed clumping method is used in LD pruning in which SNPs with the strongest evidence of association within the LD window are retained while the rest of them are discarded.

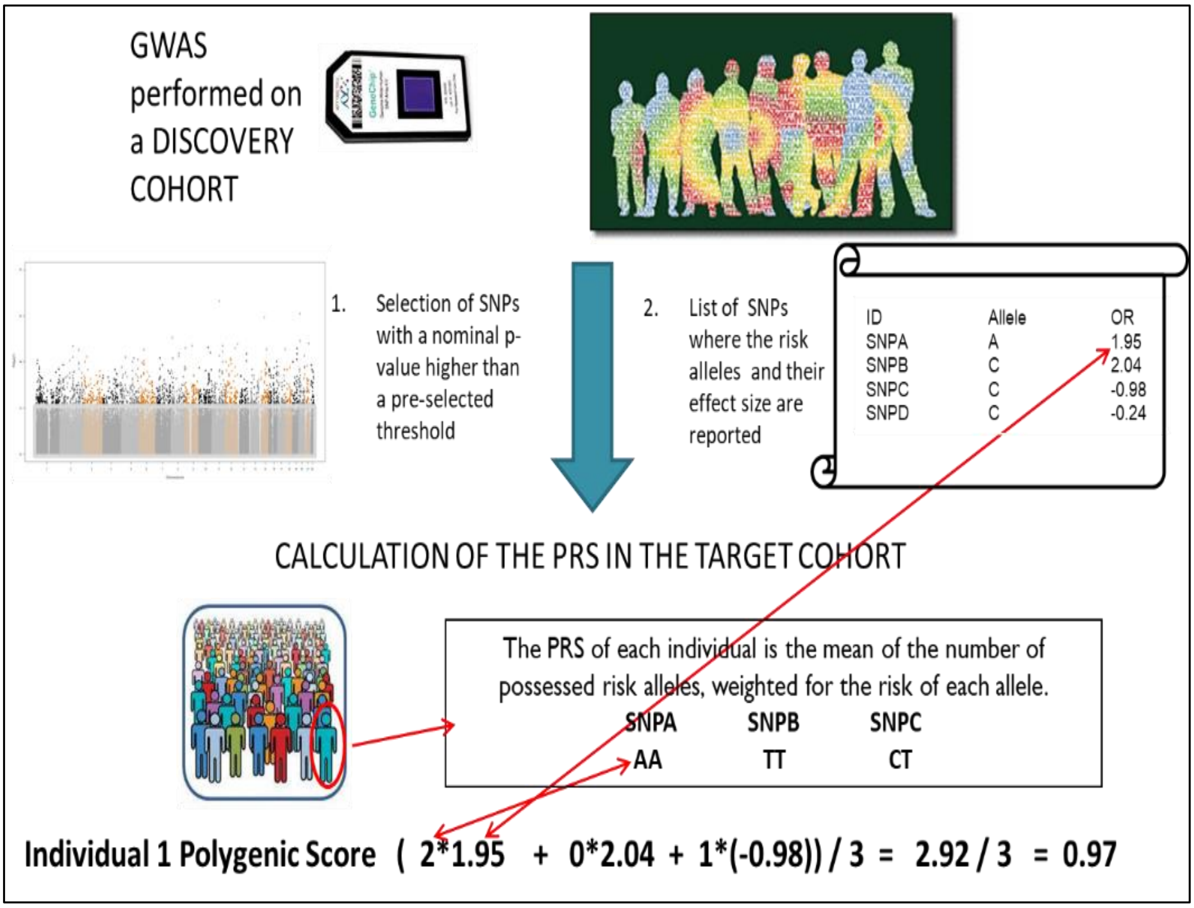

GWAS performed on a DISCOVERY COHORT
1. Selection of SNPs with a nominal p-value higher than a pre-selected threshold
2. List of SNPs where the risk alleles and their effect size are reported
ID Allele OR
SNPA A 1.95
SNPB C 2.04
SNPC C -0.98
SNPD C -0.24
CALCULATION OF THE PRS IN THE TARGET COHORT
The PRS of each individual is the mean of the number of possessed risk alleles, weighted for the risk of each allele.
SNPA SNPB SNPC
AA TT CT
Individual 1 Polygenic Score (2*1.95 + 0*2.04 + 1*(-0.98)) / 3 = 2.92 / 3 = 0.97

Figure 2.3. To calculate the PRS we need two cohorts: a discovery cohort where we identify the risk alleles and their effect size and a target cohort where the PRS is effectively calculated. In the discovery cohort after performing GWAS, we obtain a list of SNPs with a p-value. The SNPs are ranked according to their p-value in ascending order, and only risk alleles with a p-value higher than a pre-selected threshold will be selected. We obtain in this way a list of risk SNPs with the risk allele and its effect size. This information will be used to calculate the PRS in the target cohort. For each individual, the PRS is calculated as the mean number of risk alleles weighted for the risk of each allele.

For conducting PRS analysis, trait-specific weights (log of the odds ratios for binary traits and beta values for continuous traits) are acquired from a discovery GWAS (Marees et al. 2018). The genotype of individuals in an independent validation sample is weighted based on allele effect sizes from the discovery GWAS. Further, these effects are summed across multiple SNPs and represent a PRS value.

PRS in the form of the equation can be given as:

$$PRS_j = \sum_{j=1}^{m} X_j \beta_j \tag{7}$$

Where each individual's score, PRS_j, is calculated by the sum of an individual's risk alleles X_j weighted by risk alleles effect sizes, βj, derived from GWAS summary statistics across m SNPs.

All common SNPs can be used in PRS analysis, but it is important to first clump the SNPs from GWAS results before calculating the risk scores. Theoretically, for a polygenic trait all SNPs are informative. P-value thresholding is done because the best fitting model can be more oligogenic or more polygenic and this is not known in advance. Basically, a model optimization is performed. It is a common practice among researchers to perform multiple PRS analyses each accounting for varying p-value thresholds. The PRS is tested for association with a disease or control status using the logistic regression method by considering principal components as covariates. The prediction accuracy of PRS in the target sample is given in terms of pseudo R^2 for logistic regression (Marees et al. 2018).

The widely used software package for performing PRS analysis is PRS-ice (Euesden, Lewis, and O'Reilly 2015) as it has built-in options for clumping, p-value thresholds, principal components consideration, and plotting of attractive graphs. There are also other methods and tools available for computing PRS. For example lasso regression (LASSOSUM) and Bayesian approaches (LDPred) (Kulm, Mezey, and Elemento 2020)

2.3 Transcriptome-wide association study (TWAS)

One possible mechanism that a genetic variant may influence the associated trait is through regulating the gene expression of its neighbour genes. A method developed to investigate such a potential mechanism is the transcriptome-wide association study (TWAS). In these studies, rather than directly testing SNPs for association with phenotype, the SNPs are used to predict gene expression levels, and the predicted gene expression measures are then tested for association with the phenotype.

TWAS utilizes reference panels, such as the Genotype tissue expression portal (GTEx) datasets (GTEx Consortium 2013), where both SNP genotypes and gene expression profiles have been measured in a variety of relevant tissues to develop prediction models for gene expression. These matrices are then used to impute the expression profile of a target dataset based only on genotype information. In the final step, statistical associations between genetic variants and the trait under investigation are estimated. The three main steps of TWAS are depicted in Figure 2.4 and described below.

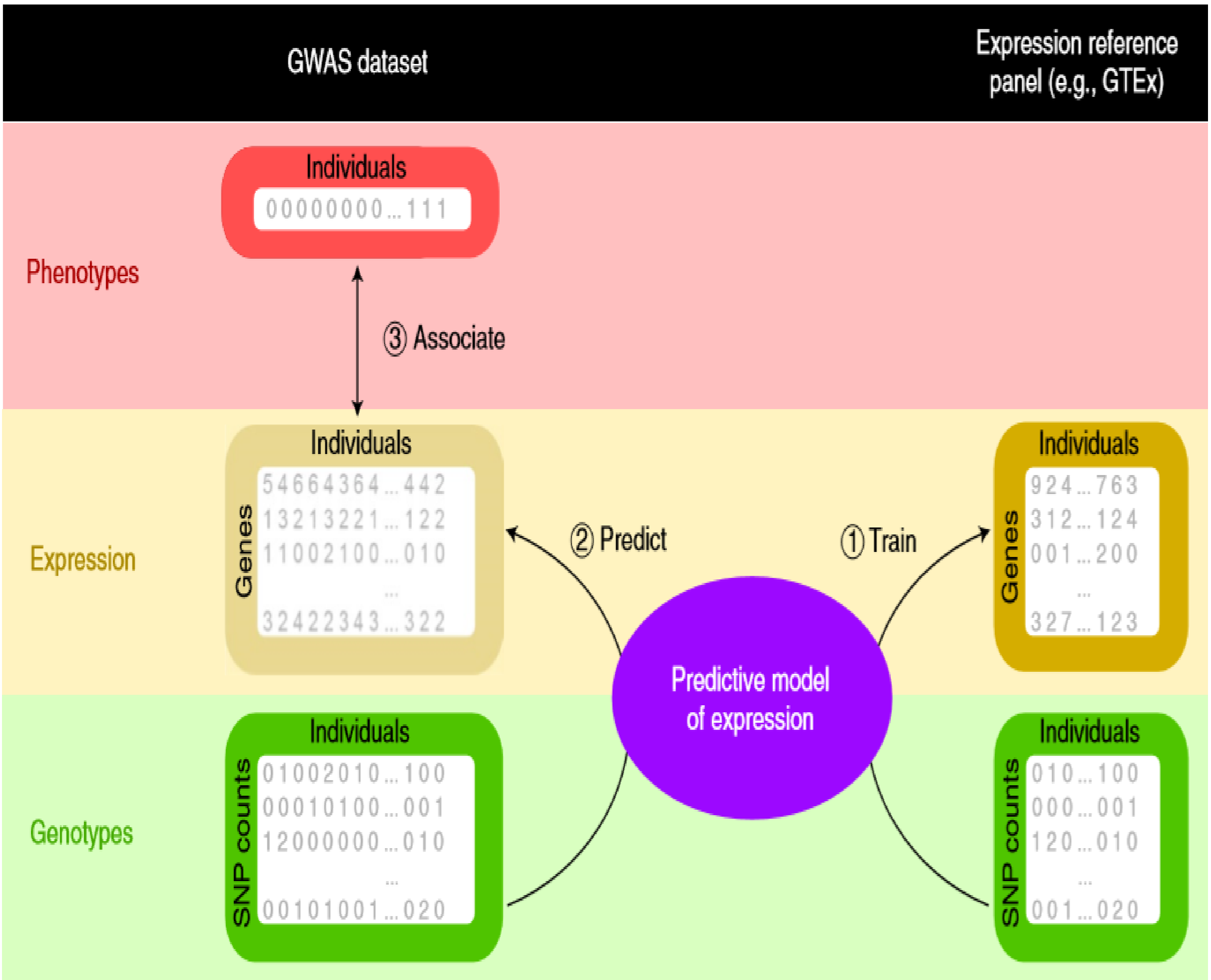


Figure 2.4. Three major steps of TWAS. (1) Training of predictive models using GTEx data. (2) Implementation of models to the GWAS cohort (3) Establishing the association between predicted expression and trait using statistical methods (Wainberg et al. 2019).

(1) Training of predictive models using GTEx data

In the first step, predictive models are trained in reference panels (e.g. GTEx data) using SNP and gene expression information. The learning of models is based on the expression variation for each gene using allele counts of genetic variants in the vicinity of the gene (typically 500 kb or 1 MB around the gene). Subsequently, these models estimate weights based on the correlation between SNPs and gene expression values in the training data while accounting for linkage disequilibrium (LD) among SNPs. In S-PrediXcan (Barbeira et al. 2018) the models were trained using the elastic net approach and the authors of this software have deposited the weights and SNP covariances in a publicly available resource (http://predictdb.org/)

(2) Application of TWAS models to the GWAS cohort and association testing

After estimating the effect sizes of genetic variants relative to their impact on gene expression levels in the reference panel, there can be two different ways in which a TWAS is performed to predict gene expression in target samples. One possible route is to predict gene expression at the individual level using effect sizes of cis-SNPs in the reference panel and measuring the association between the predicted gene expression and a trait. Individual-level prediction can be performed using the PrediXcan software (Barbeira et al. 2018). The gene expression value can be predicted using equation 7, where T_g is the gene expression level, w_{lg} is the weight of SNP l responsible for the expression of gene *g,* calculated by a trained model and X_l is the individual SNP dosage. The regression coefficients of phenotype Y on each gene's predicted expression can be calculated using equation 8:

$$\boldsymbol{T}_{\boldsymbol{g}} = \sum_{\boldsymbol{l}\,\boldsymbol{\varepsilon}\,\boldsymbol{model}\,\boldsymbol{g}} \boldsymbol{W}_{\boldsymbol{lg}}\boldsymbol{X}_{\boldsymbol{l}} \tag{7}$$

$$\boldsymbol{Y} = \boldsymbol{T}_{\boldsymbol{g}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{8}$$

Another route involves the estimation of the association between the predicted gene expression levels and a trait, by employing a weighted linear combination of SNP-trait standardized effect sizes (z-scores) and LD between SNPs (Figure 2.5). Fusion and S-PrediXcan software predict gene expression profiles for a group of individuals using GWAS summary statistics as an input file (Gusev et al. 2016)(Barbeira et al. 2018). The expression z-score Z_g of a gene can be calculated using equation 9,

$$\boldsymbol{Zg} = \sum_{\boldsymbol{l}\,\varepsilon\,\boldsymbol{model}\,\boldsymbol{g}} \boldsymbol{wlg}\,\frac{\boldsymbol{\sigma l}}{\boldsymbol{\sigma g}}\,\frac{\boldsymbol{\beta l}}{\boldsymbol{SE(\beta l)}} \tag{9}$$

The $\boldsymbol{\beta_l}$ term represents the regression coefficients of the SNPs in the regression model built on the GWAS data, with the phenotype/trait as dependent variable. The weights of SNPs $\boldsymbol{W_{lg}}$ and variance ratio $\boldsymbol{\sigma_l/\sigma_g}$ can be estimated from the training data sets by the software. As mentioned previously, the training set can be any reference transcriptomic dataset (1000 Genomes or GTEx data) where the prediction models are trained using the elastic-net method (Barbeira et al. 2018).

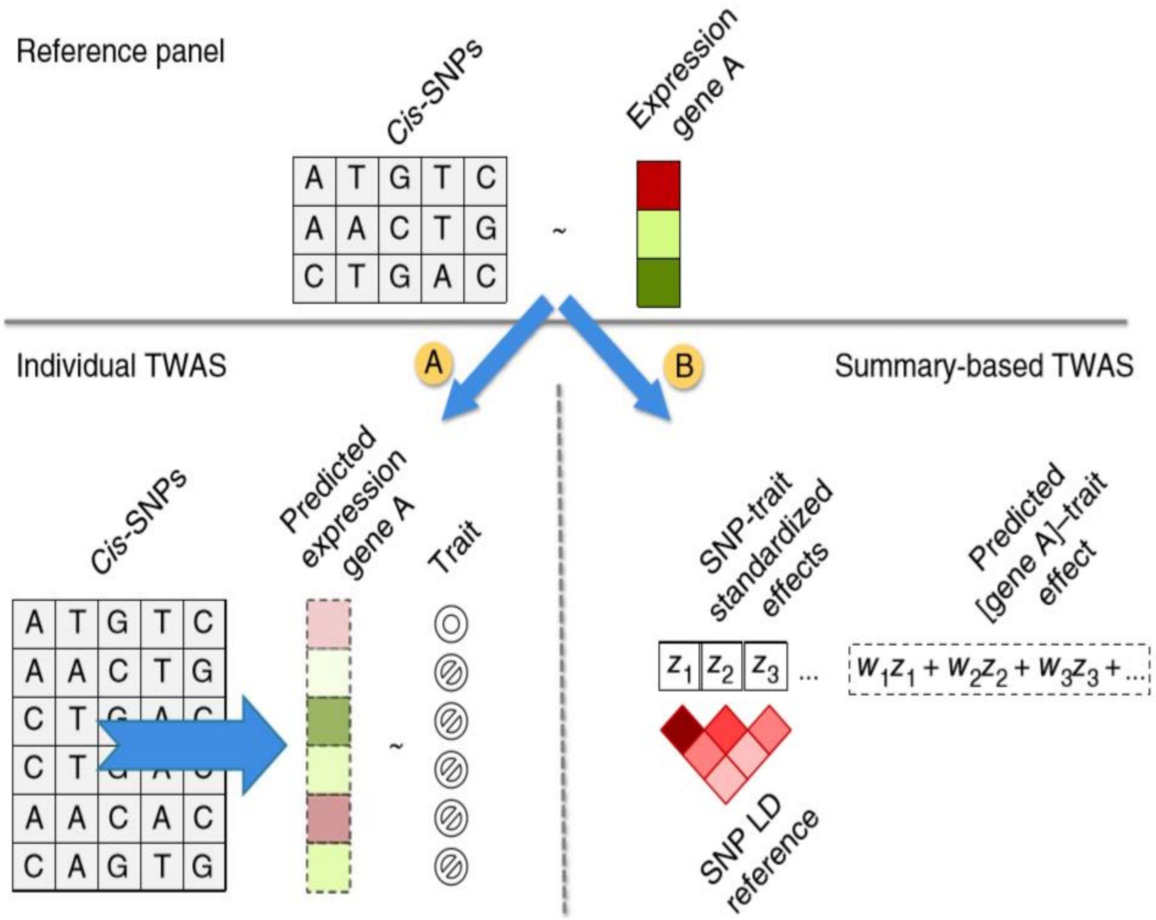


Fig 2.5. Two possible approaches for carrying out TWAS for imputing gene expression in target individuals (Gusev et al. 2016).

There are several advantages of the individual TWAS technique compared to traditional GWAS and expression studies (So et al. 2017).

(a) An advantage of transcriptomic-wide association studies is that the number of genes to test is generally much lower than the number of SNPs measured in a GWAS, making multiple testing correction less impactful on the raw results.

(b) TWAS sample sizes are usually significantly larger than those used in conventional expression studies and summary statistics are easily accessible for a number of traits.

(c) The expression profiles can be imputed for different tissues which can help to comprehend biological mechanisms at the tissue level.

There are also some disadvantages related to the TWAS method (Wainberg et al. 2019)

a) Only a small proportion of the heritability explained by SNPs can be attributed to gene expression regulation. Thus, with TWAS a significant part of the genetic signal available with GWAS is removed. (https://opain.github.io/GenoPred/Functionally_informed_prediction.html)

b) Genes whose expression is mainly driven by environmental factor are not considered in TWAS.

c) The model prediction accuracy might be affected by the sample size used for training the models in the tissue/cell.

d) eQTL could be also context-specific, that is their effect over gene expression might be due to gene-environment interactions and using TWAS model trained on reference tissue cannot detect these effects. This could be an advantage as pointed out (e.g., no confounding effect of treatment) but also a limitation (a disease could also lead to biological changes with specific eQTL regulation, e.g., inflammation related effects).

3. The Connectivity Map and its applications in pharmacogenomics.

The Connectivity Map (CMap) (Lamb et al. 2006) is a publicly available comprehensive library of transcriptional expression data that can be used to analyze relationships between drugs, cellular physiology, and disease states. For developing CMap database, researchers used the gene expression profiling method which is low cost and highly reproducible as compared to gene expression microarrays and RNA sequencing techniques. The most recent version of CMap has entered into the second phase of implementation as part of NIH's Library of Integrated Network-Based Cellular Signatures (LINCS) program (Subramanian et al. 2017). The current version of LINCS (also known as LINCS-L1000) contains 591,697 expression profiles generated from 29,668 compounds and genetic modifications (collectively known as 'perturbagens') across 98 different cell lines (N. Lim and Pavlidis 2019). Cell lines with the highest number of profiles are listed in Table 3.1. The CMap dataset consists of comparative data from human cells that treated with perturbagens and untreated cells (corresponding vehicle controls) representing a useful resource to identify differentially expressed genes (DEGs) in terms of z-scores. The CMap data can be downloaded from the CLUE Data Library (https://clue.io/data) or from the Gene Expression Omnibus repository (accession number GSE92742).

Table 3.1. Cell lines with the highest number of profiles in the LINCS CMap database.

Cell line	Tissue type	Profile count
A375	Skin	33,656
A549	Lung	37,577
HCC515	Lung	23,714
HA1E	Kidney	26,164
HEPG2	Liver	21,032
HT29	Colon	30,449
MCF7	Breast	52,373
PC3	Prostate	21,032
VCAP	Prostate	21,032

Note. Table adapted from (Musa, Tripathi, et al. 2018)

3.1 Major components of the CMap analysis pipeline

In brief, the analysis pipeline of CMap is comprised of three steps, described in following paragraphs.

(1) Gene expression signature of the biological state of interest

In the first step, a list of differentially expressed genes ranked on the basis of their z-score is obtained by comparing gene expression profiles of cases and controls (Fig 3.1A). This list will define the gene expression signature of the trait and will be used as an input query to the reference database. Depending on the nature of the study, the gene expression profile can be derived from human subjects or animal models of diseases. Alternatively, the expression profile can be suggested by disease experts who are able to identify the directionality of expression of specific genes corresponding to disease (Fig 3.1A). Defining a query signature is an essential requirement for CMap analysis. There is no 'gold standard' for generating expression profiles of the phenotype of interest and they can be obtained through traditional RNA sequencing methods or bioinformatics approaches based on genome-wide association data (Musa, Ghoraie, et al. 2018).

(2) Reference database

The gene expression signature representing a biological state is given as a query to CMap, which comprises gene expression profiles obtained from the treatment of cultured human cell lines with a large number of perturbagens (see above) (Lamb et al. 2006) (Fig 3.1B). Briefly, by comparing the expression profiles of each cell line before and after the treatment with the perturbagens is possible to know how these perturbagens modulate gene expression and to obtain the expression signature (reference profiles) of each perturbagen in each cell line.

