GENETIC PREDICTORS FOR EPILEPSY DEVELOPMENT, TREATMENT RESPONSE AND DOSING

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of

Doctor in Philosophy

by

Kanvel Shazadi

March 2013

DECLARATION

This thesis is the result of my own work. The material contained within this thesis has not been presented, nor is currently being presented, either wholly or in part of any other degree or qualification.

Kanvel Shazadi

This research was carried out in the Department of Molecular and Clinical Pharmacology, in the Institute of Translational Medicine, at the University of Liverpool.

CONTENTS

ABSTRACT

Genetic predictors for epilepsy development, treatment response and dosing

Antiepileptic drug (AED) treatment is the first line strategy for seizure control in the majority of individuals with epilepsy but remains challenging, not least because of interindividual variability in efficacy, tolerability and dosing. The studies presented in this thesis set out to explore that variability from a genomic perspective in patients with newly diagnosed epilepsy from across the UK. Single nucleotide polymorphisms (SNPs) in genes encoding drug metabolising enzymes (DMEs) may be associated with the dose of carbamazepine (CBZ) required for seizure control. A cohort of 159 individuals who were seizure-free for 12 months on a stable dose of CBZ monotherapy was genotyped for 51 SNPs across six DMEs. Haplotype analysis identified 8 haplotype blocks across the genes. No single SNPs or haplotype blocks were associated with CBZ dose. Thus, it is unlikely that genetic variability in DMEs accounts for the individual differences in CBZ dose requirement.

A splice site SNP (rs3812718) in the *SCN1A* gene was previously shown to influence maximum doses of AEDs. This SNP was genotyped in 817 patients and tested for association with maximum and maintenance doses of several AEDs. An association was identified between rs3812718 and maximum AED dose, with an interaction analysis suggestive of a drug specific effect. These findings suggest that this *SCN1A* variant contributes to variability in the limit of tolerability to AEDs.

Response to AED treatment is multifactorial and likely to be influenced by multiple genes. Five SNPs previously reported to predict treatment outcome in epilepsy were genotyped in 772 patients and the resulting data, together with data from an Australian cohort, incorporated into a predictive algorithm. The algorithm failed to predict treatment outcome in general but was partially successful in identifying responders to CBZ and valproate. These five SNPs may be relevant to the prognosis of epilepsy, particularly when treated with specific AEDs.

Primary generalised epilepsies (PGEs) are highly heritable and believed to be polygenic in origin. Predictive algorithms were employed to explore genetic influences on seizure (absence vs. myoclonus) and epilepsy (PGE vs. focal) type using 1,840 SNP genotypes available from 436 patients with PGE. Although the algorithms failed to distinguish PGE patients on the basis of genetic variants, they showed improved association over univariate methods of analysis. Such an approach may be suitable for future investigations using large genomic datasets.

A recent genome-wide association study identified multiple genetic variants that approached genome-wide significance for association with 12 month remission from seizures. Five of these SNPs were genotyped in an independent cohort of 424 patients and tested for association with remission and time to remission. No significant associations were found, questioning the validity of the original observation or the method of replication. Further work is required to understand this outcome.

In conclusion, the genetic bases of epilepsy, AED response and AED dose requirement are multigenic and thus far undetectable using traditional association studies in modestly-sized patient cohorts. Further advances in genomic, bioinformatics and statistical methodologies are required before the genetic contribution to heterogeneity in epilepsy-related phenotypes can be translated into improved clinical care.

MANUSCRIPTS AND COMMUNICATIONS

Manuscripts in preparation

Shazadi K, Petrovski S, Roten A, Huggins RM, Brodie MJ, Pirmohamed M, Johnson MR, Marson AG, O'Brien TJ, Sills GJ. (2013). Validation of a multigenic model for treatment response in new-onset epilepsy

Shazadi K, Zhang JE, Carr D, Jorgensen AL, Alfirevic A, Wilson E, Brodie MJ, Pirmohamed M, Marson AG, Sills GJ. (2013). Common genetic variation in drug metabolising enzymes does not influence carbamazepine dose requirement in patients with newly diagnosed epilepsy

Shazadi K, Sills G J, Jorgensen AL, Alfirevic A, Pirmohamed M, Marson AG. (2013). A common variant in the sodium channel SCN1A gene influences the maximum dose of antiepileptic drugs prescribed to patients with newly treated epilepsy

Oral presentations

"Epilepsy genomics; Investigations into common genetic predictors of efficacy and dosing". Wellcome Trust Leena Peltonen Summer School of Genomics, Hinxton, 25th August 2011

"Genetic variability in drug metabolising enzymes as a determinant of carbamazepine dose requirement in newly diagnosed epilepsy". 9th European Congress on Epileptology, Rhodes, 30th June 2010

"Genetic variability in drug metabolising enzymes as determinant of carbamazepine dose requirement in newly diagnosed epilepsy". Annual Scientific Meeting of The UK Chapter of The International League Against Epilepsy, Sheffield, 8th October 2009

Poster presentations

Kanvel Shazadi, Slave Petrovski, Graeme J Sills, Munir Pirmohamed, Terence O'Brien, Anthony G Marson: Validation of the Australian five-SNP genetic classifier of early treatment outcomes in newly-diagnosed epilepsy. Wellcome Trust Leena Peltonen School of Genomics, Hinxton, 21st-25th August 2011.

Kanvel Shazadi, Slave Petrovski, Graeme J Sills, Munir Pirmohamed, Terence O'Brien, Anthony G Marson: Validation of the Australian five-SNP genetic classifier of early treatment outcomes in newly-diagnosed epilepsy. International Congress on Epileptology, Rome, 29th August 2011.

Kanvel Shazadi Graeme J Sills, Anthony G Marson, Munir Pirmohamed: Influence of a functional SCN1A polymorphism on the maximum and maintenance dose of antiepileptic drugs administered to patients with newly diagnosed epilepsy. International Congress on Epileptology, Rome, 29th August 2011.

Kanvel Shazadi, Slave Petrovski, Graeme J Sills, Munir Pirmohamed, Terence O'Brien, Anthony G Marson: Validation of the Australian five-SNP genetic classifier of early treatment outcomes in newly-diagnosed epilepsy. Wellcome Trust meeting on Genomic disorders, Hinxton, 23rd-25th March 2011.