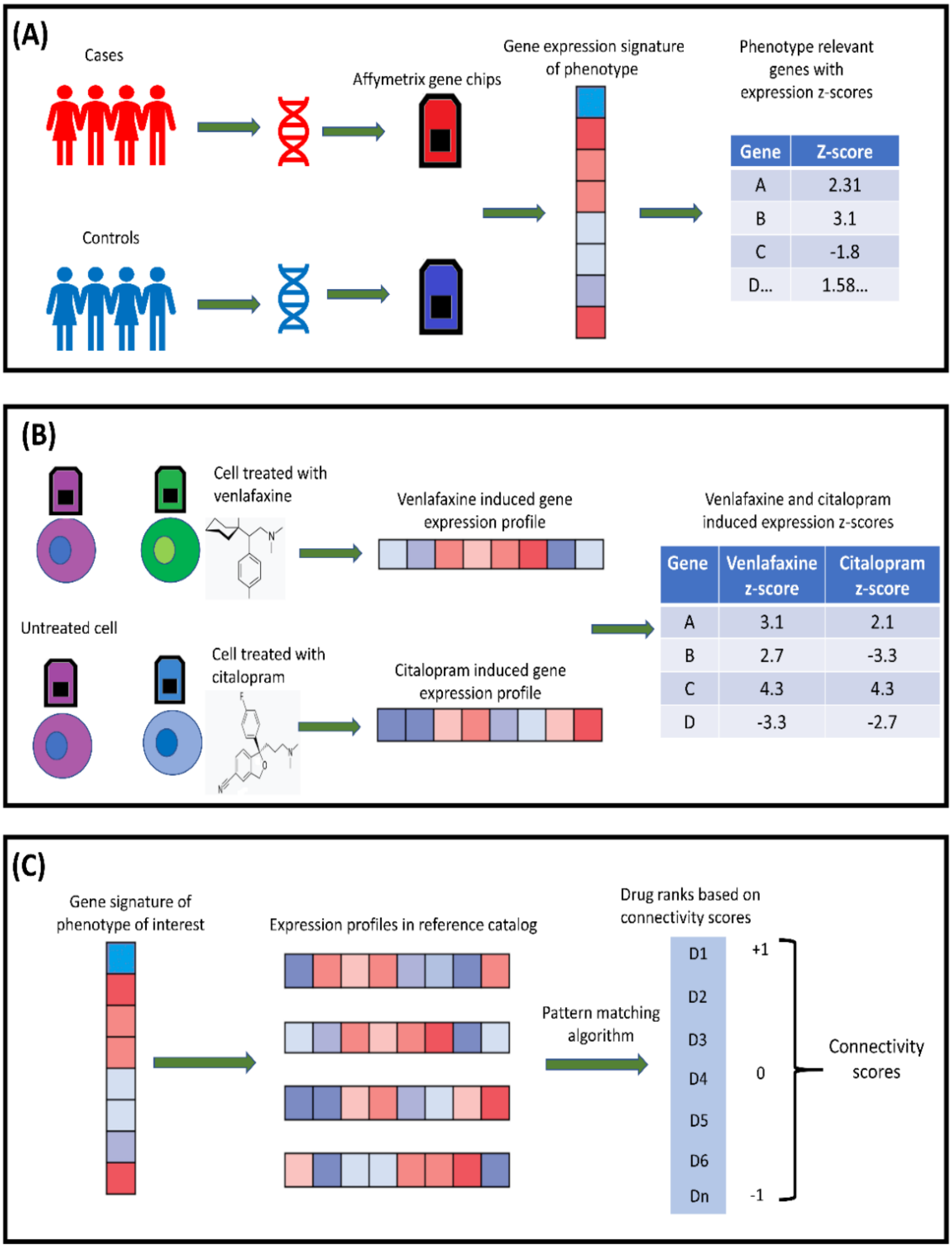
(3) Pattern matching algorithm (gene set enrichment analysis)

Finally, a pattern-matching algorithm is used to compare the gene expression profile of the trait of interest to gene expression profiles included in the reference database and to obtain for each comparison a connectivity score. Connectivity scores are a measure of the similarity between the expression profile of the trait under analysis and the expression profiles included in the CMap database.

The connectivity scores are calculated by means of a gene set enrichment analysis (GSEA, described in the following paragraph), based on Kolmogorov-Smirnov statistics (Lamb et al. 2006) (Fig 3.1C).

3.2 Gene Set Enrichment Analysis (GSEA) to estimate connectivity scores.

In the CMap pipeline, the goal of GSEA is to determine whether the genes of a given query signature are randomly distributed in the reference signature gene list or primarily located at the extremes (top or bottom). The resulting connectivity score is calculated by walking down the reference signature gene list increasing a running sum statistics when a gene available in query is encountered in the reference list and decreasing it when the gene of query signature is not present in the reference gene list. The connectivity score corresponds to the weighted Kolmogorov-Smirnov statistics. The magnitude of increment depends on the correlation of profile with the phenotype of interest (z-scores) (Subramanian et al. 2005). The significance value of the calculated score is obtained through the permutation procedure. The values of connectivity scores range between +1 and -1. A positive connectivity score represents a positive correlation, and a negative connectivity score represents a negative correlation between the query signature and drug profiles in the reference database. A null connectivity score occurs when there is no correlation between the query profile and the profiles in the reference catalog (Lamb et al. 2006).


(A)
Cases
Affymetrix gene chips
Gene expression signature of phenotype
Phenotype relevant genes with expression z-scores
Controls

Gene	Z-score
A	2.31
B	3.1
C	-1.8
D...	1.58...

(B)
Cell treated with venlafaxine
Venlafaxine induced gene expression profile
Venlafaxine and citalopram induced expression z-scores
Untreated cell
Cell treated with citalopram
Citalopram induced gene expression profile

Gene	Venlafaxine z-score	Citalopram z-score
A	3.1	2.1
B	2.7	-3.3
C	4.3	4.3
D	-3.3	-2.7

(C)
Gene signature of phenotype of interest
Expression profiles in reference catalog
Pattern matching algorithm
Drug ranks based on connectivity scores
D1
D2
D3
D4
D5
D6
Dn
+1
0
-1
Connectivity scores

Fig 3.1. Overview of CMap concept. (A) Generating the expression signature of the biological state of interest. (B) Mechanism of drug profiles obtained by treating various cell types with perturbagens for CMap database. (C) Major steps of CMap analysis.

3.3 Application of CMap in pharmacogenomics

The CMap database has been used in Pharmacogenomics since its introduction in 2006. The concept of CMap has been used in the identification of new therapeutic options for the treatment of several diseases. Moreover, multiple studies adopted this approach to evaluate the drugs' mechanisms of action (MoA) and to find new implications of already existing, approved drugs (drug repurposing).

3.3.1 Finding novel therapeutic options

The most important application of the CMap database is the discovery of novel treatment options for diseases. In a recent study by Lim et al, a gastric cancer gene signature was used to query CMap (S. M. Lim, Lim, and Cho 2014). The authors of the study found vorinostat and trichostatin A as potential candidates for treating gastric cancer. These findings were validated in experimental settings where vorinostat significantly inhibited cell viability in gastric cancer cell lines. Another study conducted by Siu et al mentioned polyphyllin D as a therapeutic option for non-small cell lung cancer (Siu et al. 2008). This study demonstrated that polyphyllin D can activate the pathways involved in the apoptosis mediated by the estrogen receptor and mitochondria.

3.3.2 Evaluating a drug's mechanism of action

In pharmacology, studying the exact impact of a drug on the biological systems is important to understand its mechanism of action and it is also crucial to identify new compounds that are active towards specific targets. With the help of CMap, it is possible to identify signalling pathways affected by a drug, which can further help us to comprehend the biology of drug-disease relations. CMap was used to construct a drug network (DN) based on a distance metric with the ability to score the similarity between gene expression profiles and drug treatment (Iorio et al. 2010). The investigators used the graph theory to partition the DN to identify the drugs that were highly interconnected, hence, they have similar MoA and therapeutic purposes. Notably, the study by Iorio et al also found that the MoA of fasudil, a Rho-kinase inhibitor, was highly interconnected with autophagy. The drug, previously used in the treatment of cerebral vasospasm, was experimentally validated and its usage as a repurposed drug was confirmed for other disease.

3.3.3 Drug repurposing

Conventional drug development is a time-consuming process and involves high costs. CMap has been used extensively in the domain of drug repurposing. For instance, Johnstone and his colleagues suggested a possible mechansim of calmodulin signaling using piperazine as promoters of central nervous system (CNS) neurite growth (Johnstone et al. 2012). Moreover, using a CMap-based approach, they found that piperazine and phenothiazine, two antipsychotics, have the potential to be repurposed for neuron regeneration. Similarly, using CMap, Jin and colleagues presented a novel drug repurposing strategy for treating type II diabetes (Jin et al. 2014). They reported that the combination of Trolox C and Cytisine is effective for the treatment of type 2 diabetes, but if these drugs are used separately, neither of them helps in achieving the desired outcome.

Research Part

4. Project aims and outline of the thesis

The main purpose of this thesis work was to investigate computational methods contributing to precision psychiatry. Previous studies have focused on the identification of genetic variants associated with AD efficacy (Uher et al. 2013) (Wigmore et al. 2020), and here we expand the focus to transcriptomic profiles derived from TWAS using data from eQTL studies. The Ph.D. project described here can be divided into two major studies. In the first study, we developed and tested an *in-silico* approach aiming at the identification of the gene expression profile associated with remission in MDD following citalopram treatment in the STAR*D study, and comparing this profile with various AD-induced (including citalopram-induced) transcriptional profiles available in the CMap database. In the second study, we developed polygenic models using supervised learning algorithms based on the genetic associations of the individual profile of STAR*D subjects and in-vitro antidepressant profiles from CMap for the prediction of AD response.

The two aforementioned studies described in this thesis were developed along the following steps:

First study

1. After classifying STAR*D participants as remitters or non-remitters to citalopram, a GWAS was conducted to test the association of SNPs with remission from MDD
2. From the GWAS summary statistics data, a TWAS was performed (using the FUSION software) to identify the gene expression profile corresponding to remission in MDD in STAR*D individuals
3. We extrapolated and processed various AD-induced expression profiles across several human cell lines from the CMap database
4. Finally, we ranked ADs based on the likelihood of efficacy by comparing their associated expression profiles to the one induced by the treatment with citalopram and associated with remission in MDD.

Second study

1. Individual expression profiles were imputed in STAR*D first stage individuals (citalopram remitter and non-remitters) using the Predixcan tool.
2. Development of polygenic models using supervised learning approach to predict citalopram remission based on the genetic associations of individual genetic profiles and in-vitro drug expression profiles.

5. Investigating an *in-silico* approach for prioritizing antidepressant drug prescription based on drug-induced expression profiles and predicted gene expression

5.1 Introduction

Antidepressant (AD) prescription is currently based on international guidelines, previous clinical experience, and the presence of co-morbidities, but is largely a 'trial and error' practice (Leuchter et al. 2010). SSRIs represent the most common AD class currently used for treating MDD with, however, a highly variable response to SSRI treatment between patients (Biernacka et al. 2015). Any attempt to achieve symptom remission by switching between treatment or combining multiple pharmacotherapies can take weeks to be fully evaluated in terms of clinical effectiveness.

The availability of objective and reproducible predictors of AD response could reduce the time needed to perform such evaluation, and, in turn, to achieve remission and relieve patients' suffering (Leuchter et al. 2010). Prior studies suggest that AD response and remission are heritable traits (Tansey et al. 2013b), offering the opportunity to use genetic markers in the attempt to predict optimal drug prescription. In light of these observations, the development of strategies taking into account various factors related to the clinical presentation, the patient's genotype and their metabolism at baseline has been warranted (Gandal et al. 2016).

In this rapidly evolving context, the purpose of the work presented here was to develop a new approach aiming to contribute to precision psychiatry precisely by providing effective ways to predict patients' response to therapeutic agents. As previously mentioned, several studies published in the last decade have focused on the identification of genetic variants predictive of AD efficacy (Uher et al. 2013)(Wigmore et al. 2020), and here we applied TWAS concept to obtain gene expression profiles using GWAS summary statistics. Transcriptomic profiles associated with the efficacy of specific ADs in clinical trials can be compared with the *in vitro* AD-induced gene expression changes, in order to test if drug-induced gene expression signatures could be used as markers of clinical efficacy of specific ADs. The main aim of this study was to develop and test this approach by computing gene expression profile associated with remission to citalopram in the STAR*D study and comparing this profile with citalopram and other ADs induced transcriptional responses available from the Connectivity Map (CMap) database. CMap is a genome-scale library of cellular signatures and a catalogue of transcriptional responses to chemical and genetic perturbations (Lamb et al. 2006). A positive correlation between expression profiles of citalopram remission and *in vitro* citalopram induced gene changes was hypothesized to be indicative of the potential utility of our approach. Within this analysis, we have hypothesized that results specific to escitalopram would replicate those of citalopram, being the former the therapeutically active enantiomer of the latter (Tsuchimine et al. 2018). In a second step, we have applied the same approach on control drugs (no known AD effect) to provide a proof of principle of the usefulness of the method, since control drugs were expected to have less similarity to citalopram remission gene expression profiles than (es)citalopram and other ADs.

5.2 Methods

5.2.1 Study Population

This study is based on the STAR*D data (Fava et al. 2003). The STAR*D is a clinical trial of protocol-guided antidepressant treatment for outpatients with MDD. The study included 4,041 treatment-seeking adult outpatients, recruited in 18 primary care and 23 psychiatric clinical sites across the United States, comprised in the STAR*D cohort. Genotyping was performed in 1,948 participants (Garriock et al. 2010). Our analysis used data from the first treatment step (level 1), which consisted of protocol-guided citalopram (20–60 mg/day). Remission was defined as a score < 6 on the Quick Inventory of Depressive Symptomatology clinician-rated (QIDS-C) scale at level 1 exit (after 12 weeks of citalopram treatment), in line with the previous literature (Novick et al. 2015)(Rush et al. 2003). The choice of remission over symptom improvement as main endpoint of the study as the former was associated with better disease prognosis and lower risk of relapse in STAR*D participants(Bradley N. Gaynes, M.D. et al. 2009). STAR*D genotype and phenotype data are available through the National Institute of Mental Health Human Genetics Initiative (https://www.nimhgenetics.org/). The STAR*D study recruited non-psychotic MDD patients aged 18-75 years from psychiatric and primary health care clinics, followed up between 2000 and 2004 (Gaynes et al. 2008). The study design of STAR*D comprised of four treatment levels to assess treatment response, with each level consisting in 14 weeks of treatment. All STAR*D patients were initially treated with the level 1 treatment, and all patients not achieving significant remission by the end of each treatment level were entered the higher-tier level (Fava et al. 2003) (Trivedi et al. 2006). Alternatively, treatment was suspended on patients of each level with significant symptomatic improvement or remission and then they were followed up for one year. Genetic material was collected from 1,948 (48%) participants; of whom 1,491 (37% of the original STAR*D sample, including 980 of white/European ancestry) passed quality control and were included in previously reported genome-wide analyses (Garriock et al. 2010). The study was approved by institutional ethics review boards at all centres. Written consent was obtained from all participants after the procedures and any associated risks were explained.

5.2.2 Genotyping, quality control, and imputation

Details on the genotyping procedure can be found elsewhere (Garriock et al. 2010). Individual genotype data for the STAR*D cohort was processed using the PGC "RICOPILI" pipeline for standardized quality control, imputation, and association analysis (Lam et al. 2019). Quality control and imputation were performed according to the standards from the PGC. The default parameters for retaining SNPs and subjects were: SNP missingness < 0.05 and 0.02 for samples, before and after sample removal; subject missingness < 0.02 for SNPs; autosomal heterozygosity deviation ($|F_{het}|$<0.2); difference in SNP missingness between cases and controls < 0.02; and SNP Hardy-Weinberg equilibrium (P>10^{-6} in controls or P>10^{-10} in cases).

Genotype imputation was performed using the pre-phasing/imputation stepwise approach implemented in IMPUTE2 / SHAPEIT (chunk size of 3 Mb and default parameters). The imputation reference set consisted of 2,186 phased haplotypes from the 1000 Genomes Project dataset (August 2012, 30,069,288 variants, release "v3.macGT1"). After imputation, we identified SNPs with optimal imputation quality (INFO >0.8) and missingness (<1%) values included in the PCA from which the resulting principal components were subsequently used as covariates in the final association analysis. SNPs underwent linkage disequilibrium-based pruning (r^2 > 0.02) and frequency filtering (MAF > 0.05). This SNP set was used for robust relatedness testing and population structure analysis. Relatedness testing aided identification of duplicated samples and pairs of subjects with $\hat{\pi}$ > 0.2, where one randomly selected member of each pair was removed, with the only preference being the retention of cases over controls.

5.2.3 Statistical analysis

(i) Genome-wide Association Study

As previously mentioned, a GWAS was conducted using the RICOPILI pipeline to test the association of each SNP with remission to citalopram, classifying STAR*D participants (*N=1163*) as remitters or non-remitters. The logistic regression analysis included the covariates of sex, age, baseline QIDS-C score, and the first 20 population principal components from the PCA performed on the genotypes. The GWAS summary statistics were then converted to LD-score regression format using the munge_sumstats.sh script, removing SNPs with an INFO < 0.3 (https://github.com/bulik/ldsc/wiki).

(ii) Transcriptome-wide Association Study

We used STAR*D GWAS summary statistics to perform a TWAS using the FUSION software (Gusev et al. 2016). Briefly, FUSION requires pre-computed gene expression SNP-weights and GWAS summary statistics to predict the association between the expression of each gene and the phenotype of interest. SNP-weights from the dorsolateral prefrontal cortex data of the CommonMind Consortium (DLPFC), 48 tissues including 13 brain regions within GTEx, the Young Finn study, the Netherlands Twin Registry, and the Metabolic Syndrome in Men study datasets were considered (Table 5.1). The SNP weights were previously derived by FUSION authors (Gusev et al. 2016)(Gusev et al. 2018)(Mancuso et al. 2018). Since possible confounding factors were considered when estimating the SNP weights by including known and hidden covariates (Stegle et al. 2010), we assumed medication usage in donors of the above-mentioned data sources was unlikely to have had an impact on the SNP weights. From the GTEx database, we considered a wide range of tissues in addition to brain regions because of their larger sample sizes and the presence of a moderate correlation of cis-expression quantitative trait loci (eQTL) effects among different tissues (Consortium 2017). All gene expression SNP-weights were downloaded from the FUSION website (http://gusevlab.org/projects/fusion/). This study uses the term SNP-weight sets to define SNP-weights from a given sample and tissue (e.g. GTEx hippocampus, CMC DLPFC). Furthermore, each gene within a given SNP-weight set constitutes a *feature* or *gene-tissue pair.* We combined the FUSION output for all SNP-weight sets, using the TWAS associations (*z*-scores) to represent the gene expression signature of citalopram remitters. The 52 SNP-weight sets in this study contained 252,878 features, representing 26,363 unique genes. Where multiple features for a single gene were available, only the feature providing the highest cross-validation coefficient of determination (CV-R^2) was retained. We did not define any CV-R^2 threshold for feature selection. Similar criteria have been implemented elsewhere (Pain et al. 2019).

Table 5.1. Tissues considered for TWAS

GTEx v7 multi-tissue (RNA-seq)	
Tissue	**No**
Adipose - Subcutaneous	385
Adipose - Visceral (Omentum)	313
Adrenal Gland	175
Artery - Aorta	267
Artery - Coronary	152
Artery - Tibial	388
Brain - Amygdala	88
Brain - Anterior cingulate cortex (BA24)	109
Brain - Caudate (basal ganglia)	144
Brain - Cerebellar Hemisphere	125
Brain - Cerebellum	154
Brain - Cortex	136
Brain - Frontal Cortex (BA9)	118
Brain - Hippocampus	111
Brain - Hypothalamus	108

Brain - Nucleus accumbens (basal ganglia)	130
Brain - Putamen (basal ganglia)	111
Brain - Spinal cord (cervical c-1)	83
Brain - Substantia nigra	80
Breast - Mammary Tissue	251
Blood - EBV-transformed lymphocytes	117
Skin - Transformed fibroblasts	300
Colon - Sigmoid	203
Colon - Transverse	246
Esophagus - Gastroesophageal Junction	213
Esophagus - Mucosa	358
Esophagus - Muscularis	335
Heart - Atrial Appendage	264
Heart - Left Ventricle	272
Liver	153
Lung	383
Minor Salivary Gland	85
Muscle - Skeletal	491
Nerve - Tibial	361
Ovary	122
Pancreas	220
Pituitary	157
Prostate	132
Skin - Not Sun Exposed (Suprapubic)	335
Skin - Sun Exposed (Lower leg)	414
Small Intestine - Terminal Ileum	122
Spleen	146
Stomach	237
Testis	225
Thyroid	399
Uterus	101
Vagina	106
Whole Blood	369
Common mind consortium (RNA seq)	
Brain prefrontal cortex	452
Metabolic Syndrome in men (RNA seq)	
Adipose	563
Young Finns Study (Expression microarray)	
Blood	1264
Netherland twin registry (Expression microarray)	
Blood	1247

(iii) Comparison of TWAS results with *in vitro* AD-induced gene expression

We evaluated the correlation between the TWAS expression profile of citalopram remitters extrapolated from the STAR*D cohort data with the *in vitro* gene expression profiles associated with various antidepressants available in CMap (Phase II data) (Figure 5.1).

We considered the expression profiles of 21 ADs (Table 5.2) in 5 human cell lines available in Phase II of CMap ((a) A375, Human malignant melanoma (b) MCF7, Breast cancer (c) PC3, prostate cancer (d) HA1E, kidney (e) HT29, colon cancer).