Kanvel Shazadi, Graeme J Sills, Anthony G Marson, Munir Pirmohamed: Common genetic variation in drug metabolising enzymes as a determinant of carbamazepine dose requirement in newly diagnosed epilepsy. Annual American Epilepsy Society meeting, San Antonio, 6th December 2010.

Kanvel Shazadi, Graeme J Sills, Anthony G Marson, Munir Pirmohamed: Influence of a functional SCN1A polymorphism on the maximum and maintenance dose of antiepileptic drugs administered to patients with newly diagnosed epilepsy. Annual Scientific Meeting of the International League Against Epilepsy, Brighton, 14th October 2010.

ACKNOWLEDGEMENTS

Foremost I thank an ultimate thanks to Our Creator the All Mighty for His countless blessings one can never sum up. I would then also like to thank my primary supervisors Dr Graeme Sills and Professor Anthony Marson, not only for whom without I would not have had the opportunity to undertake a PhD but also for the endless opportunities given to me to develop as a research student. The support they gave to complete this thesis, their kindness, and encouragement and also the patience they offered over the past year for which I am ever grateful. Each of my secondary supervisors deserves a special thank you for their valuable involvement in my studies. Dr Ana Alfirevic, Professor Munir Pirmohamed, Professor Anthony Marson and Dr Andrea Jorgensen. Not only have you made it a pleasure to work in the Department of Personalised Medicine and part of the epilepsy research group, but you have contributed your time and effort to my work and inspired me though out the years. Your knowledge and enthusiasm for medical statistics and your passion for better patient care has undoubtedly motivated me during my research work and my future ambitions. A thank you to the Australian collaborators Professor Terence O'Brien and Dr Slave Petrovski, Professor Samuel Berkovic, Professor Ingrid Scheffer, for allowing me to visit their research departments and their contribution towards my thesis.

I also thank those colleagues and friends who have made my experience at the Department of Personalised Medicine during my time in Liverpool an enjoyable one. This includes Natalie, Vivian, Andrew, Steffen, Hayley, Philippe, Shaunik, Prathibha, Lucy and the many others I have encountered. From the epilepsy group a thanks to both Rebecca's, Alison and Arif for your friendship and kindness, and Pete for your all your administration help. For their treasured friendship, care, concern and their hugs, I thank from the bottom of my heart, Azizah, Maike and Elena. I am also indebted to Dan, Fabio, Sudeep and Helen for their invaluable guidance during the different works described in this thesis. Also to the clinicians Mas, Richard and Tav for allowing me to observe them during their ward rounds, their advice and kindness. I gained so much understanding from this short experience, for which I am ever grateful.

However unique and cherished thanks are reserved for my adopted sister Eunice Zhang. I very much appreciate your support both in my personal life as well as the guidance throughout my PhD. I can never repay you for your kindness and memorable friendship. I express utter gratitude for letting me be your flat mate, for the long chats and the laughter. Without your friendship and generous nature many of us would have not survived our long hours in the lab. You are truly an inspiration. Also to my dear old friend Rosy, whom again without I wouldn't have endured the last year. Thank you endlessly for your precious time and support.

I am blessed with my many family members who never gave up hope and struggled through alongside me. A deep thank you (Jazaakillaahu khayran) to my sister Arzoo for her prayers, encouragement and beneficial knowledge. Lastly to my dearest parents a heartfelt thank you. Words cannot full fill my gratitude to you both. To my father for your tough talk, patience, generosity and support over the past two years. I have been gifted with a fantastic mother, without whose faith I would be lost entirely. Thank you for your reassurance and your endless love throughout my life. May Allaah bless this work and make it of benefit to those who read it.

("…Waqul rabbizidnee 'ilmaa").

ABBREVIATIONS

ABC	ATP-Binding Cassette
ABCB1	ATP-Binding Cassette Sub-Family B Member 1
ABCC	ATP-Binding Cassette Sub-Family C
ABCC1	ATP-Binding Cassette Sub-Family C Member 1
ABCC2	ATP-Binding Cassette Sub-Family C Member 2
ABCG2	ATP-Binding Cassette Sub-Family G Member 2
ADR	Adverse Drug Reactions
AE	Adverse Effects
AE3	Anion Exchanger Isoform 3
AED	Antiepileptic drugs
AFF2	AF4/FMR2 Family, Member 2
AMPA	Alpha-amino-3-hydroxy-5-methyl-4- isoxazole-propionic acid
ANOVA	Analysis Of Variance
Arg	Arginine
AUC	Area Under Curve
BBB	Blood Brain Barrier
BCRP	Breast Cancer Resistance Protein
BNFIS	Benign Focal Epilepsy of Infancy
BP	Base Pair
BRD2	Bromodomain Containing 2
BZD	Benzodiazepines

Ca^{2+}	Calcium
CACNA1H	Calcium Channel, Voltage-dependent, T type, Alpha 1H
CACNA	Calcium Channel, Voltage-dependent
CAE	Childhood Absence Epilepsy
CBZ	Carbamazepine
CBZ-E	Carbamazepine Epoxide
CDCV	Common Disease Common Variant
CEPH	Centre d'Etude du Polymorphisme Humain
CHRN	Nicotinic Cholinergic Receptor
CHRNA4	Neuronal Cholinergic Receptor, Nicotinic Alpha 4
CHRNA7	Neuronal Cholinergic Receptor, Nicotinic Alpha 7
CI	Confidence Interval
CLCN2	Chloride Channel, Voltage-sensitive 2
CNS	Central Nervous System
CNV	Copy Number Variant
CNVs	Copy Number Variants
CSF	Cerebrospinal Fluid
CT	Computerised Tomography
CYP	Cytochrome
CYP1A2	Cytochrome P450, Family 1, Subfamily A, Polypeptide 2
CYP2C8	Cytochrome P450, Family 2, Subfamily C, Polypeptide 8
CYP2C9	Cytochrome P450, Family 2, Subfamily C, Polypeptide 9
CYP2C19	Cytochrome P450, Family 2, Subfamily C, Polypeptide 19

CYP2D6 Cytochrome P450, Family 2, Subfamily D, Polypeptide 6

CYP3A4 Cytochrome P450, Family 3, Subfamily A, Polypeptide 4

CYP3A5 Cytochrome P450, Family 3, Subfamily A, Polypeptide 5

CXorf40A Chromosome X Open Reading Frame 40a

DAT Dopamine Transporter

DDD Defined Daily Dose

DLT Decision Learning Tree

DME Drug Metabolising Enzyme

dNTP Deoxy-ribonucleotide Triphosphate

dsDNA Double Stranded DNA

EDTA Ethylene-diamine-tetra-acetic Acid

EEG Electroencephalogram

EFHC1 EF-Hand Domain (C-Terminal) Containing 1

EGFR Epidermal Growth Factor Receptor

EM Extensive Metaboliser

EPHX1 Epoxide Hydrolase 1, Microsomal

ESE Splicing Site Enhancers

ESE-finder Exonic Splicing Site Enhancer Finder

ESM Ethosuximide

ESL Eslicarbazepine

ESS Splicing Site Silencer

F	Forward
FAST SNP	Function Analysis and Selection Tool for Single Nucleotide Polymorphisms
FBM	Felbamate
FDR	False Discovery Rate
FEB4	Febrile Convulsions 4
FN	False Negative
FP	False Positive
GABA	Gamma-Aminobutyric Acid
GABBR	$GABA_B$ Receptor
GABBR2	$GABA_B$ Receptor 2
GABRA1	Gamma-Aminobutyric Acid A Receptor, Alpha 1
GABRG2	Gamma-Aminobutyric Acid Receptor
GAT	Gamma-Aminobutyric Acid Transporter
GBP	Gabapentin
GEFS$^+$	Generalised Epilepsy with Febrile Seizures
GPLD1	Glycosylphosphatidylinositol Specific Phospholipase D1
GRIK1	Glutamate Receptor, Ionotropic, Kainate 1
GSH	Glutathione
GST	Glutatione-S-Transferase Family
GSTA-4	Glutathione-S-Transferase Alpha 4
GSTM1	Glutathione-S-Transferase Mu 1

GTCS	Generalised Tonic Clonic Seizures
GWA	Genome Wide Association
GWAS	Genome Wide Association Study
HER2	Human Epidermal Growth Factor Receptor 2
HGP	Human Genome Project
His	Histidine
HLA-A*3101	Human Leukocyte Antigen Class I, A Serotype Group 31, Allele *3101
HLA-B*1502	Human Leukocyte Antigen Class I, B Serotype Group 15, Allele *1502
HWE	Hardy Weinburg Equilibrium
IDSP1	Iduronate 2-Sulfatase Pseudogene 1
IGE	Idiopathic Generalised Epilepsies
ILAE	International League Against Epilepsy
Ile	Isoleucine
IM	Intermediate Metaboliser
JAE	Juvenile Absence Epilepsy
JME	Juvenile Myoclonic Epilepsy
K^+	Potassium
Kb	Kilo Bases
KCNJ3	Potassium Inwardly-rectifying Channel, Subfamily J, Member 3

KCNQ1	Potassium Voltage-gated Channel, KQT-like Subfamily, Member 1
KCNQ2	Potassium Voltage-gated Channel, KQT-like Subfamily, Member 2
***k*NN**	*k*-Nearest Neighbour
K_V7.1	Voltage-dependent Potassium Channel A-Subunit
LCM	Locosamide
LD	Linkage Disequilibrium
Leu	Leucine
LEV	Levetiracetam
LGI1	Leucine-rich Glioma Inactivated Protein 1
LGI4	Leucine-rich Glioma Inactivated Protein 4
LOGREG	Logistic Regression
LRE	Localised Related Epilepsy
LRT	Log Ratio Test
LTG	Lamotrigine
Lys	Lysine
MAF	Minor Allele Frequency
Magea9b	Melanoma Antigen Family A, 9b
MALDI-TOF	Matrix-assisted Laser Desorption/Ionization- Time Of Flight
Mass1	Monogenic audio-genic seizure-susceptible protein
MDR	Multifactor Dimensionality Reduction
ME2	Malic Enzyme 2, NAD(+)-dependent, Mitochondrial

MEH	Microsomal Epoxide Hydrolase
$Mgcl_2$	Magnesium Chloride
Mglur	Metabotropic glutamate receptor
MGST1	Microsomal Glutathione S-transferase 1
MGST2	Microsomal Glutathione S-transferase 2
MGST3	Microsomal Glutathione S-transferase 3
MJ	Myoclonic Jerk
ML	Machine Learning
MLA	Machine Learning Algorithm
MNP	Multiple Nucleotide Polymorphisms
MRI	Magnetic Resonance Imaging
MRP	Multi-Drug Resistance Related Protein
MS	Mass Spectroscopy
n	Number
Na^+	Sodium
NAT1	N-acetyltransferase 1
NAT2	N-acetyltransferase 2
Na_V	Voltage Gated Sodium Channel
NaV1.1	Brain Sodium Channel, Voltage-gated, Type I Alpha Subunit
NaV1.15A	Brain Sodium Channel, Voltage-gated, Type I Alpha Subunit Exon 5N Splice Site Variant
NaV1.15N	Brain Sodium Channel, Voltage-gated, Type 1 Alpha Subunit Exon 5N Splice Site Variant