Table 5.2. List of Antidepressants and drug class

Antidepressants	Drug Class
Citalopram Escitalopram Fluoxetine Fluvoxamine Paroxetine Sertraline	Selective serotonin reuptake inhibitor
Trazodone Duloxetine Venlafaxine	Serotonin-norepinephrine reuptake Inhibitor
Amitriptyline Imipramine Nortriptyline Trimipramine Clomipramine Dosulepin	Tricyclic antidepressant
Maprotiline Mianserin Mirtazapine	Tetracyclic antidepressant
Tranylcypromine Selegiline	Monoamine oxidase inhibitor
Reboxetine	Noradrenaline reuptake inhibitor

Drug-induced expression profiles were evaluated in cells treated for 24 hours with 10µm drug concentration. We used CMap's GEO series (GSE70138) data and extracted relevant expression profiles using cmapR package (https://github.com/cmap/cmapR). Of the 12,328 genes within the CMap profiles, 10,027 were captured by the SNP-weight included in the citalopram remitter TWAS. We compared the expression profiles of the 21 ADs with the profile of citalopram remitters obtained from the TWAS using an approach described in a previous study (So et al. 2017). As an example, some differentially expressed genes represented in terms of z-scores of citalopram remitters and drug-induced profiles are reported in Table 5.3

Table 5.3 Gene expression value in terms of z-scores in the TWAS and drug-induced profile

Gene	Citalopram-remitters expression Z-score (TWAS)	Citalopram-induced expression Z-score (TWAS)	Venlafaxine-induced expression Z-score (CMap)
SLC31A2	-1.24	-0.58	-1.8

EPCAM	3.57	0.33	2.37
RBM6	-2.16	-0.49	0.25
CBR3	-0.41	2.38	-0.75
HMGCS1	0.64	3.04	3.46
⋮	⋮	⋮	⋮

The differential expression profiles of remission from STAR*D and drugs from CMap were analysed using R code (https://sites.google.com/site/honcheongso/software/gwascmap) (So et al. 2017), according to the following procedure:

a. *Evaluating the relationship between AD-induced gene expression and expression profiles of citalopram remitters*. Patterns of expressions were tested by analysing all and the most up-regulated and down-regulated genes in the TWAS (*k* = 50, 100, 250, 500). The correlation between CMap antidepressant profiles and the STAR*D remitter profile was assessed for each drug using both the Spearman's and the Pearson's correlation coefficients using all the aforementioned *k* genes. Furthermore, we adopted also the KS test – as reported by the original CMap study (Subramanian et al. 2005) – to compare the expression patterns of AD and citalopram remitters and calculated connectivity scores (Lamb et al. 2006). The 21 tested ADs were ranked based on the results of each test (Pearson, Spearman, and KS), and then the average rank across the tests for each drug were computed. Drugs were ranked in ascending order of correlation coefficients. As an example, Table 5.4 and 5.5 report the rank of citalopram and venlafaxine estimated with the three different methods and the average rank calculation.

Table 5.4 Average rank correlation results for Citalopram using Pearson, Spearman, and KS test.

Gene Subset	Pearson		Spearman		KS-test	
	Coefficient	Rank	Coefficient	Rank	Connectivity score	Rank
All	0.5	1	0.4	1	-	-
50	0.3	1	0.4	1	0.2	2
100	0.4	1	0.3	1	0.2	1.5
250	0.1	1	0.2	2	0.1	2
500	0.2	1.5	0.1	1.5	0.1	1
Mean rank	-	1.1	-	1.3	-	1.62
Average rank of Citalopram = 1.34						

Table 5.5 Average rank correlation results for Venlafaxine using Pearson, Spearman, and KS test.

Gene Subset	Pearson		Spearman		KS-test	
	Coefficient	Rank	Coefficient	Rank	Connectivity score	Rank
All	0.2	3	0.3	3.2	-	-
50	0.1	3	0.1	2	0.2	3.7
100	0.3	2	0.4	3	0.3	2.6
250	0.1	3	0.3	3	0.1	2.6
500	0.4	2.5	0.1	2.8	0.1	3.2
Mean rank	-	2.7	-	2.8	-	3.02
Average rank of Venlafaxine = 2.84						

b. *Significance of ranks using permutation.* To estimate the significance of the ranks, a one-sided permutation procedure was performed by shuffling the z-scores obtained in the TWAS and calculating the corresponding rank of each drug by repeating the procedure in step a. One hundred permutations were performed to calculate the distribution of ranks under the null hypothesis and estimated the p-value of the observed ranks.

c. *Calculation of ranks probability for each AD across cell lines using the Genome Scan Meta-Analysis (GSMA) method*. We combined ranks of each AD in five cell lines by adding them and calculated the sum of ranks probability using GSMA, a non-parametric method for meta-analysing ranks (Wise, Lanchbury, and Lewis 1999)

Finally, we repeated the process described above in (a), (b), and (c) for five control drugs (Table. 5.6) having hypothetically no antidepressant effect to validate the proposed method.

Table 5.6. List of control agents and drug class

Control Drugs	Drug Class
Pantoprazole	Proton pump inhibitors
Clofibrate	Fibrates
Rifaximin	Antibiotic
Acarbose	Alpha-glucosidase inhibitors
Ipriflavone	Isoflavone

The major steps of the applied in-silico method are shown in Figure 5.1

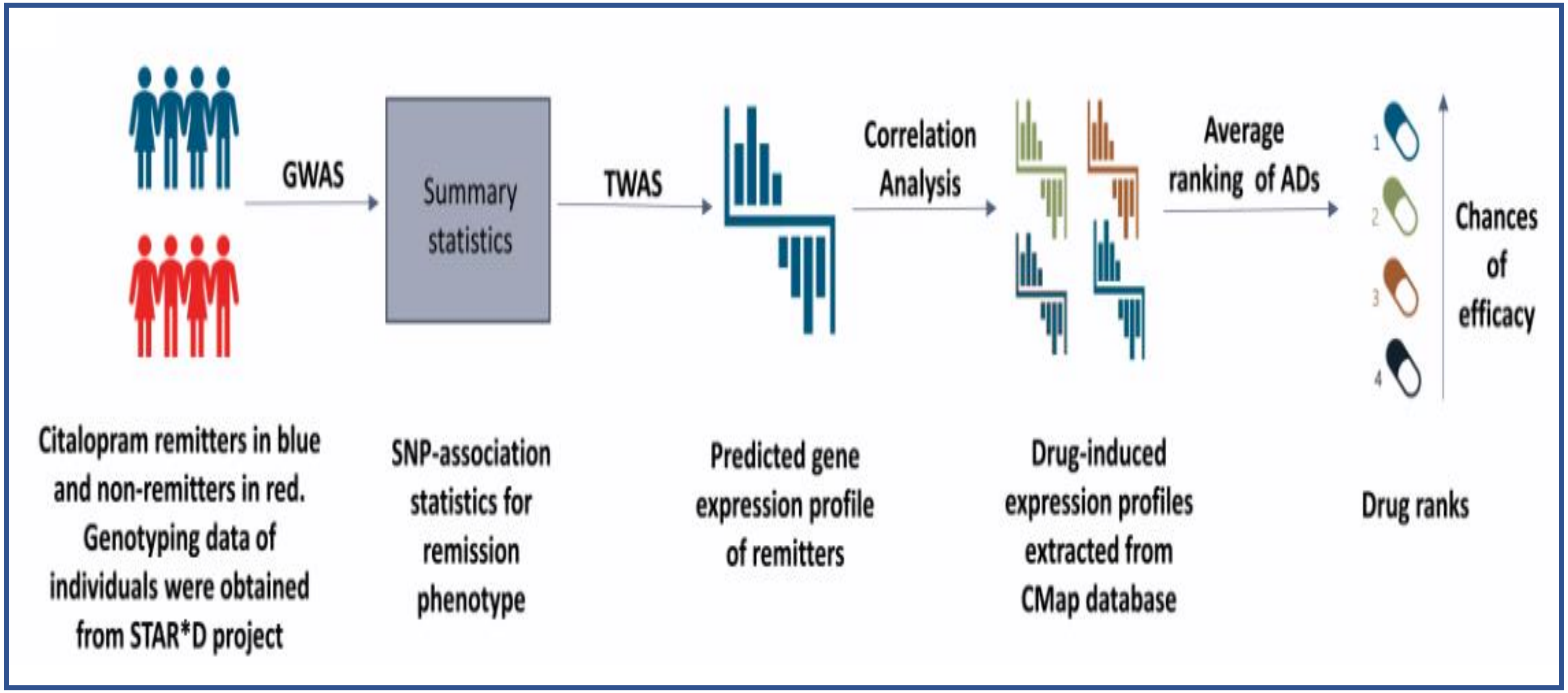

GWAS
Summary statistics
TWAS
Correlation Analysis
Average ranking of ADs
1
2
3
4
Chances of efficacy
Citalopram remitters in blue and non-remitters in red. Genotyping data of individuals were obtained from STAR*D project
SNP-association statistics for remission phenotype
Predicted gene expression profile of remitters
Drug-induced expression profiles extracted from CMap database
Drug ranks

Figure 5.1. Illustration of the major steps of the proposed in-silico method

5.3 Results

STAR*D data included 506 citalopram remitters and 657 non-remitters. The main clinical-demographic characteristics of the samples are shown in Table 5.7.

Table 5.7 Main clinical demographic characteristics of STAR*D

Number of individuals N	1163
Level 1 citalopram remitters	506
Level 1 citalopram non-remitters	657
Female proportion	0.58
Mean age (SD)	43.33 (13.49)
Mean baseline QIDS-C score (SD)	16.14 (3.16)

As reported in Figures 5.2, the GWAS and TWAS Q-Q plots showed no evidence of confounding, therefore data were used for the comparison with the AD induced expression profile available in the CMap database.

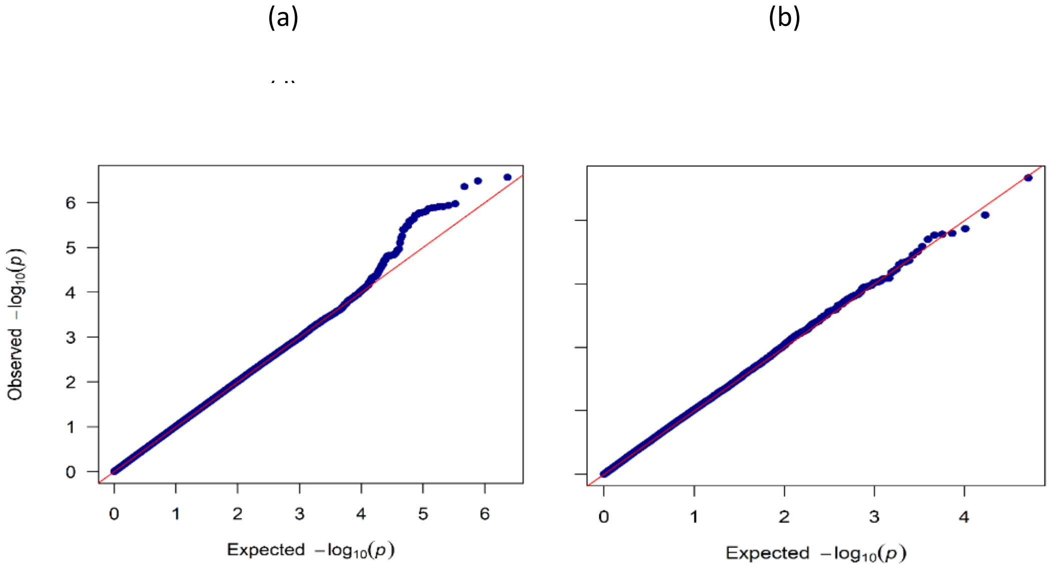


Figure 5.2. (a) QQ plot of GWAS p-values for *citalopram remission N* (p-values) = 1158655*. (b) TWAS p-values for citalopram remission* N (p-values) = 26363.

The full results obtained on the 5 different cell lines for all the 21 tested ADs are displayed in Table 5.8. The average rank across tests for ADs showed that escitalopram (S-enantiomer of citalopram) was the AD with the highest average rank, followed by amitriptyline in MCF7 (breast cancer cell line). In A375 (human malignant melanoma) escitalopram ranked second after trimipramine, whereas in PC3 cells (prostate cancer), citalopram was the second-highest ranked AD after mirtazapine. In HT29 cells (colon cancer), citalopram ranked third after trimipramine and dosulepin, while inHA1E cells (kidney) escitalopram and citalopram did not rank in top positions, and the two highest-ranked ADs were imipramine and fluvoxamine. In the analysis of combined ranks across cell lines, we found sertraline, trimipramine, and venlafaxine as drugs with the best sum of ranks and a nominally significant combined p-value across the five cell lines. Citalopram globally ranked 4th, with a near-significant p-value (0.057) (Table 5.8). In summary, despite the marked variability in terms of same-AD ranking among the different cell lines, (es)citalopram-induced expression signatures were found to be significantly correlated with the citalopram remission profile in three cell lines (A375, MCF7, and PC3).

We attempted to validate our approach by also analysing the expression profiles of five control drugs, without any known AD effect. In A375 all control drugs ranked after (es)citalopram. In PC3, three control drugs ranked after (es)citalopram. In MCF7, four control drugs were ranked after escitalopram and one control drug was ranked after citalopram. In HT29 and HA1E, four control compounds ranked before the (es)citalopram. In the combined rank analysis, rifaximin, trimipramine, and clofibrate had the top sum ranks across cell lines, though the only rifaximin was nominally significant ($p<0.05$) (Table 5.9). Our hypothesis of higher correlation of (es)citalopram-induced gene expression with citalopram remission TWAS results compared to control drugs was partially confirmed only in two cell lines (A375 and in MCF7).

Table 5.8. Ranking of ADs in five human cell lines.

Cell lines	**A375**		**MCF7**		**PC3**		**HA1E**		**HT29**		**Combined cell lines**	
Drug	**Rank**	**Perm.p.value**	**Rank**	**Perm.p.value**	**Rank**	**Perm.p.value**	**Rank**	**Perm.p.value**	**Rank**	**Perm.p.value**	**Rank Sum**	**p.value**
Amitriptyline	16.9	0.8610	2.4*	0.0290	12.6	0.6138	17.6	0.8948	12.9	0.6224	62.4	0.7049
Citalopram	7.6	0.2862	7.7	0.2890	2.3*	0.0276	10.6	0.4776	4.4	0.1114	32.6	0.0571
Clomipramine	11.1	0.5229	4.1	0.0890	9.1	0.3714	18.6	0.9371	16.8	0.8590	59.7	0.6533
Dosulepin	17.3	0.8824	18.1	0.9257	15	0.7605	8.8	0.3571	2.6*	0.0405	61.8	0.7049
Duloxetine	11	0.5181	20.9	0.9971	17.2	0.8776	11.3	0.5305	4.8	0.1300	65.2	0.7753
Escitalopram	2.6*	0.0310	1.2*	0.0014	6.9	0.2395	14.5	0.7324	17.6	0.8976	42.8	0.2034
Fluoxetine	13.8	0.6843	10.8	0.4890	17.8	0.9033	6.2	0.1976	16.7	0.8562	65.3	0.7752
Fluvoxamine	19.6	0.9729	6.2	0.2010	2.9	0.0443	3.2	0.0529	9.6	0.4005	41.5	0.1833
Imipramine	6.2	0.2014	12.6	0.6052	14.6	0.7395	1.2*	0.0062	5.2	0.1490	39.8	0.1468
Maprotiline	3.7	0.0667	6.8	0.2362	9.8	0.4200	11.2	0.5229	10.5	0.4590	42	0.1833
Mianserin	19.6	0.9729	15.6	0.8029	17.8	0.9033	17.9	0.9062	19.9	0.9795	90.8	0.9979
Mirtazapine	9.8	0.4362	18.8	0.9529	2*	0.0181	7.7	0.2871	5.2	0.1490	43.5	0.2247
Nortriptyline	11.6	0.5576	15.2	0.7833	12	0.5748	18.1	0.9167	7.9	0.2990	64.8	0.7752
Paroxetine	18.3	0.9248	14.4	0.7390	6.4	0.2129	14.7	0.7448	16.7	0.8562	70.5	0.8695
Reboxetine	11.8	0.5681	10.2	0.4476	10.7	0.4895	19.8	0.9800	13.3	0.6552	65.8	0.7965
Selegiline	10.6	0.4890	17	0.8667	19.2	0.9581	15	0.7671	15.6	0.8005	77.4	0.9512
Sertraline	4.4	0.1033	6.3	0.2052	9.1	0.3714	3.6	0.0662	7.4	0.2743	30.8*	0.0412
Tranylcypromine	16.9	0.8610	14.8	0.7619	18	0.9119	9.5	0.3929	16.9	0.8614	76.1	0.9428
Trazodone	10.3	0.4657	10.9	0.4943	13	0.6424	8	0.3043	16.6	0.8529	58.8	0.6264
Trimipramine	2.4*	0.0267	13.4	0.6614	4	0.0919	9.7	0.4086	1.4*	0.0105	30.9*	0.0412
Venlafaxine	5.5	0.1567	3.6	0.0676	10.6	0.4819	3.8	0.0810	9	0.3657	32.5*	0.0487

Note. P-values were obtained through the permutation procedure described in section 5.2.3. Top ranks are marked with an asterisk and p-values < 0.05 are highlighted in gray in each cell line and the combined cell line results. The ADs (es)citalopram are mentioned in bold.

Table 5.9. Ranking of ADs and control drugs in five human cell lines.

Cell lines	**A375**		**MCF7**		**PC3**		**HA1E**		**HT29**		**Combined cell lines**	
Drug	**Rank**	**Perm.p.value**	**Rank**	**Perm.p.value**	**Rank**	**Perm.p.value**	**Rank**	**Perm.p.value**	**Rank**	**Perm.p.value**	**Rank Sum**	**p.value**
Acarbose	15.9	0.6392	10.2	0.3238	17.7	0.7196	11.7	0.4108	1.4*	0.0062	56.9	0.2809
Amitriptyline	19.4	0.8054	3.6*	0.0477	16.4	0.6554	22.1	0.9088	16.9	0.6892	78.4	0.7385
Citalopram	6.6	0.1581	11.1	0.3723	4.1	0.0596	13.8	0.5146	7.9	0.2227	43.5	0.0878
Clofibrate	10.8	0.3504	7.2	0.1885	17.2	0.6954	3.6*	0.0442	3.1*	0.0377	41.9	0.0698
Clomipramine	14.7	0.5777	6.7	0.1596	11.7	0.4069	23.4	0.9523	21	0.8662	77.5	0.7385
Dosulepin	21.9	0.9000	23.1	0.9442	19.2	0.7954	12	0.4219	5.3	0.1108	81.5	0.8089
Duloxetine	15.4	0.6135	25.9	0.9988	22.2	0.9123	14.9	0.5738	7.7	0.2154	86.1	0.8670
Escitalopram	2.8*	0.0296	2*	0.0085	9.2	0.2812	18.5	0.7662	21.9	0.8935	54.4	0.2248
Fluoxetine	16	0.6488	14.4	0.5527	22.5	0.9238	8.8	0.2569	20.7	0.8523	82.4	0.8089
Fluvoxamine	24.2	0.9715	8.6	0.2492	4.2	0.0638	4.9	0.0915	13.6	0.5035	55.5	0.2615
Imipramine	5.5	0.1127	17.2	0.6931	18.8	0.7723	1.8*	0.0092	9	0.2758	52.3	0.1911
Ipriflavone	17.6	0.7162	19.1	0.7908	3.2*	0.0362	21.2	0.8742	5.2	0.1069	66.3	0.4770
Maprotiline	4	0.0642	10.1	0.3192	12.6	0.4581	14.6	0.5592	14.4	0.5462	55.7	0.2615
Mianserin	23.5	0.9527	20.2	0.8427	22.2	0.9123	22.7	0.9323	24.1	0.9631	112.7	0.9983
Mirtazapine	13.5	0.5123	23.8	0.9662	3.5*	0.0431	11	0.3677	8.5	0.2508	60.3	0.3426
Nortriptyline	15.3	0.6073	19.2	0.7965	16	0.6331	22.9	0.9377	11.7	0.4058	85.1	0.8537
Pantoprazole	10.5	0.3373	6.4	0.1473	10.6	0.3492	1.8*	0.0092	25.2	0.9908	54.5	0.2248
Paroxetine	22.7	0.9285	18.8	0.7742	8.3	0.2381	18.7	0.7769	20.6	0.8500	89.1	0.9021
Reboxetine	15.4	0.6135	13.8	0.5212	13.6	0.5146	24.6	0.9812	17.3	0.7054	84.7	0.8537
Rifaximin	13.2	0.4977	1.6*	0.0050	1.7*	0.0062	10.5	0.3404	7.1	0.1862	34.1*	0.0234
Selegiline	14	0.5381	21.6	0.9004	24.2	0.9723	19.3	0.8023	19.7	0.8165	98.8	0.9727
Sertraline	5.3	0.1085	9.2	0.2804	12.1	0.4285	5.6	0.1204	11.4	0.3938	43.6	0.0878
Tranylcypromine	21.3	0.8815	19.2	0.7965	22.8	0.9354	12.7	0.4623	20.9	0.8619	96.9	0.9635
Trazodone	13.5	0.5123	15.2	0.6027	17.2	0.6954	11.2	0.3792	20.6	0.8500	77.7	0.7385
Trimipramine	2.6*	0.0250	17.6	0.7146	6	0.1327	13.1	0.4815	2.7*	0.0285	42	0.0698
Venlafaxine	5.4	0.1115	5.2	0.1004	13.8	0.5288	5.6	0.1204	13.1	0.4738	43.1	0.0784

Note. P-values were obtained through the permutation procedure described in section 5.2.3. Top ranks are marked with an asterisk and p-values < 0.05 are highlighted in gray in each cell line and the combined cell line results. The ADs (es)citalopram are mentioned in bold and control drugs are in italic and underlined.