Nav1.2	Brain Sodium Channel, Voltage-gated, Type II Alpha Subunit
Nav1.3	Brain Sodium Channel, Voltage-gated, Type III Alpha Subunit
Nav1.4	Brain Sodium Channel, Voltage-gated, Type IV Alpha Subunit
Nav1.5	Brain Sodium Channel, Voltage-gated, Type V Alpha Subunit
Nav1.6	Brain Sodium Channel, Voltage-gated, Type VI Alpha Subunit
Nav1.7	Brain Sodium Channel, Voltage-gated, Type VII Alpha Subunit
Nav1.8	Brain Sodium Channel, Voltage-gated, Type VIII Alpha Subunit
Nav1.9	Brain Sodium Channel, Voltage-gated, Type IX Alpha Subunit
Navβ1	Brain Sodium Channel, Voltage-gated, Type I Beta Subunit
Navβ2	Brain Sodium Channel, Voltage-gated, Type II Beta Subunit
Navβ3	Brain Sodium Channel, Voltage-gated, Type III Beta Subunit
Navβ4	Brain Sodium Channel, Voltage-gated, Type IV Beta Subunit
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NICE	National Institute for Health and Clinical Excellence
NMDA	N-Methyl-D-aspartate
NN	Neural Network
NPV	Negative Predictive Value
OPRM	Opioid Receptor, Mu 1
OXC	Oxcarbazepine
PCR	Polymerase Chain Reaction

PD Pharmacodynamics

PDD Prescribed Daily Dose

PGB Pregabalin

PGE Primary Generalised Epilepsy

PGx Pharmacogenetics

PHT Phenytoin

PK Pharmacokinetics

PM Poor Metaboliser

PNS Peripheral Nervous System

PolyPhen Polymorphism Phenotyping v2

PPV Positive Predictive Value

QC Quality Control

R Reverse

RF Random Forest

RFM Rufinamide

RLIP76/RALBP1 RAL-A binding protein 1

rs Reference Sequence

RT Real-Time

RTB Retigabine

S14 Ribosomal Protein S14 Pseudogene 3

SANAD Standard and New Antiepileptic Drugs

SAP	Shrimp Alkaline Phosphatase
SCN	Sodium Channel Voltage-gated
SCN1A	Sodium Channel Voltage-gated Type I Alpha Subunit
SCN2A	Sodium Channel Voltage-gated Type II Alpha Subunit
SCN3A	Sodium Channel Voltage-gated Type III Alpha Subunit
SCN4A	Sodium Channel Voltage-gated Type IV Alpha Subunit
SCN5A	Sodium Channel Voltage-gated Type V Alpha Subunit
SCN6A/7A	Sodium Channel Voltage-gated Type VI/VII Alpha Subunit
SCN8A	Sodium Channel Voltage-gated Type VIII Alpha Subunit
SCN9A	Sodium Channel Voltage-gated Type IX Alpha Subunit
SCN10A	Sodium Channel Voltage-gated Type X Alpha Subunit
SCN11A	Sodium Channel Voltage-gated Type XI Alpha Subunit
SCN1B	Sodium Channel Voltage-gated Type I Beta Subunit
SCN2B	Sodium Channel Voltage-gated Type II Beta Subunit
SCN3B	Sodium Channel Voltage-gated Type III Beta Subunit
SCN4B	Sodium Channel Voltage-gated Type IV Beta Subunit
SEM	Standard Error of Mean
SIFT	Sorting Intolerant From Tolerant
SJS	Stevens Johnsons Syndrome
SLC	Solute Carrier
SLC1A3	Solute Carrier Protein Family 1A3
SMEI	Severe Myoclonic Epilepsy of Infancy
SNP	Single Nucleotide Polymorphisms

SS	Splice Site
SUDEP	Sudden Unexpected Death in Epilepsy
SV2A	Synaptic Vesicle Glycoprotein 2a
SVM	Support Vector Machine
TA	Transient Attack
TDM	Therapeutic Drug Monitoring
TE	Tris-EDTA buffer
TESS	Transcription Element Search System
TF	Transcription Factor
TFBS	transcription factor binding site
TFSEARCH	Transcriptional Factor Search
TGB	Tiagabine
TMEM114	Transmembrane Protein 114
TN	True Negative
TOF	Time Of Flight
TOF-MS	Time Of Flight Mass Spectroscopy
TP	True Positive
TPM	Topirmate
Try	Tryptophan
tSNPs	Tagging Single Nucleotide Polymorphism
Tyr	Tyrosine

UGT	UDP-glucuronosyltransferases
UGT1A4	UDP Glucuronosyltransferase 1 Family, Polypeptide A4
UGT2B7	UDP Glucuronosyltransferase 2 Family, Polypeptide B7
UM	Ultra-rapid Metaboliser
UNC	Unclassified Epilepsy
UTR	Untranslated Region
VGB	Vigabatrin
VPA	Valproate
WHO	World Health Organisation
ZNS	Zonisimide

CHAPTER ONE

GENERAL INTRODUCTION

CONTENTS

1.1 Epilepsy

Epilepsy is a common serious neurological disorder experienced by millions and a cause of substantial morbidity and mortality. The disorder is found in all ages from neonates to the elderly and affects approximately 0.75% of the population with an estimated prevalence of 8.5 per 1,000 individuals (www.who.int/mediacenre/factsheets/fs999). In the UK the prevalence rate is 6.2 per 1,000 population and it is diagnosed in about 80 individuals each day (Shorvon, 2009)(www.who.int/mediacenre/factsheets/fs999). Costing around two billion pounds a year, and known for its potentially devastating social consequences and poor health outcomes, untreated epilepsy is also a critical public health issue. The long standing stigma associated with epilepsy has resulted in many persons having lower employment and education levels and lower socioeconomic status (Duncan et al., 2006).s Additional issues include higher psychological distress, more physical injuries such as fractures and burns, and increased mortality than the general population (Shneker and Fountain, 2003, Fisher et al., 2005).

The history and treatment of epilepsy dates back some 4000 years (Chaudhary et al., 2011), with the term epilepsy originating from the Ancient Greek word 'epilambanein', which means "to seize" or "to attack". In these ancient times however, epilepsy was considered to have a religious origin; among existing theories were demonic possession and divine experience. Hippocrates became the first physician to define epilepsy as a "disease" and originally attributed the disorder to brain dysfunction (Fatovic-Ferencic and Durrigl, 2001). He was also the first to accurately describe epilepsy symptoms in both adults and children (Magiorkinis et al., 2010, Chaudhary et al., 2011).

The early remedies used to treat epilepsy were mainly empirical and reflective of this early notion of a spiritual basis (Magiorkinis et al., 2010). A more rational scientific view of epilepsy didn't appear until the 17th century when advancements in anatomy, physiology, and chemistry of the modern era were established and wherein nerve action was first associated with seizure causation (Magiorkinis et al., 2010, Chaudhary et al., 2011).

Today epilepsy is considered to be one of the most common serious neurological conditions and is defined as "a disorder of the brain characterised by an enduring predisposition to generate epileptic seizures": and requiring the occurrence of at least one epileptic seizure (Fisher et al., 2005). Epilepsy is currently not considered a uniform disorder but a manifestation of underlying brain dysfunction that comprises of a collection of several seizure-related syndromes, varying in their aetiologies, clinical features, treatment, and prognosis (Shneker and Fountain, 2003, Engel, 2006b).

1.1.1 Aetiology of epilepsy

The key manifestation of all epilepsies is recurrent seizures, though the aetiologies that give rise to these seizures are notoriously diverse, varying both worldwide and with age (Beck and Elger, 2008). Epilepsy is commonly associated with overt causes, these are often referred to as symptomatic or structural aetiologies and include central nervous system (CNS) tumors, neurodevelopmental abnormalities, CNS trauma and inflammation (Shneker and Fountain, 2003, Beck and Elger, 2008). In the UK the most common causes of epilepsy were cerebrovascular disease (15%), cerebral tumour (6%), alcohol-related (6%) and post-traumatic (2%) basis (Sisodiya and Duncan, 2004, Steinlein, 2008).

In a small number of patients a mutation in a single gene suffices to cause chronic seizures and this group of rare monogenic or Mendelian epilepsies thus are genetic in origin. More than 200 Mendelian epilepsies exist, however, in total they only account for around 1% of all epilepsy cases (Steinlein, 2008, Bhalla et al., 2011). In addition to the symptomatic and rare monogenic epilepsies there is a large group of common epilepsies that have a yet unknown aetiology (approximately two-thirds of all epilepsy cases). These epilepsies are thought to have some genetic contribution though are assumed polygenic and have an overall multifactorial basis (Sisodiya and Duncan 2004; Steinlein 2008).

1.1.2 Classification system

The classification of epilepsy is important for understanding its natural history, prognosis, diagnostic testing and treatment (Shneker and Fountain, 2003). Several classification systems have been proposed over the years and these continue to evolve over time to modify those definitions that predate modern neuroimaging, genomic technologies, and current concepts in molecular biology (Engel, 2006a, Berg et al., 2010). The most recent universally employed classifications of epilepsy seizures and syndromes were published by the International League Against Epilepsy (ILAE) in 1981 and 1989 respectively and although a new ILAE Classification system has since been proposed in 2001 and more recently in 2010; these latest versions remain complex and thus controversial as to their superiority for clinical usage (Commission on Classification and Terminology of the International League against Epilepsy, 1981, Commission on Classification and Terminology of the International League against Epilepsy, 1989, Berg et al., 2010). See Table 1.1 for the ILAE classification of seizures (Berg et al., 2010).

Table 1.1 Classification of seizures (based on 1989 ILAE classification). Adapted from Engel *et* al 2001)

1 Generalised seizures

1.1 Tonic-clonic seizures

1.2 Clonic seizures
- 1.2.1 *Without tonic features*
- 1.2.2 *With tonic features*

1.3 Typical absence seizures
1.4 Atypical absence seizures
1.5 Myoclonic absence seizures
1.6 Tonic seizures
1.7 Spasms
1.8 Myoclonic seizures

1.9 Eyelid myoclonia
- 1.9.1 *Without absences*
- 1.9.2 *With absences*

1.10 Myoclonic atonic seizures
1.11 Negative myoclonus
1.12 Atonic seizures

2 Focal seizures

2.1 Focal sensory seizures
- *With elementary sensory symptoms*
- *With experiential sensory symptoms*

2.2 Focal motor seizures
- *With elementary clonic motor signs*
- *With asymmetrical tonic motor seizures*
- *With typical (temporal lobe) automatisms*
- *With hyperkinetic automatisms*
- *With focal negative myoclonus*
- *With inhibitory motor seizures*

2.3 Gelastic seizures
2.4 Hemiclonic seizures
2.5 Secondarily generalised seizures

1.1.3 Seizures in epilepsy

Epileptic seizures have been defined as the transient occurrence of signs and/or symptoms due to involuntary, abnormal excessive or synchronous neuronal activity in the brain as defined by the ILAE (Shneker and Fountain, 2003, Fisher et al., 2005). Epileptic seizures are first broadly classified into partial seizures or generalised seizures, and this division is based entirely on the site of the abnormal neuronal activity or seizure initiation, and then separated further by their individual clinical presentations (Shneker and Fountain, 2003). Partial seizures originate in a small area of the brain (one or more localised foci) and can individually be characterised according to degree of impairment or loss of consciousness during seizure onset. Generalised seizures on the other hand occur simultaneously in both cerebral hemispheres, with no localised foci, they produce loss of consciousness, either briefly or for a longer period of time and are individually characterised by presence of motor activity (Browne and Holmes, 2001, Shorvon, 2009). Both seizure types are classified using their specific clinical and encephalogram (EEG) manifestations (Browne and Holmes, 2001).

There are three broad categories of partial seizures: i) simple partial seizures, where individuals remain fully conscious, ii) complex partial seizures, where consciousness is impaired or lost and iii) partial seizures with secondary generalisation (partial seizures that spread across the entire brain and evolve into a generalised seizure (Browne and Holmes, 2001, Shneker and Fountain, 2003). Generalised seizures are divided into two overall categories: either as i) presenting major motor symptoms, as for generalised tonic-clonic seizures (GTCS), atonic seizures, tonic seizures, clonic seizures and myoclonic seizures or ii) having a lack of motor activity, as for typical and atypical absence seizures (Browne and Holmes, 2001, Shneker and Fountain, 2003).

1.1.4 Pathogenesis

Normal cerebro-cortical function in humans has been well characterised, but the neurochemical basis of the processes underlying seizure generation is not well defined (Duncan et al., 2006). Seizures in epilepsy are thought to result from multiple mechanisms that appear diverse in nature, making their pathogenesis difficult to clarify (March, 1998). A common consideration however is that seizures are possibly the end-result of many different pathological processes that disrupt the normal function of the brain (McCormick and Contreras, 2001).