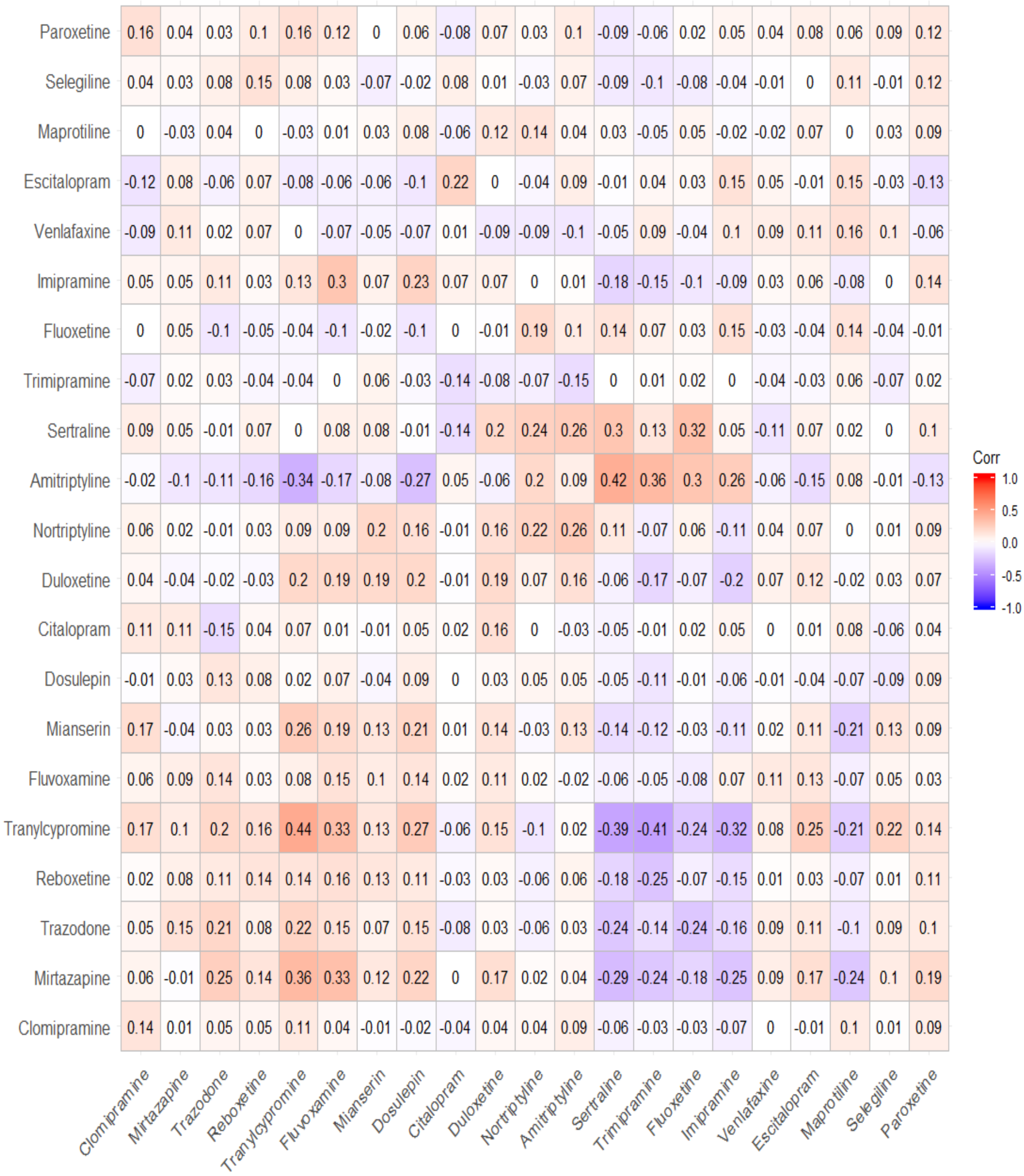
We observed that AD-induced expression profiles vary across the five analysed cell lines. Interestingly, citalopram and escitalopram have distinctive signatures, with a weak correlation between them in four cell lines (A375, MCF7, PC3, and HT29), and moderate correlation in one cell line (r=0.200 for HA1E) (Table. 5.10). The observed variability in drug-induced gene expression among cell lines likely contributes to the differences in ADs ranking across cell lines.

Table 5.10. Spearman correlation between the drug-induced expression profiles of citalopram and escitalopram in each cell line.

Cell lines	Correlation coefficient	P-value
A375	-0.008	0.405
MCF7	0.038	9.74E-05
PC3	-0.03	0.002
HT29	-0.01	0.298
HA1E	0.2	1.34E-91

Note. P-values suggesting the correlation coefficient being different from zero

For the illustration purpose, the variability of all AD-induced profiles between A375 and MCF7 cell lines are reported as a correlation matrix in Figure 5.3. Similarly, we observed small correlations between AD profiles among other cell lines.



	Clomipramine	Mirtazapine	Trazodone	Reboxetine	Tranylcypromine	Fluvoxamine	Mianserin	Dosulepin	Citalopram	Duloxetine	Nortriptyline	Amitriptyline	Sertraline	Trimipramine	Fluoxetine	Imipramine	Venlafaxine	Escitalopram	Maprotiline	Selegiline	Paroxetine
Paroxetine	0.16	0.04	0.03	0.1	0.16	0.12	0	0.06	-0.08	0.07	0.03	0.1	-0.09	-0.06	0.02	0.05	0.04	0.08	0.06	0.09	0.12
Selegiline	0.04	0.03	0.08	0.15	0.08	0.03	-0.07	-0.02	0.08	0.01	-0.03	0.07	-0.09	-0.1	-0.08	-0.04	-0.01	0	0.11	-0.01	0.12
Maprotiline	0	-0.03	0.04	0	-0.03	0.01	0.03	0.08	-0.06	0.12	0.14	0.04	0.03	-0.05	0.05	-0.02	-0.02	0.07	0	0.03	0.09
Escitalopram	-0.12	0.08	-0.06	0.07	-0.08	-0.06	-0.06	-0.1	0.22	0	-0.04	0.09	-0.01	0.04	0.03	0.15	0.05	-0.01	0.15	-0.03	-0.13
Venlafaxine	-0.09	0.11	0.02	0.07	0	-0.07	-0.05	-0.07	0.01	-0.09	-0.09	-0.1	-0.05	0.09	-0.04	0.1	0.09	0.11	0.16	0.1	-0.06
Imipramine	0.05	0.05	0.11	0.03	0.13	0.3	0.07	0.23	0.07	0.07	0	0.01	-0.18	-0.15	-0.1	-0.09	0.03	0.06	-0.08	0	0.14
Fluoxetine	0	0.05	-0.1	-0.05	-0.04	-0.1	-0.02	-0.1	0	-0.01	0.19	0.1	0.14	0.07	0.03	0.15	-0.03	-0.04	0.14	-0.04	-0.01
Trimipramine	-0.07	0.02	0.03	-0.04	-0.04	0	0.06	-0.03	-0.14	-0.08	-0.07	-0.15	0	0.01	0.02	0	-0.04	-0.03	0.06	-0.07	0.02
Sertraline	0.09	0.05	-0.01	0.07	0	0.08	0.08	-0.01	-0.14	0.2	0.24	0.26	0.3	0.13	0.32	0.05	-0.11	0.07	0.02	0	0.1
Amitriptyline	-0.02	-0.1	-0.11	-0.16	-0.34	-0.17	-0.08	-0.27	0.05	-0.06	0.2	0.09	0.42	0.36	0.3	0.26	-0.06	-0.15	0.08	-0.01	-0.13
Nortriptyline	0.06	0.02	-0.01	0.03	0.09	0.09	0.2	0.16	-0.01	0.16	0.22	0.26	0.11	-0.07	0.06	-0.11	0.04	0.07	0	0.01	0.09
Duloxetine	0.04	-0.04	-0.02	-0.03	0.2	0.19	0.19	0.2	-0.01	0.19	0.07	0.16	-0.06	-0.17	-0.07	-0.2	0.07	0.12	-0.02	0.03	0.07
Citalopram	0.11	0.11	-0.15	0.04	0.07	0.01	-0.01	0.05	0.02	0.16	0	-0.03	-0.05	-0.01	0.02	0.05	0	0.01	0.08	-0.06	0.04
Dosulepin	-0.01	0.03	0.13	0.08	0.02	0.07	-0.04	0.09	0	0.03	0.05	0.05	-0.05	-0.11	-0.01	-0.06	-0.01	-0.04	-0.07	-0.09	0.09
Mianserin	0.17	-0.04	0.03	0.03	0.26	0.19	0.13	0.21	0.01	0.14	-0.03	0.13	-0.14	-0.12	-0.03	-0.11	0.02	0.11	-0.21	0.13	0.09
Fluvoxamine	0.06	0.09	0.14	0.03	0.08	0.15	0.1	0.14	0.02	0.11	0.02	-0.02	-0.06	-0.05	-0.08	0.07	0.11	0.13	-0.07	0.05	0.03
Tranylcypromine	0.17	0.1	0.2	0.16	0.44	0.33	0.13	0.27	-0.06	0.15	-0.1	0.02	-0.39	-0.41	-0.24	-0.32	0.08	0.25	-0.21	0.22	0.14
Reboxetine	0.02	0.08	0.11	0.14	0.14	0.16	0.13	0.11	-0.03	0.03	-0.06	0.06	-0.18	-0.25	-0.07	-0.15	0.01	0.03	-0.07	0.01	0.11
Trazodone	0.05	0.15	0.21	0.08	0.22	0.15	0.07	0.15	-0.08	0.03	-0.06	0.03	-0.24	-0.14	-0.24	-0.16	0.09	0.11	-0.1	0.09	0.1
Mirtazapine	0.06	-0.01	0.25	0.14	0.36	0.33	0.12	0.22	0	0.17	0.02	0.04	-0.29	-0.24	-0.18	-0.25	0.09	0.17	-0.24	0.1	0.19
Clomipramine	0.14	0.01	0.05	0.05	0.11	0.04	-0.01	-0.02	-0.04	0.04	0.04	0.09	-0.06	-0.03	-0.03	-0.07	0	-0.01	0.1	0.01	0.09

Figure 5.3. Correlation Matrix plot between AD signatures of A375 and MCF7

5.4 Discussion

AD response is heterogeneous among MDD patients and more than 60% of patients fail to achieve remission after treatment with the first antidepressant (James et al. 2018b). Although numerous studies have advanced our understanding of the role of cytochrome P450 genes in ADs metabolism and the modulation of AD treatment outcomes (Hicks et al. 2017) (Chiara Fabbri et al. 2018), biomarkers considering sources of variability other than AD pharmacokinetics are still lacking. In this study, we evaluated an in-silico approach that can be used to prioritize ADs to treat a specific condition, utilizing gene expression profiles imputed from GWAS data and AD-induced transcriptional profiles available in CMap. In our case, the positive correlation between gene expression profiles of citalopram remitters and (es)citalopram induced expression profiles in three cell lines (A375, MCF7, and PC3) suggests that the predicted gene expression profile of a remitter is correlated with in vitro expression profiles induced by the same ADs. No previous study has tested this hypothesis and our results show that this approach might be used to rank ADs based on their likelihood of efficacy for an individual.

Generally speaking, the analysis of transcriptional profiles of drugs and disease signatures is already an established approach in the domain of drug repositioning. The studies which were based on the disease and drug profile comparison are mentioned in the section 1.2.4 of the thesis. In our study, we applied a similar strategy, but instead of disease-associated gene expression signatures we used TWAS predicted expression profiles associated with remission to a known AD drug. This way, ADs that are already available on the market and that are effective in treating the condition subject of the analysis may be identified, providing a time- and cost-effective alternative to the identification of new ADs. We indeed hypothesized that AD-induced expression profiles in vitro may correlate with gene expression profiles (predicted using GWAS data, in this case) in remitters following treatment with the same drug and/or similar drugs.

One of the most relevant observations emerging from our results is the pronounced difference in the ranking of ADs across different cell lines, suggesting the critical importance of carefully selecting the most appropriate cell line(s) when adopting this approach. The notable differences detected between cell lines can be likely explained by the inter-cellular drug-induced expression signature variability, as reported by Subramanian and colleagues. According to this study, only 15% of all the drug compounds produced highly similar signatures across multiple cell lines, whereas the vast majority (85%) produced cell line-specific signatures (Subramanian et al. 2017). The heterogeneity of drug signatures depends greatly on the cellular pathways that are particularly important in the physiological functions relevant to the cell type in question. In this study, we observed that A375 and MCF7 provided results that were more consistent with our hypothesis compared with other cell lines, for both ADs and control drugs. This may be explained by the similar embryological origin of cell lines A375 and MCF7 (skin and breast cancer cell lines, respectively), as both skin and breast cells originate from the ectoderm (outermost layer of the embryo), the same layer from which nervous tissues originate (Jiménez-Rojo et al. 2012). This hypothesis suggests that the use of brain cell lines would have been ideal in our study.

Despite the low reproducibility of gene expression profiles across cell lines, we calculated the significance of cumulative ranks across cell lines. By combining ranks of drugs in the evaluated cell lines, we aimed at the identification of ADs other than (es)citalopram – but associated with a similar treatment-induced expression profile – from which patients may benefit.

Citalopram is a racemic mixture comprised of two enantiomers, R and S-citalopram (escitalopram) in equal proportions. However, the signatures of citalopram and escitalopram are only weakly correlated in the five analysed cell lines. This can be due to the differences in modulated genes and pathways by these drugs in vitro, as reported by Sakka et al. Their study suggests that citalopram and escitalopram modulated 69 and 42 pathways, respectively, and 10 pathways were differentially modulated by the two ADs in a neuroblastoma cell line (Sakka et al. 2017). In other words, the in vitro gene expression profile of citalopram is influenced by both escitalopram and R-citalopram to a similar extent, making it different compared to the profile of escitalopram alone. On the other hand, the in vivo gene expression signature of citalopram remitters is hypothetically highly dependent on genes regulated by escitalopram rather than R-citalopram, since escitalopram has a 50-fold higher affinity for the serotonin transporter compared to R-citalopram and it is considered the main driver of the therapeutic effects of citalopram (Jacobsen et al. 2014).

It is to be noted, however, that none of the p-values relative to the correlation of rank-sum statistics across all tested cell lines would survive multiple testing correction, included that of escitalopram. As discussed above, we hypothesize the main reasons for this to lie especially in the choice of the cell line(s).

As a result, this is to be intended as an explorative analysis serving as proof-of-principle with regards to the validity of this approach, while larger studies, ideally based on brain-related cell lines, are warranted.

Irrespective of this, the presented approach is characterized by important strengths and can contribute to a deeper understanding of the genetic architecture of disease. As an example, in this case it reflected well the polygenic architecture of AD response, characterized by multiple effects of small size (Wigmore et al. 2020). Second, while RNA sequencing remains the gold standard for measuring gene expression in vivo, this approach may represent an advantageous proxy. Patients affected with psychiatric conditions and included in expression studies are for the vast majority receiving treatment, which itself is a modifier of gene expression, therefore representing a confounder. Furthermore, brain tissues can only be acquired from post-mortem samples, representing a major obstacle in performing large-scale expression studies on such patients. On the contrary, genotype-predicted gene expression profiles are not susceptible to alteration due to medications because this approach only captures the heritable component of gene expression. Furthermore, this approach is time- and cost-effective compared with any approach based on RNA sequencing and relying on tissue collection from post-mortem samples. Last, our method is computationally simple, and it can be applied to virtually any trait or condition.

There are also some limitations to the proposed methodology. First, this method can be applied only at a population level (on data from an aggregated sample of individuals), and application at the individual level would require specific adaptations (see next chapter). Second, we could not test our method in neuronal progenitor cells or differentiated neurons from the CMap transcriptional catalogue since, unfortunately, AD-induced gene expression for these cell lines was not available. A prior CMap-based study suggested that neuronal cell lines are different compared to cancer cell lines in terms of the drug expression profiles, but also showed how neuropsychiatric diseases could be reasonably modelled using cancer cell lines (Subramanian et al. 2017)(Lamb et al. 2006). However, the relevant differences between expression profiles of different cell lines found by our study and previous studies suggest that the selection of the cell line(s) most directly involved in the trait/disease of interest is of critical importance. Neural cell lines may indeed be characterized by significant expression patterns of genes taking part in those pathways that are most relevant to AD action. Additionally, gene expression signatures available in the L1000 CMap database show various challenges in terms of their analysis and usage as discussed in previous work (Musa, Tripathi, et al. 2018). Furthermore, due to the limited availability of transcriptional information for the drugs of interest across multiple cell lines, time points, and dosages, our analysis was restricted to expression profiles of 21 ADs in five cell lines treated with 10 micromolar drug concentration for 24 hours' time length.

This study indicates that there is the correlation between (es)citalopram-induced expression profiles and predicted expression associated with remission to citalopram seems to be specific to some cell lines. These limitations may, at least in part, be overcome by enhancing this approach for application at the individual level by investigating the correlation between a drug-induced expression profile and an individual’s predicted gene expression profile, which can be used to rank drugs by their predicted efficacy. Hence, the given method can be improved by considering genotype data at the individual level and using expression signatures of brain cell lines.

6. Prediction of antidepressant response using a supervised learning approach leveraging predicted gene expression and in-vitro drug-related expression profiles

6.1 Introduction

The lack of reproducible biomarkers predicting AD response is a primary challenge in depression treatment (Labermaier, Masana, and Müller 2013). Due to the heterogeneous and polygenic nature of AD response in MDD, possible research strategies to disentangle the factors modulating this phenotype are GWAS and computational models (Musker and Wong 2019) (Adam et al. 2020). Identifying optimal treatments using computational models to personalize drug response prediction may substantially improve treatment success. However, the computational task of predicting drug response holds multiple challenges due, for example, to limited data availability and implicit weaknesses of the various approaches (Adam et al. 2020).

Conventional statistical methods and machine learning tools have been used to establish drug response prediction models in both clinical and preclinical settings (Perez-Gracia et al. 2017)(Dhandapani and Goldman 2017). As described in the previous chapter of this thesis and in a recently published paper (Shoaib et al, 2020), we found a positive correlation between the predicted gene expression profiles of citalopram remission with the in-vitro (es)citalopram-induced expression profiles in various human cell types.

As a continuation of the work mentioned above, we aimed to test if our findings could be extended to develop prediction models of drug response at the individual level.

In this study, we predicted gene expression values in each STAR*D first stage participant included in GWAS of citalopram/escitalopram remission using their genotypic data and relying on results of various eQTL studies. Furthermore, we integrated these values with citalopram and other ADs-induced transcriptional responses in human cancer cell types obtained from CMap.

This project aims to develop supervised learning algorithms for predicting remission based on the combination of individual genetic profiles with the in-vitro drug expression profiles, in the hypothesis that these models could be valuable in the prediction of AD clinical efficacy.

6.2 Methods

6.2.1 Datasets

*a) STAR*D cohort*

For predicting drug response at the individual level, we used the genotyping data of STAR*D first stage participants treated with citalopram. The genotyping data was comprised of 503 citalopram remitters and 657 non-remitters. Details of STAR*D data acquisition, and its pre-processing are mentioned in Chapter 5 of this thesis (sections 5.2.1 and 5.2.2).

b) CMap drug profiles

As previously described (Chapter 5, section 5.2.3) Cmap phase II data were downloaded and 21 antidepressant profiles of 5 human cell lines were extracted. A detailed description of the CMap data we used can be found in section 5.2.3.

6.2.2 Individual-level gene expression prediction

Gene expression profiles corresponding to the remission status in all individuals of STAR*D were predicted using pre-computed gene expression and SNP weights from GTEx and other consortia derived by FUSION's authors (Gusev et al. 2016) (Figure 6.1). In each individual, imputed gene expression was calculated as the sum of SNPs alleles weighted by their known effect on gene expression (eQTL information) adjusted for linkage disequilibrium among SNPs:

$$Gene\ exp\ (g)\ = \sum_{l=1}^{m} \boldsymbol{W_{lg}X_l}$$

where W_{lg} is the expression weight for SNP *l* on gene *g* and X_l is the number of reference alleles for SNP *l*.

Individual gene expression profiles were estimated using 52 reference panels from GTEx, Common mind consortium (CMC), Netherland twin registry (NTR), and Young finns study (YFS) (Table 5.1). For individual-level prediction, gene expression values (z-scores) across all panels were selected using the criteria based on the coefficient of determination (CV-R^2) as implemented elsewhere (Pain et al. 2019). The 52 SNP-weight sets in this study contained 252,878 features, representing 26,363 unique genes. Where multiple features for a single gene were available, only the feature providing the highest cross-validation CV-R^2 was retained. We did not define any CV-R^2 threshold for feature selection also mentioned in section 5.2.3.

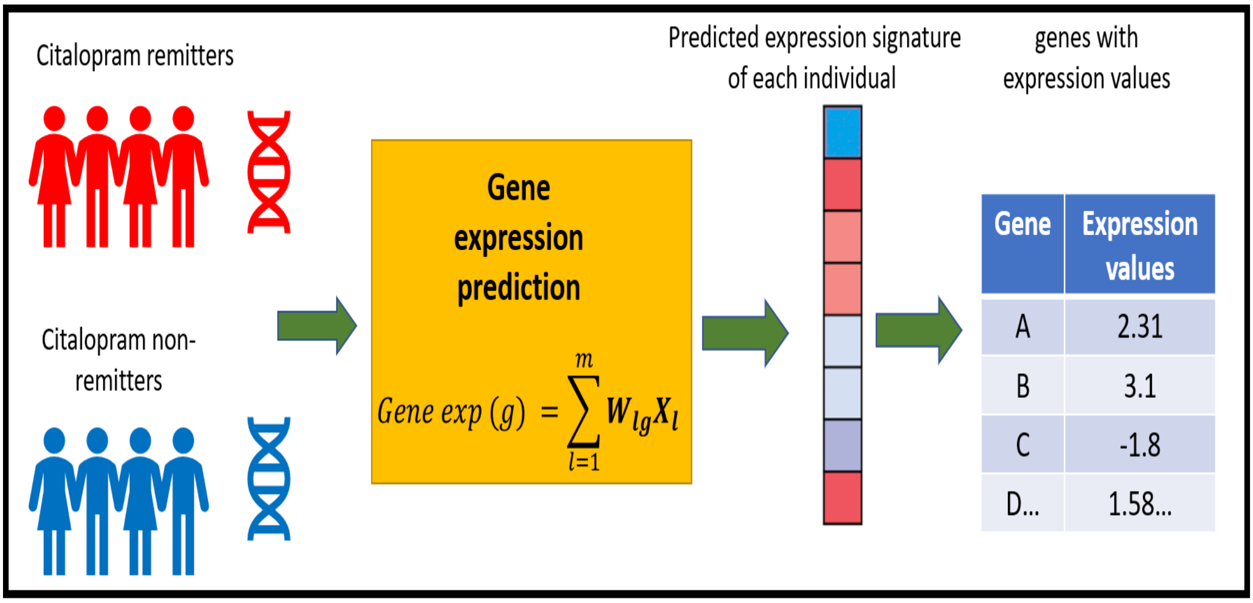

Citalopram remitters
Citalopram non-remitters
Gene expression prediction
Gene exp (g) = $\sum_{l=1}^{m} W_{lg}X_l$
Predicted expression signature of each individual
genes with expression values
Gene
Expression values
A
2.31
B
3.1
C
-1.8
D...
1.58...