At a basic level, it is increasingly becoming evident that epileptogenic activity is most likely to be the consequence of a disruption of mechanisms that control the balance between excitation and inhibition in selected brain regions (Dichter and Ayala, 1987, Scharfman, 2007). The transition from normal brain neural networks to hyper-excitable networks, a

process known as epileptogenesis is also not yet completely understood. Theories suggest a greater spread in the activation and recruitment of neurones in addition to enhanced connectivity, enhanced excitatory transmission, a failure of inhibitory mechanisms and changes in intrinsic neuronal properties (March, 1998, Duncan et al., 2006).

1.1.5 Epilepsy and its syndromes

The 1989 ILAE classification system defines epileptic syndromes as "an epileptic disorder characterised by a cluster of signs and symptoms customarily occurring together; these include type of seizure, aetiology, anatomy, precipitating factors, age of onset, severity, chronicity, diurnal and circadian cycling and sometimes prognosis" . In the widely used 1989 ILAE classification (Commission on Classification and Terminology of the International League against Epilepsy, 1989), epilepsies are principally divided according to overall seizure type i.e. whether they are i) Generalised epilepsies, ii) Localisation-related epilepsies, iii) those that on the basis of clinical features cannot be assigned to either focal or generalised categories (unclassified) and iv) special syndromes, then sub-divided according to causation. There are three main causes used for classification; symptomatic epilepsies are those presumed to have an acquired cause, genetic epilepsies are those with a presumed genetic basis and crypotogenic epilepsies are presumed symptomatic but have an overall unknown cause (Commission on Classification and Terminology of the International League against Epilepsy, 1989). Refer to Tables 1.1 and 1.2 for 1989 Classifications.

Table 1.2 Classification of Epilepsies and Epileptic Syndromes and Related Seizure Disorders (based on 1989 ILAE classification)

1 Localisation-related (local, focal, partial) epilepsies and syndromes

1.1 Idiopathic (with age related onset)

Benign childhood epilepsy with centrotemporal spikes

Childhood epilepsy with occipital paroxysms

Primary reading epilepsy

1.2 Symptomatic

Chronic progressive epilepsia partialis continua

Syndromes characterized by seizures with specific modes of precipitation

Temporal lobe epilepsies

Frontal lobe epilepsies

Parietal lobe epilepsies

Occipital lobe epilepsies

1.3 Cryptogenic

2 Generalised epilepsies and syndromes

2.1 Idiopathic (with age-related onset)

Benign neonatal familial convulsions

Benign neonatal convulsions

Benign myoclonic epilepsy in infancy

Childhood absence epilepsy

Juvenile myoclonic epilepsy

Epilepsy with generalized tonic-clonic seizures on awakening

Other generalized idiopathic epilepsies

Epilepsies with seizures precipitated by specific modes of activation

2.2 Cryptogenic or symptomatic

West syndrome

Lennox-Gastaut syndrome

Epilepsy with myoclonic-astatic seizures

Epilepsy with myoclonic seizures

2.3 Symptomatic

Nonspecific etiology

Early myoclonic encephalopathy

Early infantile epileptic encephalopathy with suppression burst

Other symptomatic generalised epilepsies

Specific syndromes

Epileptic seizures complicating other disease states

3 Epilepsies and syndromes undetermined whether focal or generalised

3.1 With both generalised and focal seizures

Neonatal seizures

Severe myoclonic epilepsy of infancy

Epilepsy with continuous spike waves during slow-wave sleep

Acquired epileptic epilepsies

Other undetermined epilepsies

3.2 Without unequivocal generalised or focal features

4 Special syndromes

4.1 Situation-related seizures

Febrile convulsions

Isolated seizures or isolated status epilepticus

Seizures occurring only with acute metabolic or toxic events

(Commission on Classification and Terminology of the International League Against Epilepsy, 1989) The 1989 classification of syndromes was adopted in this thesis, due to the primary use of these definitions for patient classification for the various UK epilepsy cohorts

1.1.6 Epidemiology

Epilepsy affects approximately 50 million people globally. The prevalence of active epilepsy is approximately 5-10 per 1000 population in most locations (Sander, 2003a). In the UK, epilepsy is diagnosed in about 80 individuals each day; 350,000 have active epilepsy (defined as the occurrence of a seizure during the previous 2 years and/or the taking of antiepileptic drugs) and 100,000 have refractory epilepsy (Sander, 2003a). However studies have shown that the disorder is not evenly distributed, with the age-adjusted incidence of epilepsy in developed countries ranging from 24 to 54 new cases per 100,000 population and a higher rate presumed for developing countries (recent reports of 49.3 to 190 per 100,000 population)(Sander, 2003a)(www.who.int/mediacentre/factsheets/fs999). In the UK, around 450,000 individuals have epilepsy and the age-standardised prevalence is estimated as 7.5 per 1000 population (Sisodiya and Duncan, 2004). 50% of individuals who develop epilepsy do

so before the age of 15 years though prevalence increases with age (around 3 per 1,000 in under 16s and 12 per 1,000 in over 65s) (www.who.int/medicentre/factsheets/fs 999/en/index.html).

1.1.7 Co-morbidities and risk factors

There are numerous comorbidities that complicate both the assessment and treatment of epilepsy. These include psychological and/or psychiatric problems, having a learning disability and/or a concomitant medical condition(s) (Shneker and Fountain, 2003, Duncan et al., 2006). The disorder is additionally often associated with serious physical implications from a heightened accumulation of brain damage and/or neurological deficits. Individuals with epilepsy thus generally carry a greater risk of injury and/or sudden unexpected death in epilepsy (known as SUDEP) than that of the general population (Duncan et al., 2006). Although most people with epilepsy are able to lead a normal emotional and cognitive life, neurobehavioral problems can be found in a large number of patients (Torta and Keller, 1999).

1.2 Prognosis of epilepsy

The risk of recurrence is greatest in the first few months after a first seizure years (Hauser et al., 1998). About 50% of those who have suffered a single unprovoked seizure have a further seizure within 5 years (Hauser et al., 1998) and about 75% of those with two initial unprovoked seizures suffer further seizures within the first four years (Hauser et al., 1998, Sisodiya and Duncan, 2004). In general however, the prognosis for complete seizure control is good as approximately 70% of patients do eventually achieve long-term remission within the first 5 years of diagnosis (Sander, 1993, Cockerell et al., 1997). Prognostic factors include age of onset, number of seizures in the early stages of the condition, early response to antiepileptic drugs (AEDs) and certain epilepsy specific EEG findings (Sander, 1993, Kwan and Brodie, 2000a, MacDonald et al., 2000, Kwan and Brodie, 2001a). In accordance with the association between early seizure control and long-term remission the prospect of seizure cessation has also been indicated to decrease as time elapses (Brodie and French, 2000, MacDonald et al., 2000, Kwan and Brodie, 2001a, Sander, 2003a).

AEDs are highly successful in suppressing seizures in most but little is known about the role of AED treatment on the outcome of epilepsy (Duncan et al., 2006). The recent assumption is that an inherent element to both treatment response and outcome may exist and so for some chronic epilepsy patients, remission could be impossible from onset (MacDonald et al., 2000, Sander, 2003a).

1.3 Treatment of epilepsy

Pharmacotherapy is the mainstay of treatment for people with epilepsy (Panayiotopoulos and International League against Epilepsy., 2005). Non-pharmacological options include ketogenic diet, vagal nerve stimulation and brain surgery, although these are only feasible in selected individuals and are usually considered once drug treatment has failed; with the latter used as a last resort in severe or chronic epilepsy cases (Sander, 2004, Duncan et al., 2006).

1.3.1 Pharmacological management

AEDs are primarily intended to prevent epileptic seizures and generally function to increase inhibition, decrease excitation, and/or prevent aberrant burst firing of neurones that is often associated with seizure generation. The ultimate goal of pharmacological management in epilepsy is to achieve complete seizure freedom as quickly as possible (Vajda, 2007), without any drug-related adverse effects (AEs) (i.e. nausea, dizziness, weight gain), adverse drug reactions (ADRs) (i.e. hepatotoxicity, haemotoxicity, dermatotoxicity and teratogenicity) (Sander, 2004, Mann and Pons, 2007) and using only a single AED (Beghi and Perucca, 1995).

As previously mentioned AEDs are greatly effective in abolishing seizures in up to 70% of patients (Kwan and Brodie, 2000a, 2001a). Significant seizure control reduces the morbidity and premature mortality (Sander and Bell, 2004); (Mohanraj et al., 2006) often associated with unpredictable and continuous seizures, and so greatly enhances overall quality of life (Birbeck et al., 2002).

1.3.2 Antiepileptic drug treatment

There are currently over 20 AEDs available, differing in their chemical structure and/or mode of action (Table 1.3). The majority of the available AEDs were developed in the 1980-1990's (Schachter, 2007) and several new AEDs are in clinical trials or have been recently approved (Bialer and White, 2010). These drugs have been developed either through serendipity, such as the early discovery of the anticonvulsive properties of bromide and barbiturates (Porter and Rogawski, 1992), or through the screening of new compounds in experimental animal models of epilepsy. The progress in these animal studies and drug trials have, over the past 20 years, allowed the introduction of numerous AEDs to the clinic (Duncan et al., 2006, Smith et al., 2007).

1.3.3 History and effectiveness of anticonvulsants

Drugs introduced up to the early 1970s included the benzodiazepines (BZDs) (such as diazepam), carbamazepine (CBZ), ethosuximide (ESM), phenytoin (PHT) and valproic acid (VPA) (Schachter, 2007). These early AEDs are known as older generation drugs and were considered an advancement over the initially available barbiturates, due to their markedly improved tolerability and a broader spectrum of efficacy against different seizure types (Schachter, 2007). A more rational approach was taken to subsequent AED development (Kupferberg, 2001, Smith et al., 2007). This produced the following 'new generation of AEDs': felbamate (FBM), gabapentin (GBP), lamotrigine (LTG), levetiracetam (LEV), oxcarbazepine (OXC), pregabalin (PGB), topiramate (TPM), tiagabine (TGB), vigabatrin (VGB) and zonisamide (ZNS) (Schachter, 2007) and more recently lacosamide (LCM), eslicarbazepine (ESL), rufinamide (RUF) and retigabine (RTG) (Bialer et al., 2007, Bialer and White, 2010). Of these TGB and VGB were designed with specific mechanisms of action (Bialer et al., 2007, Bialer and White, 2010).

Despite this ever-growing list of anti-seizure agents, issues with efficacy and tolerability largely remain (Kwan and Brodie, 2001a). The available evidence indicates that efficacy and tolerability to drug treatment in epilepsy has not substantially improved (Loscher and Schmidt, 2011). There is no compelling evidence that third-generation AEDs, have made clinically relevant advances in the day to- day tolerability of current epilepsy treatment. Some AEDs of these newer drugs do however have advantages; namely linear pharmacokinetics (PK), an improved tolerability profile, lessened drug interaction potential, a lower risk of hypersensitivity reactions and fewer AED-associated AEs (Perucca, 2001a, Loscher and Schmidt, 2011). Data in this regard thus remains to be accumulated over the coming years before any definite conclusions on the success of modern AEDs can be drawn.

1.3.4 Pathways of drug action

Several distinct molecular and cellular events occur during a seizure that involve sodium (Na^+), calcium (Ca^{2+}) and potassium (K^+) ions (McNamara, 1994). These are not only critical for normal neuronal function and signaling pathways, but are also important in the initiation of seizures, spread of seizure activity and arrest of the seizure (McNamara, 1994). Na^+ conductance is important for the initiation and maintenance of seizure activity as is Ca^{2+} conductance, which also contributes to neuronal injury, and K^+ conductance is essential in the arrest of seizure discharge (McNamara, 1994, Scharfman, 2007). Synaptic transmission also plays a critical role in maintaining the balance between excitation and inhibition, and so perturbation in this process can likewise lead to seizure generation (Scharfman, 2007). The principal neurotransmitters involved in synaptic transmission are gamma-aminobutyric acid

(GABA) and the excitatory amino acid glutamate (Scharfman, 2007).