Fig 6.1. Individual-level gene expression prediction using pre-computed weights from eQTL data.

6.2.3 Combination of individual gene expression and in vitro AD-induced gene expression using gene expression risk scores

We evaluated the association between the individual predicted expression profiles corresponding to citalopram remission status and *in vitro* gene expression profiles of 21 antidepressants, selected based on their availability in CMap. We considered the expression profiles of 21 ADs in 5 human cell lines available in Phase II of CMap as we did in the previous work (Chapter 5). In our previous study, we used nonparametric methods (e.g. Spearman's correlation) to test the association between remission gene expression profiles and CMap drug profiles (Chapter 5) (Shoaib et al. 2020). Here, we propose another method: we calculated gene expression risk scores (GeRS) to combine the strength and directionality of individuals' gene expression profiles with ADs-induced expression profiles. GeRS can be considered as an equivalent of polygenic risk scores, where instead of a weighted sum of SNPs, GeRS calculate a sum of gene expression values associated with a trait (citalopram remission status in this case) weighted by CMap z-scores representing drug-induced changes (Table 6.1)

The GeRS in each individual can be calculated by summing weighted gene expression values, as follows:

$$\boldsymbol{GeRS} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{g}} \boldsymbol{Gene\ Exp_j\ X\ weight_j}$$

where $Gene\ Exp_j$ is the predicted gene expression value in an individual and $weight_j$ is the z-score (absolute value) corresponding to the AD-induced expression changes of the same gene in CMap. We considered for each AD available in CMap data the top k modulated genes (k=50, 100, 250, 500).

Therefore, the GeRS can be used to combine the information provided by individual expression profiles of patients and ADs-induced expression profiles from CMap. The proposed gene-expression regulation based polygenic models have the ability to capture the polygenic nature of AD remission.

Table 6.1. Example of GeRS calculation. The higher the GeRS value for a remitter, the more similar their profile is to the imipramine-induced profile, whereas the opposite is true in the case of a non-remitter.

Genes	Expression z-scores of genes in a remitter profile	Expression z-scores of genes in a non-remitter profile	Expression z-scores of imipramine induced expression profile from CMap
SLC31A2	1.2	1.83	0.76
EBCAM	-0.85	1.7	1.85
CBR3	2.1	-1.9	1.7
GeRS of a remitter for AD imipramine = (1.2 x 0.76) + (-0.85 x 1.85) + (2.1 x -1.7) = 2.90			
GeRS of a non-remitter for AD imipramine = (1.83 x 0.76) + (1.7 x 1.85) + (-1.9 x 1.7) = 1.30			

6.2.4 Prediction of remission status using supervised learning algorithms

a) Logistic regression

Firstly, we fitted a logistic regression model using the remission status in STAR*D as the dependent variable and each of the 525 GeRS as predictors (21 ADs with 5 gene subsets [k values] across 5 cell lines available in CMap), using the R package glmnet (https://cran.rproject.org/web/packages/glmnet/index.html). We considered the top 20 population principal components as covariates to account for population structure. Nagelkerke R2 values of the model were computed with and without the population principal components in the STAR*D dataset.

b) Elastic net regression

We developed a model for the prediction of remission using a combination of all the available GeRS, firstly in each cell line and then combining all the cell lines. We used regularization with logistic regression to fit the model, to balance the risks of overfitting and excessive noise, as this approach decreases the total number of predictors and selects the most discriminative ones, other than applying a certain degree of shrinkage to the coefficients of the predictors left in the model. We trained a lasso-ridge logistic regression (elastic-net) model in STAR*D, using a 10 fold cross-validation (cv.glmnet function of the R glmnet package), using training sets for both groups of citalopram remitters and non-remitters. Negelkerke R2 values of the model were computed with and without the population principal components, as we performed for logistic regression.

6.3 Results

a) Logistic regression

For STAR*D subjects (*n =1163*), we found that GeRS corresponding to 30 CMap AD expression profiles were associated (p-value < 0.05) with either citalopram remission or non-remission status (Table 6.2) in different cell lines (HA1E, PC3, A375), with several of the ADs being TCAs. Notably, in HA1E and A375 cell lines GeRS corresponding to CMap amitriptyline-, clomipramine-, and nortriptyline-induced expression profiles were associated with citalopram non-remission. In A375 and PC3 cell lines, other TCA GeRS including dosulepin, imipramine and trimipramine showed an association with citalopram remission. After TCAs, most of the GeRS associated with citalopram remission were computed from SSRIs and tetracyclic antidepressants (TeCA) expression profiles, and they mostly showed an association with non-remission. GeRS computed using (es)citalopram expression profile were not found to be associated with remission.

b) Elastic net regression

When considering each of the tested cell lines separately, GeRS predicted citalopram remission only in the HA1E cell line (p-value = 0.002). When all the cell lines were considered together, the corresponding model showed a similar effect (p-value = 0.003) (Table. 6.3)

Table 6.2. Significant predictors (p-value < 0.05) obtained using logistic regression model. The coefficient estimates, standard error and p-values are calculated for the full model including all the covariates.

Cell	AD	Class	Subset	Full_R2 (with covariates)	Estimates	SE	P	GeRS_R2 (without covariates)
HA1E	Amitriptyline	TCA	100	0.0448	-0.2226	0.0612	0.0003	0.0149
HA1E	Amitriptyline	TCA	250	0.0389	-0.1721	0.0606	0.0045	0.0090
HA1E	Amitriptyline	TCA	10503	0.0378	-0.1595	0.0599	0.0078	0.0079
HA1E	Amitriptyline	TCA	50	0.0363	-0.1444	0.0604	0.0168	0.0064
HA1E	Amitriptyline	TCA	500	0.0354	-0.1337	0.0601	0.0260	0.0055
A375	Amitriptyline	TCA	250	0.0343	-0.1209	0.0609	0.0471	0.0044
HA1E	Clomipramine	TCA	10503	0.0351	-0.1297	0.0600	0.0307	0.0052
PC3	Dosulepin	TCA	500	0.0363	0.1460	0.0612	0.0170	0.0064
MCF7	Duloxetine	SNRIS	10503	0.0455	-0.2270	0.0609	0.0002	0.0156
MCF7	Duloxetine	SNRIS	500	0.0382	-0.1645	0.0604	0.0065	0.0083
MCF7	Duloxetine	SNRIS	250	0.0348	-0.1262	0.0601	0.0358	0.0049
HA1E	Fluoxetine	SSRI	500	0.0352	-0.1302	0.0600	0.0300	0.0053
HT29	Fluoxetine	SSRI	50	0.0350	-0.1288	0.0605	0.0333	0.0051
PC3	Fluvoxamine	SSRI	10503	0.0375	0.1577	0.0605	0.0091	0.0076
A375	Imipramine	TCA	50	0.0356	0.1366	0.0605	0.0240	0.0057
HA1E	Maprotiline	TeCA	500	0.0346	-0.1247	0.0607	0.0400	0.0047
HT29	Maprotiline	TeCA	500	0.0344	0.1205	0.0603	0.0459	0.0044
HT29	Mianserin	TeCA	500	0.0352	-0.1302	0.0599	0.0298	0.0053
HA1E	Mianserin	TeCA	250	0.0349	0.1264	0.0597	0.0344	0.0050
A375	Nortriptyline	TCA	500	0.0354	-0.1342	0.0606	0.0268	0.0055
HA1E	Nortriptyline	TCA	10503	0.0348	-0.1273	0.0605	0.0354	0.0049
HA1E	Reboxetine	NARI	10503	0.0389	-0.1706	0.0602	0.0046	0.0090
HA1E	Reboxetine	NARI	500	0.0353	-0.1323	0.0601	0.0278	0.0054
HT29	Reboxetine	NARI	250	0.0350	-0.1292	0.0606	0.0330	0.0051
HT29	Reboxetine	NARI	500	0.0343	-0.1215	0.0609	0.0460	0.0044
MCF7	Sertraline	SSRI	50	0.0354	-0.1331	0.0602	0.0271	0.0054
HA1E	Sertraline	SSRI	500	0.0351	-0.1305	0.0603	0.0305	0.0052
HA1E	Sertraline	SSRI	100	0.0348	-0.1269	0.0606	0.0363	0.0049
PC3	Tranylcypromine	MAOIs	500	0.0344	0.1208	0.0602	0.0448	0.0045
A375	Trimipramine	TCA	50	0.0356	0.1366	0.0605	0.0240	0.0057

Table 6.3. Elastic net results for individual and combined cell lines. The coefficient estimates, standard error and p-values are calculated for the full model including all the covariates.

Cell	Estimate	SE	P	Full_R2 (with covariates)	GeRS_R2 (without covariates)
A375	0.0326	0.0599	0.5863	0.0302	0.0003
HA1E	0.1864	0.0605	0.0021	0.0405	0.0106
HT29	0.0958	0.0601	0.1112	0.0327	0.0028
MCF7	0.0808	0.0605	0.1818	0.0319	0.0020
PC3	-0.0395	0.0603	0.5119	0.0304	0.0005
All	0.1749	0.0601	0.0036	0.0393	0.0094

6.4 Discussion

Predicting the clinical response to AD is a major obstacle for MDD treatment. In precision medicine, it is essential to identify biomarkers of disease signatures and match them with therapeutic interventions that are the most likely to be effective. Over the past few decades, the success of precision medicine has been witnessed for the treatment of certain somatic diseases. For instance, five-year survival in children and adolescents suffering from acute lymphatic leukaemia raised from 10% in 1990 to 90% in 2005 (Hunger et al. 2012). Despite such impact in specific diseases, we still need to explore the implications of precision medicine in neuropsychiatric diseases, as it has the potential to characterize underlying disease pathways and identify compounds that can restore the corresponding molecular targets. Hence, the focus of our second study was to predict AD remission using data from clinical trial and a supervised learning approach that is based on the genetic determinants of AD-induced expression levels and imputed individual expression profiles from STAR*D. In previous studies, MDD and drug-related gene expression profiles were compared using methods such as Pearson's correlation and Kolmogorov Smirnov statistics (Shoaib et al. 2020)(So et al. 2017), while in the present work we proposed GeRS as an alternative approach, which combines in a single metric the gene expression values in individuals treated with ADs and AD-induced gene expression in vitro. The approach we used for the implementation of polygenic models using penalized regression methods shows important strengths. The calculation of GeRS is computationally simple and easy to implement, as compared to previously reported methods, and it is widely applicable, virtually enabling analysis of any trait or disease with a genetic component. Another advantage of using GeRS in our study is that it also accounts for the polygenicity of remission trait as it incorporates the effect sizes of all expressed genes.

There are also downsides to the proposed methodology. Like our previous study, we could not test the models in neuronal progenitor cells or differentiated neurons from the CMap transcriptional catalog because of AD-induced gene expression profiles unavailability in the CMap database. In the present study, we considered individual level estimated profiles and we found TCAs to be associated with the non-remission citalopram profiles of STAR*D individuals. We couldn't find (es)citalopram among the significant predictors by means of logistic regression analysis, and in this respect these findings are not in line with our previous study (Chapter 5) in which we imputed a citalopram remission profile from GWAS summary statistics.

In our prior work, we compared the single citalopram remission profile with the in-vitro antidepressant profiles and found evidence of correlation of citalopram remission and in-vitro (es)citalopram transcriptional changes in various cell lines, whereas, in the present study we couldn't find any association between STAR*D citalopram remitters and citalopram induced expression profiles. Hence, individual-level expression profile estimation may need further evaluation for developing drug efficacy prediction models and other methods also need to be tested for the prediction of ADs at the individual level. Moreover, our findings from logistic regression analysis also suggest AD-induced expression is more similar to the genetically regulated expression in non-remitters than remitters.

Based on the results of logistic regression, TCAs emerged as the significant predictors and they seem to induce an expression profile in the single individuals (i.e. distribution of GeRS values) that is significantly correlated with that observed in citalopram non-remitters from STAR*D. This also represent the higher efficacy of TCAs in citalopram non-remitters of STAR*D studies.

Moreover, the elastic net model suggests HA1E may be the best cell line to predict AD response in MDD patients among the 5 cell lines tested in this work.

The developed framework might also be extended to the applications of machine or deep learning approaches. Support Vector machines have been also proved to provide the same results of elastic net regularization in different instances and it can be expected also to work in the present scenario.

7. Conclusion

In our first study, we tested an in-silico approach in five human cell lines by using GWAS results and drug-induced profiles to rank ADs based on their predicted efficacy. We first predicted the gene expression profile of citalopram remitters and non-remitters from the STAR*D GWAS cohort and calculated the correlation between such profiles and those induced by 21 different ADs, measured in vitro and available in CMap. This study indicates that there is a correlation between (es)citalopram-induced expression profiles and predicted expression associated with remission to citalopram in three cell lines.

In the second study, we further extended our approach and investigated the association between a drug-induced expression profile and an individual’s predicted gene expression levels using supervised learning and GeRS methods. Based on the findings of our previous study we were expecting, at the individual level, the predicted expression of (es)citalopram remitters to be generally more associated with (es)citalopram-induced expression than that of non-remitters. However, we couldn’t find associations between (es)citalopram and STAR*D citalopram remitters from logistic regression analysis when individual-level expression profiles were considered.

In summary, we suggested a framework that can be utilized in the development of clinical prediction models with the aim to contribute to the field of precision psychiatry. These prediction models can be improved by considering drug-induced expression profiles measured in brain cell types because of their direct role in antidepressant mechanism, and by replicating the proposed approach in an independent MDD patient cohort.

8. References

1000 Genome project consortium. 2005. “A Haplotype Map of the Human Genome.” *Nature* 437 (7063): 1299–1320. https://doi.org/10.1038/nature04226.

The 1000 Genomes Project Consortium., Corresponding authors., Auton, A. *et al.*2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Adam, George, Ladislav Rampášek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. 2020. “Machine Learning Approaches to Drug Response Prediction: Challenges and Recent Progress.” *Npj Precision Oncology* 4 (1): 1–10. https://doi.org/10.1038/s41698-020-0122-1.

Associations, American- Psychiatric. 2013. *American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (5th Ed.). American Journal of Psychiatry*.

Barbeira, Alvaro N., Scott P. Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E. Wheeler, Jason M. Torres, Eric S. Torstenson, et al. 2018. “Exploring the Phenotypic Consequences of Tissue Specific Gene Expression Variation Inferred from GWAS Summary Statistics.” *Nature Communications* 9 (1): 1–20. https://doi.org/10.1038/s41467-018-03621-1.

Benjamani, Yoav;, and Yosef Hochberg. 2016. “Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing” 57 (1): 289–300.

Biernacka, J. M., K. Sangkuhl, G. Jenkins, R. M. Whaley, P. Barman, A. Batzler, R. B. Altman, et al. 2015. “The International SSRI Pharmacogenomics Consortium (ISPC): A Genome-Wide Association Study of Antidepressant Treatment Response.” *Translational Psychiatry* 5 (4): 1–9. https://doi.org/10.1038/tp.2015.47.

Bradley N. Gaynes, M.D., M.P.H., M.B.A. Diane Warden, Ph.D., M.D. Madhukar H. Trivedi, Ph.D. Stephen R. Wisniewski, M.D. Maurizio Fava, and M.D A. John Rush. 2009. “What Did STAR∗D Teach Us? Results From a Large-Scale , Practical, Clinical Trial for Patients With Depression.” *Psychiatric Services* 60 (11): 1439–45.

Breen, Gerome, Qingqin Li, Bryan L Roth, Patricio O Donnell, Michael Didriksen, Ricardo Dolmetsch, Paul F O Reilly, et al. 2016. “Translating Genome-Wide Association Findings into New Therapeutics for Psychiatry” 19 (11).

Browning, Brian L., and Sharon R. Browning. 2008. “A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals.” *American Journal of Human Genetics* 84 (2): 210–23. https://doi.org/10.1016/j.ajhg.2009.01.005.

Bush, William S., and Jason H. Moore. 2012. “Chapter 11: Genome-Wide Association Studies.” *PLoS Computational Biology* 8 (12): e1002822. https://doi.org/10.1371/journal.pcbi.1002822.

Consortium, The GTEx. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6). https://doi.org/10.1038/ng.2653.

GTEx Consortium., Lead analysts, Aguet, F. *et al* 2017. “Genetic Effects on Gene Expression across Human Tissues.” *Nature* 7675 (550): 204–13. https://doi.org/10.1038/nature24277.

Cooper, Gregory M., Julie A. Johnson, Taimour Y. Langaee, Hua Feng, Ian B. Stanaway, Ute I. Schwarz, Marylyn D. Ritchie, et al. 2008. “A Genome-Wide Scan for Common Genetic Variants with a Large Influence on Warfarin Maintenance Dose.” *Blood* 112 (4): 1022–27. https://doi.org/10.1182/blood-2008-01-134247.

Corponi, Filippo. 2019. “Pharmacogenetics and Depression: A Critical Perspective.” *Psychiatry Investigation* 16 (9): 645–53. https://doi.org/10.30773/pi.2019.06.16.

Dhandapani, Muthu, and Aaron Goldman. 2017. “Preclinical Cancer Models and Biomarkers for Drug Development: New Technologies and Emerging Tools.” *Journal of Molecular Biomarkers & Diagnosis* 5 (8). https://doi.org/10.4172/2155-9929.1000356.

Dudley, Joel T., Marina Sirota, Mohan Shenoy, Reetesh K. Pai, Silke Roedder, Annie P. Chiang, Alex A. Morgan, Minnie M. Sarwal, Pankaj Jay Pasricha, and Atul J. Butte. 2011. “Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease.” *Science Translational Medicine* 3 (96): 96ra76. https://doi.org/10.1126/scitranslmed.3002648.

Euesden, Jack, Cathryn M. Lewis, and Paul F. O’Reilly. 2015. “PRSice: Polygenic Risk Score Software.” *Bioinformatics* 31 (9): 1466–68. https://doi.org/10.1093/bioinformatics/btu848.

Fabbri, Chiara. 2014. “From Pharmacogenetics to Pharmacogenomics: The Way toward the Personalization of Antidepressant Treatment.” *Canadian Journal of Psychiatry* 59 (2): 62–75. https://doi.org/10.1177/070674371405900202.

Fabbri, Chiara, Giulia Di Girolamo, and Alessandro Serretti. 2013. “Pharmacogenetics of Antidepressant Drugs: An Update after Almost 20 Years of Research.” *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics* 162 (6): 487–520. https://doi.org/10.1002/ajmg.b.32184.

Fabbri, Chiara, Katherine E. Tansey, Roy H. Perlis, Joanna Hauser, Neven Henigsberg, Wolfgang Maier, Ole Mors, et al. 2018. “Effect of Cytochrome CYP2C19 Metabolizing Activity on Antidepressant Response and Side Effects: Meta-Analysis of Data from Genome-Wide Association Studies.” *European Neuropsychopharmacology* 28 (8): 945–54. https://doi.org/10.1016/j.euroneuro.2018.05.009.

Fasipe, OlumuyiwaJohn. 2018. “Neuropharmacological Classification of Antidepressant Agents Based on Their Mechanisms of Action.” *Archives of Medicine and Health Sciences* 1 (6): 81. https://doi.org/10.4103/amhs.amhs_7_18.

Fava, Maurizio, A. John Rush, Madhukar H. Trivedi, Andrew A. Nierenberg, Michael E. Thase, Harold A. Sackeim, Frederic M. Quitkin, et al. 2003. “Background and Rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study.” *Psychiatric Clinics of North America* 26 (2): 457–94. https://doi.org/10.1016/S0193-953X(02)00107-7.

Franchini, Linda, Alessandro Serretti, Mariangela Gasperini, and Enrico Smeraldi. 1998. “Familial Concordance of Fluvoxamine Response as a Tool for Differentiating Mood Disorder Pedigrees.” *Journal of Psychiatric Research* 32 (5): 255–59. https://doi.org/10.1016/S0022-3956(98)00004-1.

Fuchsberger, Christian, Gonçalo R. Abecasis, and David A. Hinds. 2015. “Minimac2: Faster Genotype Imputation.” *Bioinformatics* 31 (5): 782–84. https://doi.org/10.1093/bioinformatics/btu704.

Gaedigk, Andrea, Magnus Ingelman-Sundberg, Neil A. Miller, J. Steven Leeder, Michelle Whirl-Carrillo, and Teri E. Klein. 2018. “The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database.” *Clinical Pharmacology and Therapeutics* 103: S67–S67. https://doi.org/10.1002/cpt.910.

Gandal, Michael J., Virpi Leppa, Hyejung Won, Neelroop N. Parikshak, and Daniel H. Geschwind. 2016. “The Road to Precision Psychiatry: Translating Genetics into Disease Mechanisms.” *Nature Neuroscience* 19 (11): 1397–1407. https://doi.org/10.1038/nn.4409.

García-González, Judit, Katherine E. Tansey, Joanna Hauser, Neven Henigsberg, Wolfgang Maier, Ole Mors, Anna Placentino, et al. 2017. “Pharmacogenetics of Antidepressant Response: A Polygenic Approach.” *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 75: 128–34. https://doi.org/10.1016/j.pnpbp.2017.01.011.

Garriock, Holly A., Jeffrey B. Kraft, Stanley I. Shyn, Eric J. Peters, Jennifer S. Yokoyama, Gregory D. Jenkins, Megan S. Reinalda, Susan L. Slager, Patrick J. McGrath, and Steven P. Hamilton. 2010.

“A Genomewide Association Study of Citalopram Response in Major Depressive Disorder.” *Biological Psychiatry* 67 (2): 133–38. https://doi.org/10.1016/j.biopsych.2009.08.029.

Gaynes, Bradley N., A. John Rush, Madhukar H. Trivedi, Stephen R. Wisniewski, Donald Spencer, and Maurizio Fava. 2008. “The STAR*D Study: Treating Depression in the Real World.” *Cleveland Clinic Journal of Medicine*. https://doi.org/10.3949/ccjm.75.1.57.

Gusev, Alexander, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W.J.H. Penninx, Rick Jansen, et al. 2016. “Integrative Approaches for Large-Scale Transcriptome-Wide Association Studies.” *Nature Genetics* 48 (3): 245–52. https://doi.org/10.1038/ng.3506.

Gusev, Alexander, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K. Finucane, Yakir Reshef, Lingyun Song, et al. 2018. “Transcriptome-Wide Association Study of Schizophrenia and Chromatin Activity Yields Mechanistic Disease Insights.” *Nature Genetics* 50 (4): 538–48. https://doi.org/10.1038/s41588-018-0092-1.

Haines, Jonathan L., and Hauser MA. 2005. “Complement Factor h Variant Increases the Risk for Early Age-Related Macular Degeneration.” *Science* 5720 (308): 419–21. https://doi.org/10.1097/IAE.0b013e318184661d.

Hicks, J. K., K. Sangkuhl, J. J. Swen, V. L. Ellingrod, D. J. Müller, K. Shimoda, J. R. Bishop, et al. 2017. “Clinical Pharmacogenetics Implementation Consortium Guideline (CPIC) for CYP2D6 and CYP2C19 Genotypes and Dosing of Tricyclic Antidepressants: 2016 Update.” *Clinical Pharmacology and Therapeutics* 102 (1): 37–44. https://doi.org/10.1002/cpt.597.

Hochberg, Yosef. 1988. “A Sharper Bonferroni Procedure for Multiple Tests of Significance.” *Biometrika* 75 (4): 800–802. https://doi.org/10.1093/biomet/75.4.800.

Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R. Abecasis. 2012. “Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-Phasing.” *Nature Genetics* 44 (8): 955–59. https://doi.org/10.1038/ng.2354.

Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. “A Flexible and Accurate Genotype Imputation Method for the next Generation of Genome-Wide Association Studies.” *PLoS Genetics* 5 (6): e1000529. https://doi.org/10.1371/journal.pgen.1000529.

Iorio, Francesco, Roberta Bosotti, Emanuela Scacheri, Vincenzo Belcastro, Pratibha Mithbaokar, Rosa Ferriero, Loredana Murino, et al. 2010. “Discovery of Drug Mode of Action and Drug Repositioning from Transcriptional Responses.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (33): 14621–26.

https://doi.org/10.1073/pnas.1000138107.

Ising, Marcus, Susanne Lucae, Elisabeth B. Binder, Thomas Bettecken, Manfred Uhr, Stephan Ripke, Martin A. Kohli, et al. 2009. “A Genomewide Association Study Points to Multiple Loci That Predict Antidepressant Drug Treatment Outcome in Depression.” *Archives of General Psychiatry* 66 (9): 966–75. https://doi.org/10.1001/archgenpsychiatry.2009.95.

James, Spencer L., Degu Abate, Kalkidan Hassen Abate, Solomon M. Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, et al. 2018. “Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 354 Diseases and Injuries for 195 Countries and Territories, 1990-2017: A Systematic Analysis for the Global Burden of Disease Study 2017.” *The Lancet*. https://doi.org/10.1016/S0140-6736(18)32279-7.

Jiang, Duo, and Miaoyan Wang. 2018. “Recent Developments in Statistical Methods for Gwas and High-Throughput Sequencing Association Studies of Complex Traits.” *Biostatistics and Epidemiology* 2 (1): 132–59. https://doi.org/10.1080/24709360.2018.1529346.

Jin, Ling, Jian Tu, Jianwei Jia, Wenbin An, Huanran Tan, Qinghua Cui, and Zhixin Li. 2014. “Drug-Repurposing Identified the Combination of Trolox C and Cytisine for the Treatment of Type 2 Diabetes.” *Journal of Translational Medicine* 12 (1): 1–7. https://doi.org/10.1186/1479-5876-12-153.

Johnstone, Andrea L., Gillian W. Reierson, Robin P. Smith, Jeffrey L. Goldberg, Vance P. Lemmon, and John L. Bixby. 2012. “A Chemical Genetic Approach Identifies Piperazine Antipsychotics as Promoters of CNS Neurite Growth on Inhibitory Substrates.” *Molecular and Cellular Neuroscience* 50 (2): 125–35. https://doi.org/10.1016/j.mcn.2012.04.008.

Kisor, David F., Carrie Hoefer, and Brian S. Decker. 2019. *Pharmacogenomics and Precision Medicine*. *Clinical Pharmacy Education, Practice and Research*. Elsevier Inc. https://doi.org/10.1016/b978-0-12-814276-9.00031-3.

Kulm, Scott, Jason Mezey, and Olivier Elemento. 2020. “Benchmarking the Accuracy of Polygenic Risk Scores and Their Generative Methods.” *MedRxiv*. https://doi.org/10.1101/2020.04.06.20055574.

Labermaier, Christiana, Mercè Masana, and Marianne B. Müller. 2013. “Biomarkers Predicting Antidepressant Treatment Response: How Can We Advance the Field?” *Disease Markers* 35 (1): 23–31. https://doi.org/10.1155/2013/984845.

Lam, Max, Swapnil Awasthi, Hunna J Watson, Jackie Goldstein, Georgia Panagiotaropoulou, Vassily

Trubetskoy, Robert Karlsson, et al. 2019. "RICOPILI: Rapid Imputation for COnsortias PIpeLIne." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz633.

Lamb, Justin, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, et al. 2006. "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease." *Science* 313 (September): 1929–35.

Leuchter, Andrew F., Ian A. Cook, Steven P. Hamilton, Katherine L. Narr, Arthur Toga, Aimee M. Hunter, Kym Faull, et al. 2010. "Biomarkers to Predict Antidepressant Response." *Current Psychiatry Reports* 12 (6): 553–62. https://doi.org/10.1007/s11920-010-0160-4.

Lewis, Cathryn M., and Evangelos Vassos. 2020. "Polygenic Risk Scores: From Research Tools to Clinical Instruments." *Genome Medicine* 12 (1): 1–11. https://doi.org/10.1186/s13073-020-00742-5.

Li, Q. S., C. Tian, G. R. Seabrook, W. C. Drevets, and V. A. Narayan. 2016. "Analysis of 23andMe Antidepressant Efficacy Survey Data: Implication of Circadian Rhythm and Neuroplasticity in Bupropion Response." *Translational Psychiatry* 6 (9): e889. https://doi.org/10.1038/tp.2016.171.

Licinio, Julio, and Ma Li Wong. 2011. "Pharmacogenomics of Antidepressant Treatment Effects." *Dialogues in Clinical Neuroscience* 13 (1): 63–71.

Lim, Nathaniel, and Paul Pavlidis. 2019. "Evaluation of Connectivity Map Shows Limited Reproducibility in Drug Repositioning." *BioRxiv*, 1–34. https://doi.org/10.1101/845693.

Lim, Sun Min, Jae Yun Lim, and Jae Yong Cho. 2014. "Targeted Therapy in Gastric Cancer: Personalizing Cancer Treatment Based on Patient Genome." *World Journal of Gastroenterology* 20 (8): 2042–50. https://doi.org/10.3748/wjg.v20.i8.2042.

Lin, Eugene, Po Hsiu Kuo, Yu Li Liu, Younger W.Y. Yu, Albert C. Yang, and Shih Jen Tsai. 2018. "A Deep Learning Approach for Predicting Antidepressant Response in Major Depression Using Clinical and Genetic Biomarkers." *Frontiers in Psychiatry* 9: 290. https://doi.org/10.3389/fpsyt.2018.00290.

Lohoff, Falk W. 2010. "Overview of the Genetics of Major Depressive Disorder." *Current Psychiatry Reports* 12 (6): 539–46. https://doi.org/10.1007/s11920-010-0150-6.

Mancuso, Nicholas, Simon Gayther, Alexander Gusev, Wei Zheng, Kathryn L. Penney, Zsofia Kote-Jarai, Rosalind Eeles, et al. 2018. "Large-Scale Transcriptome-Wide Association Study Identifies New Prostate Cancer Risk Regions." *Nature Communications* 9 (1): 1–11.

https://doi.org/10.1038/s41467-018-06302-1.

Marees, Andries T., Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. 2018. "A Tutorial on Conducting Genome-Wide Association Studies: Quality Control and Statistical Analysis." *International Journal of Methods in Psychiatric Research* 27 (2): 1–10. https://doi.org/10.1002/mpr.1608.

Martin, M. A., P. H. Westfall, and S. S. Young. 1994. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* https://doi.org/10.2307/2533464.

Musa, Aliyu, Laleh Soltan Ghoraie, Shu Dong Zhang, Galina Glazko, Olli Yli-Harja, Matthias Dehmer, Benjamin Haibe-Kains, and Frank Emmert-Streib. 2018. "A Review of Connectivity Map and Computational Approaches in Pharmacogenomics." *Briefings in Bioinformatics* 19 (3): 506–23. https://doi.org/10.1093/bib/bbw112.

Musa, Aliyu, Shailesh Tripathi, Meenakshisundaram Kandhavelu, Matthias Dehmer, and Frank Emmert-streib. 2018. "Harnessing the Biological Complexity of Big Data from LINCS Gene Expression Signatures." *PLoS ONE* 13 (8): 1–16.

Musker, Michael, and Ma-Li Wong. 2019. "Treating Depression in the Era of Precision Medicine: Challenges and Perspectives." In *Neurobiology of Depression*. https://doi.org/10.1016/b978-0-12-813333-0.00023-8.

Nassan, Malik, Wayne T. Nicholson, Michelle A. Elliott, Carolyn R. Rohrer Vitek, John L. Black, and Mark A. Frye. 2016. "Pharmacokinetic Pharmacogenetic Prescribing Guidelines for Antidepressants: A Template for Psychiatric Precision Medicine." *Mayo Clinic Proceedings*. https://doi.org/10.1016/j.mayocp.2016.02.023.

Novick, Diego, Jihyung Hong, William Montgomery, Héctor Dueñas, Magdy Gado, and Josep Maria Haro. 2015. "Predictors of Remission in the Treatment of Major Depressive Disorder: Real-World Evidence from a 6-Month Prospective Observational Study." *Neuropsychiatric Disease and Treatment* 11: 197–205. https://doi.org/10.2147/NDT.S75498.

O'dushlaine, Colm, Lizzy Rossin, Phil H. Lee, Laramie Duncan, Neelroop N. Parikshak, Stephen Newhouse, Stephan Ripke, et al. 2015. "Psychiatric Genome-Wide Association Study Analyses Implicate Neuronal, Immune and Histone Pathways." *Nature Neuroscience* 18 (2): 199–209. https://doi.org/10.1038/nn.3922.

O'Reilly, Richard L., Lisa Bogue, and Shiva M. Singh. 1994. "Pharmacogenetic Response to Antidepressants in a Multicase Family with Affective Disorder." *Biological Psychiatry* 36 (7):

467–71. https://doi.org/10.1016/0006-3223(94)90642-4.

Pain, Oliver, Andrew J. Pocklington, Peter A. Holmans, Nicholas J. Bray, Heath E. O’Brien, Lynsey S. Hall, Antonio F. Pardiñas, Michael C. O’Donovan, Michael J. Owen, and Richard Anney. 2019. “Novel Insight Into the Etiology of Autism Spectrum Disorder Gained by Integrating Expression Data With Genome-Wide Association Statistics.” *Biological Psychiatry* 86 (4): 265–73. https://doi.org/10.1016/j.biopsych.2019.04.034.

Peng, Chao Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. 2002. “An Introduction to Logistic Regression Analysis and Reporting.” *Journal of Educational Research* 96 (1): 3–14. https://doi.org/10.1080/00220670209598786.

Perez-Gracia, Jose Luis, Miguel F. Sanmamed, Ana Bosch, Ana Patiño-Garcia, Kurt A. Schalper, Victor Segura, Joaquim Bellmunt, et al. 2017. “Strategies to Design Clinical Studies to Identify Predictive Biomarkers in Cancer Research.” *Cancer Treatment Reviews* 53 (2017): 79–97. https://doi.org/10.1016/j.ctrv.2016.12.005.

Pongpanich, Monnat, Patrick F. Sullivan, and Jung Ying Tzeng. 2010. “A Quality Control Algorithm for Filtering SNPs in Genome-Wide Association Studies.” *Bioinformatics* 26 (14): 1731–37. https://doi.org/10.1093/bioinformatics/btq272.

Porcu, Eleonora, Serena Sanna, Christian Fuchsberger, and Lars G. Fritsche. 2013. “Genotype Imputation in Genome-Wide Association Studies.” *Current Protocols in Human Genetics* 1 (SUPPL.78): 1–14. https://doi.org/10.1002/0471142905.hg0125s78.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75. https://doi.org/10.1086/519795.

Rush, A John, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, et al. 2003. “The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): A Psychometric Evaluation in Patients with Chronic Major Depression.” *Biological Psychiatry* 54 (5): 573–83. https://doi.org/10.1016/S0006-3223(03)01866-8.

Sanseau, P, Agarwal, P., Barnes et al. 2012. “Use of Genome-Wide Association Studies for Drug Repositioning.” *Nature Biotechnology* 30 (4): 317–20. https://doi.org/10.1038/nbt.2151.

Serretti, A., A. Drago, and Michael N. Liebman. 2009. “Pharmacogenetics of Antidepressant

Response." *Biomarkers for Psychiatric Disorders* 3 (3): 315–53. https://doi.org/10.1007/978-0-387-79251-4_14.

Serretti, Alessandro, and P. Artioli. 2004. "The Pharmacogenomics of Selective Serotonin Reuptake Inhibitors." *Pharmacogenomics Journal* 4 (4): 233–44. https://doi.org/10.1038/sj.tpj.6500250.

Shoaib, Muhammad, Edoardo Giacopuzzi, Oliver Pain, Chiara Fabbri, Chiara Magri, Alessandra Minelli, Cathryn M. Lewis, and Massimo Gennarelli. 2020. "Investigating an in Silico Approach for Prioritizing Antidepressant Drug Prescription Based on Drug-Induced Expression Profiles and Predicted Gene Expression." *Pharmacogenomics Journal*. https://doi.org/10.1038/s41397-020-00186-5.

Sirota, Marina, Joel T. Dudley, Jeewon Kim, Annie P. Chiang, Alex A. Morgan, Alejandro Sweet-Cordero, Julien Sage, and Atul J. Butte. 2011. "Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data." *Science Translational Medicine*. https://doi.org/10.1126/scitranslmed.3001318.

Siu, Fung Ming, Dik Lung Ma, Yee Wai Cheung, Chun Nam Lok, Kun Yan, Zhiqi Yang, Mengsu Yang, et al. 2008. "Proteomic and Transcriptomic Study on the Action of a Cytotoxic Saponin (Polyphyllin D): Induction of Endoplasmic Reticulum Stress and Mitochondria-Mediated Apoptotic Pathways." *Proteomics* 8 (15): 3105–17. https://doi.org/10.1002/pmic.200700829.

So, Hon Cheong, Carlos Kwan Long Chau, Wan To Chiu, Kin Sang Ho, Cho Pong Lo, Stephanie Ho Yue Yim, and Pak Chung Sham. 2017. "Analysis of Genome-Wide Association Data Highlights Candidates for Drug Repositioning in Psychiatry." *Nature Neuroscience* 20 (10): 1342–49. https://doi.org/10.1038/nn.4618.

Stegle, Oliver, Leopold Parts, Richard Durbin, and John Winn. 2010. "A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in EQTL Studies." *PLoS Computational Biology* 6 (5): 1–11. https://doi.org/10.1371/journal.pcbi.1000770.

Subramanian, Aravind, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, et al. 2017. "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles." *Cell* 171 (6): 1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of* *the National Academy of Sciences of the United States of America* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

Tansey, Katherine E., Michel Guipponi, Xiaolan Hu, Enrico Domenici, Glyn Lewis, Alain Malafosse, Jens R. Wendland, Cathryn M. Lewis, Peter McGuffin, and Rudolf Uher. 2013a. "Contribution of Common Genetic Variants to Antidepressant Response." *Biological Psychiatry* 73 (7): 679–82. https://doi.org/10.1016/j.biopsych.2012.10.030.

——2013b. "Contribution of Common Genetic Variants to Antidepressant Response." *Biological Psychiatry* 73 (7): 679–82. https://doi.org/10.1016/j.biopsych.2012.10.030.

Tansey, Katherine E., Michel Guipponi, Nader Perroud, Guido Bondolfi, Enrico Domenici, David Evans, Stephanie K. Hall, et al. 2012. "Genetic Predictors of Response to Serotonergic and Noradrenergic Antidepressants in Major Depressive Disorder: A Genome-Wide Analysis of Individual-Level Data and a Meta-Analysis." *PLoS Medicine* 9 (10). https://doi.org/10.1371/journal.pmed.1001326.

Trivedi, Madhukar H., A. John Rush, Stephen R. Wisniewski, Andrew A. Nierenberg, Diane Warden, Louise Ritz, Grayson Norquist, et al. 2006. "Evaluation of Outcomes with Citalopram for Depression Using Measurement-Based Care in STAR*D: Implications for Clinical Practice." *American Journal of Psychiatry* 163 (1): 28–40. https://doi.org/10.1176/appi.ajp.163.1.28.

Tsuchimine, Shoko, Shinichiro Ochi, Misuzu Tajiri, Yutaro Suzuki, Norio Sugawara, Yoshimasa Inoue, and Norio Yasui-Furukori. 2018. "Effects of Cytochrome P450 (CYP) 2C19 Genotypes on Steady-State Plasma Concentrations of Escitalopram and Its Desmethyl Metabolite in Japanese Patients with Depression." *Therapeutic Drug Monitoring* 40 (3): 356–61. https://doi.org/10.1097/FTD.0000000000000506.

Uher, Rudolf, Nader Perroud, Mandy Y.M. Ng, Joanna Hauser, Neven Henigsberg, Wolfgang Maier, Ole Mors, et al. 2010. "Genome-Wide Pharmacogenetics of Antidepressant Response in the GENDEP Project." *American Journal of Psychiatry* 167 (5): 555–64. https://doi.org/10.1176/appi.ajp.2009.09070932.

Uher, Rudolf, Katherine E. Tansey, Marcella Rietschel, Neven Henigsberg, Wolfgang Maier, Ole Mors, Joanna Hauser, et al. 2013. "Common Genetic Variation and Antidepressant Efficacy in Major Depressive Disorder: A Meta-Analysis of Three Genome-Wide Pharmacogenetic Studies." *American Journal of Psychiatry* 170 (2): 207–17. https://doi.org/10.1176/appi.ajp.2012.12020237.

Wainberg, Michael, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N. Barbeira, David A. Knowles, David Golan, Raili Ermel, et al. 2019. “Opportunities and Challenges for Transcriptome-Wide Association Studies.” *Nature Genetics* 51 (4): 592–99. https://doi.org/10.1038/s41588-019-0385-z.

Wang, Maggie Haitian, Heather J. Cordell, and Kristel Van Steen. 2019. “Statistical Methods for Genome-Wide Association Studies.” *Seminars in Cancer Biology* 55 (April 2018): 53–60. https://doi.org/10.1016/j.semcancer.2018.04.008.

Ward, Joey, Nicholas Graham, Rona J. Strawbridge, Amy Ferguson, Gregory Jenkins, Wenan Chen, Karen Hodgson, et al. 2018. “Polygenic Risk Scores for Major Depressive Disorder and Neuroticism as Predictors of Antidepressant Response: Meta-Analysis of Three Treatment Cohorts.” *PLoS ONE* 13 (9): e0203896. https://doi.org/10.1371/journal.pone.0203896.

Whooley, Mary A, and Wong. 2003. “Depression and Cardiovascular Disorders.” *Journal of Psychiatry & Neuroscience : JPN* 23 (3): 180–81. https://doi.org/10.1146/annurev-clinpsy-050212-185526.

Wigmore, Eleanor M., Jonathan D. Hafferty, Lynsey S. Hall, David M. Howard, Toni Kim Clarke, Chiara Fabbri, Cathryn M. Lewis, et al. 2020. “Genome-Wide Association Study of Antidepressant Treatment Resistance in a Population-Based Cohort Using Health Service Prescription Data and Meta-Analysis with GENDEP.” *Pharmacogenomics Journal* 20 (2): 329–41. https://doi.org/10.1038/s41397-019-0067-3.

Wise, L. H., J. S. Lanchbury, and C. M. Lewis. 1999. “Meta-Analysis of Genome Searches.” *Annals of Human Genetics* 63 (3): 263–72. https://doi.org/10.1017/S0003480099007526.

Wray, Naomi R., Sang Hong Lee, Divya Mehta, Anna A.E. Vinkhuyzen, Frank Dudbridge, and Christel M. Middeldorp. 2014. “Research Review: Polygenic Methods and Their Application to Psychiatric Traits.” *Journal of Child Psychology and Psychiatry and Allied Disciplines* 55 (10): 1068–87. https://doi.org/10.1111/jcpp.12295.

Wray, Naomi R., Stephan Ripke, Manuel Mattheisen, MacIej Trzaskowski, Enda M. Byrne, Abdel Abdellaoui, Mark J. Adams, et al. 2018. “Genome-Wide Association Analyses Identify 44 Risk Variants and Refine the Genetic Architecture of Major Depression.” *Nature Genetics* 50 (5): 668–81. https://doi.org/10.1038/s41588-018-0090-3.

Yun, Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. 2009. “Genotype Imputation.” *Annual Review of Genomics and Human Genetics* 10: 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242.

Zhou, Wei, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, et al. 2018. “Efficiently Controlling for Case-Control Imbalance and Sample Relatedness in Large-Scale Genetic Association Studies.” *Nature Genetics* 50 (9): 1335–41. https://doi.org/10.1038/s41588-018-0184-y.

Zwicker, Alyson, Chiara Fabbri, Marcella Rietschel, Joanna Hauser, Ole Mors, Wolfgang Maier, Astrid Zobel, et al. 2018. “Genetic Disposition to Inflammation and Response to Antidepressants in Major Depressive Disorder.” *Journal of Psychiatric Research* 105 (June): 17–22. https://doi.org/10.1016/j.jpsychires.2018.08.011.

9. Details of scientific activities performed during Ph.D.

1) International mobility

Erasmus plus traineeship at the Institute of Psychiatry, Psychology, and Neuroscience, King's College London from October 2019 to July 2020 (ten months).

2) List of Conference abstracts

(i) Shoaib, M., Giacopuzzi, E., Magri, C., Minelli, A., & Gennarelli, M. (2019, October). Genomic restricted maximum likelihood (GREML) analysis to estimate the heritability of response/ resistance in major depressive disorder (MDD). In EUROPEAN JOURNAL OF HUMAN GENETICS (Vol. 27, pp. 1677-1677). Poster presented at 52nd ESHG Conference, 15-18 June 2019, Gothenburg, Sweden.

(ii) Shoaib, M., Giacopuzzi, E., Pain, O. et al, Towards Precision Psychiatry: A Data-Driven Strategy for Antidepressant Drug Prescription Using Transcriptome-Wide Association Study and Connectivity Map Approach. In the 28th World Congress of Psychiatric Genetics (WCPG). Poster presented at Virtual world congress of Psychiatric Genetics, 19-21 Oct 2020.

3) Peer-reviewed published article

(i) Shoaib, M., Giacopuzzi, E., Pain, O. et al. Investigating an in-silico approach for prioritizing antidepressant drug prescription based on drug-induced expression profiles and predicted gene expression. Pharmacogenomics J (2020). https://doi.org/10.1038/s41397-020-00186-5 Full article available at https://rdcu.be/b7EWB.

4) Workshops/Courses

(i) Attended 'Machine Learning for Health and Bioinformatics' course lectures at the Department of Biostatistics and Health Informatics, King's College London. Dates: 11/05/2020 to 15/05/2020

(ii) Attended the workshop on 'Genetic Association studies' at the Institute of Psychiatry, Psychology, and Neuroscience, King's College London. Dates: 15/06/20202 to 16/06/2020

10. Appendix

Published paper

Shoaib, M., Giacopuzzi, E., Pain, O. *et al.* Investigating an in-silico approach for prioritizing antidepressant drug prescription based on drug-induced expression profiles and predicted gene expression. *Pharmacogenomics J* (2020). https://doi.org/10.1038/s41397-020-00186-5

The Pharmacogenomics Journal
https://doi.org/10.1038/s41397-020-00186-5

ARTICLE

Investigating an in silico approach for prioritizing antidepressant drug prescription based on drug-induced expression profiles and predicted gene expression

Muhammad Shoaib[1,2] · Edoardo Giacopuzzi[3,4] · Oliver Pain[2] · Chiara Fabbri[2] · Chiara Magri[1] · Alessandra Minelli[1] · Cathryn M. Lewis[2,5] · Massimo Gennarelli[1,4]

Received: 21 April 2020 / Revised: 13 August 2020 / Accepted: 8 September 2020

Abstract

In clinical practice, an antidepressant prescription is a trial and error approach, which is time consuming and discomforting for patients. This study investigated an in silico approach for ranking antidepressants based on their hypothetical likelihood of efficacy. We predicted the transcriptomic profile of citalopram remitters by performing an in silico transcriptomic-wide association study on STAR*D GWAS data ($N = 1163$). The transcriptional profile of remitters was compared with 21 antidepressant-induced gene expression profiles in five human cell lines available in the connectivity-map database. Spearman correlation, Pearson correlation, and the Kolmogorov–Smirnov test were used to determine the similarity between antidepressant-induced profiles and remitter profiles, subsequently calculating the average rank of antidepressants across the three methods and a p value for each rank by using a permutation procedure. The drugs with the top ranks were those having a high positive correlation with the expression profiles of remitters and that may have higher chances of efficacy in the tested patients. In MCF7 (breast cancer cell line), escitalopram had the highest average rank, with an average rank higher than expected by chance ($p = 0.0014$). In A375 (human melanoma) and PC3 (prostate cancer) cell lines, escitalopram and citalopram emerged as the second-highest ranked antidepressants, respectively ($p = 0.0310$ and 0.0276, respectively). In HA1E (kidney) and HT29 (colon cancer) cell types, citalopram and escitalopram did not fall among top antidepressants. The correlation between citalopram remitters' and (es)citalopram-induced expression profiles in three cell lines suggests that our approach may be useful and with future improvements, it can be applicable at the individual level to tailor treatment prescription.

Introduction

Major depressive disorder (MDD) is a primary health issue and the third leading cause of disability in adolescents and young adults, while the second leading cause of disability in middle-aged adults worldwide. Globally, more than 264 million people of different age groups are living with depression [1]. This heavy disease burden is partly due to the complex pathogenic mechanisms of

These authors contributed equally: Edoardo Giacopuzzi, Oliver Pain, Chiara Fabbri

These authors jointly supervised this work: Cathryn M. Lewis, Massimo Gennarelli

Supplementary information The online version of this article (https://doi.org/10.1038/s41397-020-00186-5) contains supplementary material, which is available to authorized users.

✉ Cathryn M. Lewis
cathryn.lewis@kcl.ac.uk

[1] Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy

[2] Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, UK

[3] National Institute for Health Research (NIHR), Oxford Biomedical Research Centre, Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

[4] IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy

[5] Department of Medical and Molecular Genetics, Faculty of Life Sciences and Medicine, King's College London, London, UK

Published online: 17 September 2020

MDD, the interindividual heterogeneity of antidepressant response, and the lack of reliable response predictors [2].

Antidepressant (AD) choice in MDD is based on prescription guidelines and prior clinical experience, but the lack of reproducible predictors of AD response makes it a "trial and error" approach which can take up to several weeks or months and a number of treatment changes before symptom remission is achieved. The availability of objective and reproducible predictors of AD response could reduce the time needed to achieve remission and relieve patients' suffering [3]. Prior studies suggest that AD response and remission are heritable traits [4], offering the opportunity to use genetic markers to develop predictors applicable in clinical practice to guide drug prescription. The combination of clinical presentation, genomic information, and metabolic characteristics was indeed suggested as a possible strategy for the development of precision psychiatry [5].

The purpose of this study was to develop a new approach aiming to contribute to precision psychiatry. Previous studies have focused on the identification of genetic variants associated with AD efficacy [6, 7], and here we expand the focus to transcriptomic profiles derived from transcriptomic-wide association studies (TWAS) based on GWAS summary statistics. Transcriptomic profiles associated with the efficacy of specific ADs in clinical trials can be compared with the in vitro AD-induced gene expression changes, in order to test if drug-induced gene expression signatures could be used as markers of clinical efficacy of specific ADs. The main aim of this study was to develop and test this approach by computing gene expression profile associated with remission to citalopram in the sequence treatment alternative to relieve depression (STAR*D) study and comparing this profile with citalopram and other ADs induced transcriptional responses available from the connectivity-map (CMap) database. CMap is a genome-scale library of cellular signatures and catalog of transcriptional responses to chemical and genetic perturbations [8]. A positive correlation between expression profiles of citalopram remission and in vitro citalopram-induced gene changes was hypothesized to be indicative of the potential utility of our approach. We also hypothesized that the same would be true for escitalopram since it is the therapeutically active enantiomer of citalopram [9]. Secondly, we applied the same approach adding control drugs (no known AD effect) to ADs to provide a proof of principle of the usefulness of the method, since control drugs were expected to have less similarity to citalopram remission gene expression profile than (es)citalopram and other ADs.

Methods

Study population

This study is based on a STAR*D data [10]. The STAR*D study is a trial of protocol-guided AD treatment for outpatients with MDD. The study included 4041 treatment-seeking adult outpatients, recruited in 18 primary care and 23 psychiatric clinical sites across the United States. Genotyping was performed in 1948 participants [11]. Our analysis used data from the first treatment step (level 1), which consisted of protocol-guided citalopram (20–60 mg/day). Remission was defined as a score < 6 on the quick inventory of depressive symptomatology clinician-rated (QIDS-C) scale at level 1 exit (after 12 weeks of citalopram treatment), in line with the previous literature [12, 13]. We chose remission rather than symptom improvement as it was associated with disease prognosis and lower risk of relapse in STAR*D participants, therefore it is considered the main goal of AD treatment [14]. STAR*D genotype and phenotype data are available through the National Institute of Mental Health Human Genetics Initiative (https://www.nimhgenetics.org/). Further details about the STAR*D study are available in the Supplementary Material (Section 1).

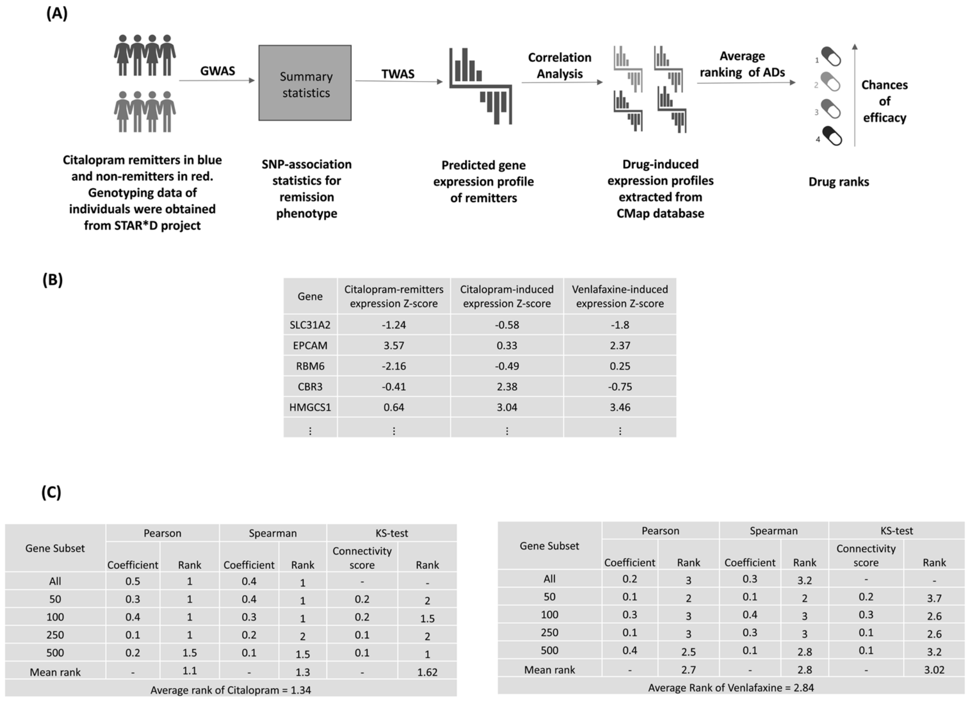
Genotyping, quality control, and imputation

Details on the genotyping procedure can be found elsewhere [11]. Individual genotype data was processed using the Psychiatric Genomics Consortium "RICOPILI" pipeline for standardized quality control and imputation [15]. Imputation of SNPs and insertion–deletion polymorphisms was performed using the 1000 Genomes Project multi-ancestry reference panel (see Supplementary Material, Section 2).

Statistical analysis

Genome-wide association study (GWAS)

A GWAS was conducted using the RICOPLI pipeline to test the association of each SNP with remission to citalopram, classifying STAR*D participants ($N = 1163$) as remitters or non remitters. The logistic regression analysis included the covariates of sex, age, baseline QIDS-C score, and the first 20 population principal components. The GWAS summary statistics were then converted to LD-score regression format using the munge_sumstats.sh script, removing SNPs with an INFO < 0.3 (https://github.com/bulik/ldsc/wiki).


(A)
Citalopram remitters in blue and non-remitters in red. Genotyping data of individuals were obtained from STAR*D project
GWAS
Summary statistics
SNP-association statistics for remission phenotype
TWAS
Predicted gene expression profile of remitters
Correlation Analysis
Drug-induced expression profiles extracted from CMap database
Average ranking of ADs
1
2
3
4
Chances of efficacy
Drug ranks
(B)
(C)

Gene	Citalopram-remitters expression Z-score	Citalopram-induced expression Z-score	Venlafaxine-induced expression Z-score
SLC31A2	-1.24	-0.58	-1.8
EPCAM	3.57	0.33	2.37
RBM6	-2.16	-0.49	0.25
CBR3	-0.41	2.38	-0.75
HMGCS1	0.64	3.04	3.46
⋮	⋮	⋮	⋮

Gene Subset	Pearson		Spearman		KS-test	
	Coefficient	Rank	Coefficient	Rank	Connectivity score	Rank
All	0.5	1	0.4	1	-	-
50	0.3	1	0.4	1	0.2	2
100	0.4	1	0.3	1	0.2	1.5
250	0.1	1	0.2	2	0.1	2
500	0.2	1.5	0.1	1.5	0.1	1
Mean rank	-	1.1	-	1.3	-	1.62
Average rank of Citalopram = 1.34						

Gene Subset	Pearson		Spearman		KS-test	
	Coefficient	Rank	Coefficient	Rank	Connectivity score	Rank
All	0.2	3	0.3	3.2	-	-
50	0.1	2	0.1	2	0.2	3.7
100	0.3	3	0.4	3	0.3	2.6
250	0.1	3	0.3	3	0.1	2.6
500	0.4	2.5	0.1	2.8	0.1	3.2
Mean rank	-	2.7	-	2.8	-	3.02
Average Rank of Venlafaxine = 2.84						

Fig. 1 Illustration of antidepressants ranking method using data from STAR*D and Connectivity-Map (CMap). a Major steps of the proposed in silico method. **b** *Z*-scores of differentially expressed genes of citalopram remitters, venlafaxine, and citalopram-induced expression profiles from CMap. **c** Description of average rank calculation method using Pearson, Spearman correlation, and KS method.

Transcriptome-wide association study (TWAS)

We used STAR*D GWAS summary statistics to perform a TWAS using FUSION software [16]. Briefly, FUSION requires precomputed gene expression SNP weights and GWAS summary statistics to predict the association between the expression of each gene and the phenotype of interest. SNP weights from CommonMind Consortium dorsolateral prefrontal cortex (DLPFC), 48 tissues including 13 brain regions within the Genotype-Tissue Expression (GTEx) consortia, Young Finn study, Netherland twin registry, and Metabolic syndrome in men study datasets were considered (Supplementary Table S1). The SNP weights were previously derived by fusion authors [16–18]. As possible confounding factors were considered when estimating the SNP weights by including known and hidden covariates [19]. Therefore, we assumed medication usage in donors of the above-mentioned data sources is unlikely to have had an impact on the SNP weights. From the GTEx database, we considered a wide range of tissues in addition to brain regions because of their larger sample sizes and the presence of moderate correlation of *cis*-eQTL effects among different tissues [20]. All gene expression SNP-weights were downloaded from the FUSION website (http://gusevlab.org/projects/fusion/). This study uses the term *SNP-weight sets* to define SNP weights from a given sample and tissue (e.g., GTEx hippocampus, CMC DLPFC). Furthermore, each gene within a given SNP-weight set constitutes a *feature* or *gene–tissue pair*. We combined the FUSION output for all SNP-weight sets, using the TWAS associations (*z*-scores) to represent the gene expression signature of citalopram remitters. The 52 SNP-weight sets in this study contained 252,878 features, representing 26,363 unique genes. Where multiple features for a single gene were available, only the feature providing the highest cross-validation coefficient of determination (CV-R^2) was retained. We did not define any CV-R^2 threshold for feature selection. Similar criteria have been implemented elsewhere [21].

Comparison of TWAS results with in vitro AD-induced gene expression

We evaluated the correlation between the TWAS expression profile of citalopram remission and in vitro gene expression profiles of 21 ADs, selected based on their availability in CMap (Phase II data) (Fig. 1a). CMap is a publicly available comprehensive library of transcriptional expression data obtained using L1000 assay, which directly measures or infers the expression levels of 12,328 genes (https://www.broadinstitute.org/connectivity-map-cmap). The database contains L1000 profiles from various perturbating agents (small molecule compounds, shRNAs, cDNAs, and biologics). More specifically, the CMap platform provides the transcriptomic information of human cultured cell lines exposed to compounds obtained from various screening libraries including drugs approved by the FDA [22].

We considered the expression profiles of 21 ADs (Supplementary Table S2) in five human cell lines available in

Phase II of CMap ((*a) A375, human malignant melanoma (b) MCF7, breast cancer (c) PC3, prostate cancer (d) HA1E, kidney (e) HT29, colon cancer)*. Drug-induced expression profiles were evaluated in cells treated for 24 h with 10 µm drug concentration. We used CMap's GEO series (GSE70138) data and extracted relevant expression profiles using cmapR package. Of the 12,328 genes within the CMap profiles, 10,027 were captured by the SNP-weight included in the citalopram remitter TWAS. We compared the expression profiles of the 21 ADs with the profile of citalopram remitters obtained from the TWAS using an approach described in a previous study [23]. The differentially expressed genes represented in terms of *z*-scores of citalopram remitters and drug-induced profiles (Fig. 1b) were analyzed using R code (https://sites.google.com/site/honcheongso/software/gwascmap), according to the following procedure:

a. *Evaluating the relationship between AD-induced gene expression and expression profiles of citalopram remitters*. Patterns of expressions were tested by analyzing all and the strongest upregulated and downregulated genes in the TWAS ($k = 50$, 100, 250, 500). The correlation between CMap AD profiles and the STAR*D remitter profile was assessed for each drug using Spearman's correlation and Pearson's correlation using all and highly modulated remitter's *k* genes. We adopted the KS test as reported by the original CMap study [24] to compare the expression patterns of AD and citalopram remitters by considering strongly up and downregulated *k* genes and calculated connectivity scores [8]. The 21 tested ADs were ranked based on the results of each test (Pearson, Spearman, and KS), and then the average rank across tests for each drug were computed. Drugs were ranked in ascending order of their correlation results (the drug with the most positive correlation was ranked first) (Fig. 1c).
b. *Significance of ranks using permutation*. To estimate the significance of the ranks, a one-sided permutation procedure was performed by shuffling the *z*-scores obtained in the TWAS and calculating the corresponding rank of each drug by repeating the procedure in step a. One hundred permutations were performed to calculate the distribution of ranks under the null hypothesis and estimated the *p* value of the observed ranks.
c. *Calculation of ranks probability for each AD across cell lines using Genome Scan Meta-Analysis (GSMA) method*. We combined ranks of each AD in five cell lines by adding them and calculated the sum of ranks probability using GSMA, a nonparametric method for meta-analyzing ranks [25].

Finally, we repeated the process described above in (a), (b), and (c) for five control drugs (Supplementary Table S3) having hypothetically no AD effect to validate the proposed method. The major steps of the applied in silico method are shown in Fig. 1a.

Results

STAR*D data included 506 citalopram remitters and 657 non remitters with genotypic data after quality control, and the main clinical–demographic characteristics are shown in (Supplementary Table S4). The GWAS and TWAS Q-Q plots showed no evidence of confounding (Supplementary Figs. S1, S2).

The average rank across tests for ADs showed that escitalopram (S-enantiomer of citalopram) was the AD with the highest average rank followed by amitriptyline in MCF7 (breast cancer cell line). In A375 (human malignant melanoma) and PC3 (prostate cancer) cell lines, escitalopram and citalopram emerged as the second-highest ranked ADs, respectively, after trimipramine and mirtazapine, respectively. In HT29 (colon cancer) cell line, citalopram ranked third after trimipramine and dosulepin. Imipramine and fluvoxamine were the top-ranked ADs in HA1E (kidney) cell line, whereas escitalopram and citalopram did not fall in the top ranks in this cell type. In the analysis of combined ranks across cell lines, we found sertraline, trimipramine, and venlafaxine as drugs with the best sum of ranks and *p* values < 0.05, while citalopram was right after them and close to the significance threshold ($p = 0.057$) (Table 1). In summary, despite relevant differences of ADs ranking among cell lines, (es)citalopram-induced expression signatures were found to have a significant correlation with a citalopram remission profile in three cell lines (A375, MCF7, and PC3).

We also attempted to validate our approach using the expression profiles of five control drugs. In A375 all control drugs ranked after (es)citalopram. In PC3, three control drugs ranked after (es)citalopram. In MCF7, four control drugs were ranked after escitalopram and one control drug was ranked after citalopram. In HT29 and HA1E, four control compounds ranked before the (es)citalopram. In the combined rank analysis, rifaximin, trimipramine and clofibrate had the top sum ranks across cell lines, though only rifaximin was nominally significant ($p < 0.05$) (Table 2). Our hypothesis of higher correlation of (es)citalopram-induced gene expression with citalopram remission TWAS results compared to control drugs was partially confirmed only in two cell lines (A375 and in MCF7).

We observed that AD-induced expression profiles vary across the five analyzed cell lines. Interestingly, citalopram

Table 1 Ranking of ADs in five human cell lines.

Cell lines	A375		MCF7		PC3		HA1E		HT29		Combined cell lines	
Drug	Rank	Perm. *p* value	Rank	Perm. *p* value	Rank	Perm. *p* value	Rank	Perm. *p* value	Rank	Perm. *p* value	Rank sum	*p* value
Amitriptyline	16.9	0.8610	2.4*	*0.0290*	12.6	0.6138	17.6	0.8948	12.9	0.6224	62.4	0.7049
Citalopram	7.6	0.2862	7.7	0.2890	2.3*	*0.0276*	10.6	0.4776	4.4	0.1114	32.6	0.0571
Clomipramine	11.1	0.5229	4.1	0.0890	9.1	0.3714	18.6	0.9371	16.8	0.8590	59.7	0.6533
Dosulepin	17.3	0.8824	18.1	0.9257	15	0.7605	8.8	0.3571	2.6*	*0.0405*	61.8	0.7049
Duloxetine	11	0.5181	20.9	0.9971	17.2	0.8776	11.3	0.5305	4.8	0.1300	65.2	0.7753
Escitalopram	2.6*	*0.0310*	1.2*	*0.0014*	6.9	0.2395	14.5	0.7324	17.6	0.8976	42.8	0.2034
Fluoxetine	13.8	0.6843	10.8	0.4890	17.8	0.9033	6.2	0.1976	16.7	0.8562	65.3	0.7752
Fluvoxamine	19.6	0.9729	6.2	0.2010	2.9	0.0443	3.2	0.0529	9.6	0.4005	41.5	0.1833
Imipramine	6.2	0.2014	12.6	0.6052	14.6	0.7395	1.2*	*0.0062*	5.2	0.1490	39.8	0.1468
Maprotiline	3.7	0.0667	6.8	0.2362	9.8	0.4200	11.2	0.5229	10.5	0.4590	42	0.1833
Mianserin	19.6	0.9729	15.6	0.8029	17.8	0.9033	17.9	0.9062	19.9	0.9795	90.8	0.9979
Mirtazapine	9.8	0.4362	18.8	0.9529	2*	*0.0181*	7.7	0.2871	5.2	0.1490	43.5	0.2247
Nortriptyline	11.6	0.5576	15.2	0.7833	12	0.5748	18.1	0.9167	7.9	0.2990	64.8	0.7752
Paroxetine	18.3	0.9248	14.4	0.7390	6.4	0.2129	14.7	0.7448	16.7	0.8562	70.5	0.8695
Reboxetine	11.8	0.5681	10.2	0.4476	10.7	0.4895	19.8	0.9800	13.3	0.6552	65.8	0.7965
Selegiline	10.6	0.4890	17	0.8667	19.2	0.9581	15	0.7671	15.6	0.8005	77.4	0.9512
Sertraline	4.4	0.1033	6.3	0.2052	9.1	0.3714	3.6	0.0662	7.4	0.2743	30.8*	*0.0412*
Tranylcypromine	16.9	0.8610	14.8	0.7619	18	0.9119	9.5	0.3929	16.9	0.8614	76.1	0.9428
Trazodone	10.3	0.4657	10.9	0.4943	13	0.6424	8	0.3043	16.6	0.8529	58.8	0.6264
Trimipramine	2.4*	*0.0267*	13.4	0.6614	4	0.0919	9.7	0.4086	1.4*	*0.0105*	30.9*	*0.0412*
Venlafaxine	5.5	0.1567	3.6	0.0676	10.6	0.4819	3.8	0.0810	9	0.3657	32.5*	*0.0487*

P values were obtained through the permutation procedure described in Statistical analysis paragraph. Top ranks are marked with an asterisk and *p* values < 0.05 are highlighted in italic in each cell line and the combined cell line results. The ADs (es)citalopram are mentioned in bold.

and escitalopram have distinctive signatures, with a weak correlation between them in four cell lines (A375, MCF7, PC3, and HT29), and moderate correlation in one cell line ($r = 0.200$ for HA1E) (Table 3). The observed variability in drug-induced gene expression among cell lines likely contributes to the differences in ADs ranking across cell lines. The variability of all AD-induced profiles across cell lines are reported as correlation matrices in the Supplementary File (Supplementary Figs. S3–12).

Discussion

AD response is heterogeneous among MDD patients and more than 60% of patients fail to achieve remission after treatment with the first AD [1]. Although numerous studies have advanced our understanding of the role of cytochrome P450 (CYP450) genes in ADs metabolism and the modulation of AD treatment outcomes [26, 27], biomarkers considering sources of variability other than AD pharmacokinetics are still lacking. In this study, we evaluated an in silico approach to prioritize ADs utilizing imputed gene expression profiles of citalopram remitters and AD-induced transcriptional profiles available in CMap. The positive correlation between gene expression of citalopram remitters and (es)citalopram-induced expression profiles in three cell lines (A375, MCF7, and PC3) suggests that the predicted gene expression profile of a remitter is correlated with in vitro expression profiles induced by the same ADs. No previous study has tested this hypothesis and our results show that approach might be used to rank ADs based on their likelihood of efficacy for an individual.

Analysis of transcriptional profiles of drugs and disease signatures is already an established approach in the domain of drug repositioning. For instance, Sirota et al. found that cimetidine showed an opposite expression pattern to that associated with lung adenocarcinoma, and experimentally validated this drug as a potential treatment [28]. Similarly, topiramate was found as a possible treatment for inflammatory bowel disease and this hypothesis was validated in an animal model [29]. In our study, we applied a similar strategy, but instead of disease-associated gene expression signatures we used TWAS predicted expression profiles associated with remission to a known AD drug to test if they

Table 2 Ranking of ADs and control drugs in five human cell lines.

Cell lines	A375		MCF7		PC3		HA1E		HT29		Combined cell lines	
Drug	Rank	Perm. *p* value	Rank	Perm. *p* value	Rank	Perm. *p* value	Rank	Perm. *p* value	Rank	Perm. *p* value	Rank sum	*p* value
Acarbose	15.9	0.6392	10.2	0.3238	17.7	0.7196	11.7	0.4108	1.4*	*0.0062*	56.9	0.2809
Amitriptyline	19.4	0.8054	3.6*	*0.0477*	16.4	0.6554	22.1	0.9088	16.9	0.6892	78.4	0.7385
Citalopram	6.6	0.1581	11.1	0.3723	4.1	0.0596	13.8	0.5146	7.9	0.2227	43.5	0.0878
Clofibrate	10.8	0.3504	7.2	0.1885	17.2	0.6954	3.6*	*0.0442*	3.1*	*0.0377*	41.9	0.0698
Clomipramine	14.7	0.5777	6.7	0.1596	11.7	0.4069	23.4	0.9523	21	0.8662	77.5	0.7385
Dosulepin	21.9	0.9000	23.1	0.9442	19.2	0.7954	12	0.4219	5.3	0.1108	81.5	0.8089
Duloxetine	15.4	0.6135	25.9	0.9988	22.2	0.9123	14.9	0.5738	7.7	0.2154	86.1	0.8670
Escitalopram	2.8*	*0.0296*	2*	*0.0085*	9.2	0.2812	18.5	0.7662	21.9	0.8935	54.4	0.2248
Fluoxetine	16	0.6488	14.4	0.5527	22.5	0.9238	8.8	0.2569	20.7	0.8523	82.4	0.8089
Fluvoxamine	24.2	0.9715	8.6	0.2492	4.2	0.0638	4.9	0.0915	13.6	0.5035	55.5	0.2615
Imipramine	5.5	0.1127	17.2	0.6931	18.8	0.7723	1.8*	*0.0092*	9	0.2758	52.3	0.1911
Ipriflavone	17.6	0.7162	19.1	0.7908	3.2*	*0.0362*	21.2	0.8742	5.2	0.1069	66.3	0.4770
Maprotiline	4	0.0642	10.1	0.3192	12.6	0.4581	14.6	0.5592	14.4	0.5462	55.7	0.2615
Mianserin	23.5	0.9527	20.2	0.8427	22.2	0.9123	22.7	0.9323	24.1	0.9631	112.7	0.9983
Mirtazapine	13.5	0.5123	23.8	0.9662	3.5*	*0.0431*	11	0.3677	8.5	0.2508	60.3	0.3426
Nortriptyline	15.3	0.6073	19.2	0.7965	16	0.6331	22.9	0.9377	11.7	0.4058	85.1	0.8537
Pantoprazole	10.5	0.3373	6.4	0.1473	10.6	0.3492	1.8*	*0.0092*	25.2	0.9908	54.5	0.2248
Paroxetine	22.7	0.9285	18.8	0.7742	8.3	0.2381	18.7	0.7769	20.6	0.8500	89.1	0.9021
Reboxetine	15.4	0.6135	13.8	0.5212	13.6	0.5146	24.6	0.9812	17.3	0.7054	84.7	0.8537
Rifaximin	13.2	0.4977	1.6*	*0.0050*	1.7*	*0.0062*	10.5	0.3404	7.1	0.1862	34.1*	*0.0234*
Selegiline	14	0.5381	21.6	0.9004	24.2	0.9723	19.3	0.8023	19.7	0.8165	98.8	0.9727
Sertraline	5.3	0.1085	9.2	0.2804	12.1	0.4285	5.6	0.1204	11.4	0.3938	43.6	0.0878
Tranylcypromine	21.3	0.8815	19.2	0.7965	22.8	0.9354	12.7	0.4623	20.9	0.8619	96.9	0.9635
Trazodone	13.5	0.5123	15.2	0.6027	17.2	0.6954	11.2	0.3792	20.6	0.8500	77.7	0.7385
Trimipramine	2.6*	*0.0250*	17.6	0.7146	6	0.1327	13.1	0.4815	2.7*	*0.0285*	42	0.0698
Venlafaxine	5.4	0.1115	5.2	0.1004	13.8	0.5288	5.6	0.1204	13.1	0.4738	43.1	0.0784

P values were obtained through the permutation procedure described in Statistical analysis paragraph. Top ranks are marked with an asterisk and *p* values < 0.05 are highlighted in italic in each cell line and the combined cell line results. The ADs (es)citalopram are mentioned in bold and control drugs are in italic.

Table 3 Spearman correlation between the drug-induced expression profiles of citalopram and escitalopram in each cell line.

Cell lines	Correlation coefficient	*p* value
A375	−0.008	0.405
MCF7	0.038	9.74E−05
PC3	−0.03	0.002
HT29	−0.01	0.298
HA1E	0.2	1.34E−91

P values suggesting the correlation coefficient being different from zero.

could be useful to prioritize the prescription of available ADs, rather than for identifying new potential ADs. We indeed hypothesized that AD-induced expression profiles in vitro may be correlated with gene expression profiles predicted using GWAS data in remitters to the same drug and similar drugs. A major observation that emerged from our results was the difference in the ranking of ADs drugs across cell lines which suggested that the selection of the most appropriate cell line(s) is a relevant step for applying our approach. The differences in results among cell lines can be explained by the intercellular drug-induced expression signature variability, as reported by Subramanian et al. According to this study, 15% of all the drug compounds produced highly similar signatures across nine cell lines, whereas the remaining drugs produced diverse signatures [22]. The heterogeneity of drug signatures depends on the cellular pathways associated with a cell type. In this study, we observed that A375 and MCF7 provided results that were more consistent with our hypothesis compared with other cell lines for both ADs and control drugs. This may be explained by the similar embryological origin of these cell lines. A375 and MCF7 are skin and breast cancer cell lines respectively, and both skin and breast cells originate from the ectoderm (outermost layer of the embryo), the same layer from which nervous tissues originate [30]. This hypothesis suggests that the use of brain cell lines may be more suitable for our study, but this was not possible as discussed among the limitations.

Despite the low comparability of gene expression profiles across cell lines, we decided to calculate the significance of cumulative ranks across cell lines. By combining ranks of drugs in the evaluated cell lines, the evidence of association of remitters' signature to ADs other than (es)citalopram might indicate that they may induce a similar gene expression profile to that observed in remitters to citalopram, thus hypothetically patients who benefit from citalopram may benefit also from these ADs. Alternatively, top drugs in the combined cells results may also represent false positive as none of the rank-sum statistics would survive after multiple testing correction.

Citalopram is a racemic mixture comprised of two enantiomers, R and S-citalopram (escitalopram) in equal proportions. However, the signatures of citalopram and escitalopram are only weakly correlated in the five analyzed cell lines. This can be due to the differences in modulated genes and pathways by these drugs in vitro, as reported by Sakka et al. Their study suggests that citalopram and escitalopram modulated 69 and 42 pathways, respectively, and ten pathways were differentially modulated in a neuroblastoma cell line [31]. In other words, the in vitro gene expression profile of citalopram is influenced by both escitalopram and R-citalopram to a similar extent, making it different compared to the profile of escitalopram alone. On the other hand, the in vivo gene expression signature of citalopram remitters is hypothetically highly dependent on genes regulated by escitalopram rather than R-citalopram, since escitalopram has a 50-fold higher affinity for the serotonin transporter compared to R-citalopram and it is considered responsible for the therapeutic effects of citalopram [32]. However, escitalopram was not close to significance in the analyses of ADs combined ranks across cell lines.

Our approach is innovative and shows important strengths. First, it reflects the polygenic architecture of AD response, characterized by multiple effects of small size [7]. Second, it is based on genotype-predicted gene expression profiles providing an advantage over traditional expression data from microarray and RNA sequencing methods. Patients from expression studies are indeed mostly medicated and brain tissues can only be acquired from post-mortem samples, hence, psychiatric medications might confound the expression results. On the contrary, genotype-predicted gene expression profiles are not susceptible to alteration due to medications because this approach only captures the heritable component of gene expression. However, ideally, RNA sequencing for measuring the transcriptome in biological specimens derived from drug-naive patients before and during treatment, after response evaluation, would represent the gold standard. Considering that this study design is expensive and time consuming, GWAS samples provide a more powerful and cheaper alternative, which is easily accessible, and there is the availability of GWAS summary statistics for a number of traits. Further, the expression profiles can be imputed for different tissues which can help to comprehend biological mechanisms at the tissue level. Last, our method is computationally simple, and it can be applied to other traits.

There are also some limitations to the proposed methodology. First, this method can be implemented only at a population level (on data from an aggregated sample of individuals), while it needs to be adapted for application at the individual level. Second, we could not test our method in neuronal progenitor cells or differentiated neurons from the CMap transcriptional catalog since AD-induced gene expression for these cell lines was not available. A prior CMap study suggested that neuronal cell lines are different compared to cancer cell lines in terms of the drug expression profiles but neuropsychiatric diseases can be reasonably modeled using cancer cell lines [22, 8]. However, the relevant differences between expression profiles of different cell lines found by our study and previous studies suggest that the identification of the most suitable cell line for the trait of interest is an important step. Neural cell lines may indeed show distinctive pathways that are relevant to AD action. Additionally, gene expression signatures available in the L1000 CMap database show various challenges in terms of their analysis and usage as discussed in previous work [33]. Due to the limited availability of transcriptional information for the drugs of interest across multiple cell lines, time points, and dosages, our analysis was restricted to expression profiles of 21 ADs in five cell lines treated with 10 micromolar drug concentration for 24 h time length.

In conclusion, we tested an in silico approach in five human cell lines by using GWAS results and drug-induced profiles to rank ADs based on their correlation, which hypothetically may reflect the chances of the efficacy of specific ADs. This study indicates that there is a correlation between (es)citalopram-induced expression profiles and predicted expression associated with remission to citalopram only in some cell lines. Therefore, at the individual level, on average the predicted expression of (es)citalopram remitters should be more correlated with (es)citalopram-induced expression than non remitters. Our approach can further be extended by investigating the correlation between a drug-induced expression profile and an individual's predicted gene expression levels which can be used to rank drugs by their predicted efficacy. Hence, the given method can be improved by considering genotype data at the individual level and using expression signatures of brain cell lines. A 'one size fits all' is not a valid strategy for the treatment of MDD and our study proposed a new approach aiming to contribute to the development of precision psychiatry.

Acknowledgements We thank the NIMH for providing the opportunity of analyzing their data on the STAR*D sample. We would like to thank the CMap team for making their data available for community research use. This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Biomedical Research Center at Oxford, South London, Maudsley NHS Foundation Trust, and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. We acknowledge the use of research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Center at South London & Maudsley and Guy's & St. Thomas' NHS Foundation Trusts, and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's & St. Thomas' Charity (TR130505).

Compliance with ethical standards

Conflict of interest CML is a member of the Scientific Advisory Board of Myriad Neurosciences. The other authors declare no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392:1789–858.
2. Fabbri C, Tansey KE, Perlis RH, Hauser J, Henigsberg N, Maier W, et al. New insights into the pharmacogenomics of antidepressant response from the GENDEP and STAR∗D studies: rare variant analysis and high-density imputation. Pharmacogenomics J. 2018;18:413–21.
3. Leuchter AF, Cook IA, Hamilton SP, Narr KL, Toga A, Hunter AM, et al. Biomarkers to predict antidepressant response. Curr Psychiatry Rep. 2010;12:553–62.
4. Tansey KE, Guipponi M, Hu X, Domenici E, Lewis G, Malafosse A, et al. Contribution of common genetic variants to antidepressant response. Biol Psychiatry. 2013;73:679–82.
5. Gandal MJ, Leppa V, Won H, Parikshak NN, Geschwind DH. The road to precision psychiatry: translating genetics into disease mechanisms. Nat Neurosci. 2016;19:1397–407.
6. Uher R, Tansey KE, Rietschel M, Henigsberg N, Maier W, Mors O, et al. Common genetic variation and antidepressant efficacy in major depressive disorder: a meta-analysis of three genome-wide pharmacogenetic studies. Am J Psychiatry. 2013;170:207–17.
7. Wigmore EM, Hafferty JD, Hall LS, Howard DM, Clarke TK, Fabbri C, et al. Genome-wide association study of antidepressant treatment resistance in a population-based cohort using health service prescription data and meta-analysis with GENDEP. Pharmacogenomics J. 2020;20:329–41.
8. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313:1929–35.
9. Tsuchimine S, Ochi S, Tajiri M, Suzuki Y, Sugawara N, Inoue Y, et al. Effects of cytochrome P450 (CYP) 2C19 genotypes on steady-state plasma concentrations of escitalopram and its desmethyl metabolite in Japanese patients with depression. Ther Drug Monit. 2018;40:356–61.
10. Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA, et al. Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study. Psychiatr Clin N Am. 2003;26:457–94.
11. Garriock HA, Kraft JB, Shyn SI, Peters EJ, Yokoyama JS, Jenkins GD, et al. A genomewide association study of citalopram response in major depressive disorder. Biol Psychiatry. 2010;67:133–8.
12. Novick D, Hong J, Montgomery W, Dueñas H, Gado M, Haro JM. Predictors of remission in the treatment of major depressive disorder: Real-world evidence from a 6-month prospective observational study. Neuropsychiatr Dis Treat. 2015;11:197–205.
13. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry. 2003;54:573–83.
14. Gaynes BN, Warden D, Trivedi MH, Wisniewski SR, Fava M, Rush AJ. What did STAR* D teach us? Results from a large-scale, practical, clinical trial for patients with depression. Psychiatric services. 2009;60:1439–45.
15. Lam M, Awasthi S, Watson HJ, Goldstein J, Panagiotaropoulou G, Trubetskoy V, et al. RICOPILI: Rapid Imputation for COnsortias PIpeLIne. Bioinformatics. 2020;36:930–3.
16. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48:245–52.
17. Gusev A, Mancuso N, Won H, Kousi M, Finucane HK, Reshef Y, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat Genet. 2018;50:538–48.
18. Mancuso N, Gayther S, Gusev A, Zheng W, Penney KL, Kote-Jarai Z, et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nat Commun. 2018;9:1–11.
19. Stegle O, Parts L, Durbin R, Winn J. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6:1–11.
20. The GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;7675:204–13.
21. Pain O, Pocklington AJ, Holmans PA, Bray NJ, O'Brien HE, Hall LS, et al. Novel insight into the etiology of autism spectrum disorder gained by integrating expression data with genome-wide association statistics. Biol Psychiatry. 2019;86:265–73.
22. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: l1000 platform and the first 1,000,000 profiles. Cell. 2017;171:1437–52.e17.
23. So HC, Chau CKL, Chiu WT, Ho KS, Lo CP, Yim SHY, et al. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. Nat Neurosci. 2017;20:1342–9.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-

based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102:15545–50.

25. Wise LH, Lanchbury JS, Lewis CM. Meta-analysis of genome searches. Ann Hum Genet. 1999;63:263–72.
26. Hicks JK, Sangkuhl K, Swen JJ, Ellingrod VL, Müller DJ, Shimoda K, et al. Clinical pharmacogenetics implementation consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants: 2016 update. Clin Pharmacol Ther. 2017;102:37–44.
27. Fabbri C, Tansey KE, Perlis RH, Hauser J, Henigsberg N, Maier W, et al. Effect of cytochrome CYP2C19 metabolizing activity on antidepressant response and side effects: Meta-analysis of data from genome-wide association studies. Eur Neuropsychopharmacol. 2018;28:945–54.
28. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med. 2011;3:96ra77.
29. Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med. 2011;3:96ra76.
30. Jiménez-Rojo L, Granchi Z, Graf D, Mitsiadis TA. Stem cell fate determination during development and regeneration of ectodermal organs. Front Physiol. 2012;3:1–11.
31. Sakka L, Delétage N, Chalus M, Aissouni Y, Sylvain-Vidal V, Gobron S, et al. Assessment of citalopram and escitalopram on neuroblastoma cell lines. Cell toxicity and gene modulation. Oncotarget. 2017;8:42789–807.
32. Jacobsen JPR, Plenge P, Sachs BD, Pehrson AL, Cajina M, Du Y, et al. The interaction of escitalopram and R-citalopram at the human serotonin transporter investigated in the mouse. Psychopharmacology. 2014;231:4527–40.
33. Musa A, Tripathi S, Kandhavelu M, Dehmer M, Emmert-streib F. Harnessing the biological complexity of Big Data from LINCS gene expression signatures. PLoS ONE. 2018;13:1–16.