To exhibit antiepileptic activity, a drug must act on one or more target molecules in the brain, such as ion channels, neurotransmitter transporters and neurotransmitter metabolic enzymes (Kwan et al., 2001). These interactions modulate the bursting properties of neurones and reduce synchronisation in localised neuronal ensembles (Kwan et al., 2001). Although the mechanisms of action of many AEDs are not fully understood, they are broadly categorised according to the following three general modes of action (Kwan et al., 2001) (based on their basic cellular mechanisms) i) modulation of voltage-dependent ion channels (including Na^+, Ca^{2+}, K^+ channels), ii) enhancement of GABA-mediated inhibitory neurotransmission and iv) attenuation of excitatory, glutamate-mediated transmission (Meldrum, 1996, Kwan et al., 2001, Schachter, 2007). For many of the newer AEDs multiple molecular mechanisms of action have been identified (White, 1999).

1.3.5 The gamma-aminobutyric acid inhibitory system

The potentiation of inhibitory neurotransmission is one of the main mechanisms of AED action and several AEDs exert their effects, at least in part, by actions on the GABAergic system (Kwan et al., 2001, Benarroch, 2007). In the CNS, inhibition is principally mediated by the neurotransmitter GABA, which functions through binding to chloride-permeable ionotropic $GABA_A$ receptors (mediators of fast inhibition) and metabotropic $GABA_B$ receptors (mediators of slow inhibition) (Benarroch 2007). Dysfunction of $GABA_A$ receptor-mediated fast inhibition is an important pathophysiological mechanism of increased neuronal excitability and has been identified in the process of epileptogenesis (Benarroch 2007; Olsen and Avoli 1997). Loss-of-function of the receptor, through mutations in $GABA_A$ subunit genes have additionally been linked to various epilepsy syndromes (Olsen and Avoli, 1997, Baulac et al., 2001, Wallace et al., 2001a, Bianchi et al., 2002, Macdonald et al., 2004).

Benzodiazepines, barbiturates, FBM and TPM, all modulate this pathway by facilitating GABA-ergic neurotransmission through increasing $GABA_A$ receptor function (Kwan et al., 2001). AEDs may alternatively enhance synthesis of the GABA neurotransmitter (as with VPA and GBP), decrease GABA degradation (as with VPA and VGB), or prevent GABA re-uptake into neurones and glia (as with TGB) (Kwan et al., 2001).

1.3.6 Glutamate neurotransmission

Excitatory neurotransmission in the brain is mediated by excitatory amino acids. Glutamate is the principal excitatory neurotransmitter in the mammalian brain and exerts it effect through one of four glutamate receptors (the ionotropic, NMDA, kainate and AMPA receptors and the metabotropic mGluR receptor) found in the CNS (Meldrum, 2000).

Abnormal glutamate receptor function has been observed in several experimental seizure models and has been implicated in both the initiation and spread of epileptic seizures (Meldrum, 1995, Chapman, 1998, 2000). Because of the role of glutamate in the pathophysiology of seizures and the substantial evidence that glutamate receptor antagonists are protective in various animal models, great effort has been devoted toward the development of novel AEDs targeting the glutamate system (Meldrum, 2000, Meldrum and Rogawski, 2007). Of the drugs mentioned above only FBM and TPM appeared to reduce the glutamate action via the blockade of ionotropic glutamate receptors in addition to their primary action on GABA neurotransmission (Upton, 1994, Macdonald and Kelly, 1995, Meldrum, 1996). Recently however perampanel has been developed and approved as a selective AMPA receptor (major ionotropic glutamate receptor subtype) antagonist for the treatment of partial onset seizures (Rogawski, 2011, Krauss et al., 2012).

1.3.7 Neuronal ion channels

The K^+, Na^+ and Ca^{2+} neuronal voltage-gated ion channels maintain neuronal function through shaping the sub-threshold electrical behaviour of the neurones, allowing action potential firing, and regulating neuronal responsiveness to synaptic signals and pre-synaptic neuronal neurotransmitter release (Scharfman, 2007). These channels are subsequently central to deregulation in neuronal signaling as evident in the generation of seizure discharges. The vast majority of the newer and established AEDs can thus exert their anticonvulsant effects through ion channel modulation (Bialer and White, 2010). Na^+ channel targeting AEDs include CBZ, ESL, FBM, PHT, LCM, LTG, OXC, RUF, TPM, VPA and ZNS; K^+ channel AEDs include RTG and TPM and Ca^{2+} channel AEDs include ESM, FBM, GBP, LEV, PGB, LTG, TPM, VPA (Shorvon, 2010).

Table 1.3 Proposed mechanisms of action of antiepileptic drugs

Drug		Main mode of action
Benzodiazepines	BZD	Potentiate GABA-mediated inhibition
Carbamazepine	CBZ	Blocks voltage-gated Na^+ channels
Clobazam	CLB	Increases inhibition by $GABA_A$
Ethosuximide	ESM	Blocks T-type Ca^{2+} channels
Felbamate	FBM	Potentiates GABA-mediated inhibition and blocks voltage-dependent Na^+ channels
Gabapentin	GBP	Binds to the α2δ subunit of neuronal voltage-gated Ca^{2+} channels inhibiting calcium flow
Lamotrigine	LTG	Blocks voltage-gated Na^+ channels
Levitiracetam	LEV	Binds to synaptic vesicle protein SV2A
Lacosamide	LCM	Blocks voltage-gated Na+ channels
Oxcarbazepine	OXC	Blocks voltage-gated Na^+ channels
Perampanel	PRM	Blocks AMPA glutamate receptor
Phenobarbital	PB	Augments the inhibitory effect of GABA by prolonging the Cl^- channel opening at the $GABA_A$ receptor
Phenytoin	PHT	Blocks voltage-gated Na^+ channels
Pregabalin	PGB	Calcium channel blocker, binds to channel to inhibit calcium flow
Retigabine	RTG	Modulation of K^+ channel, prolonging channel opening
Rufinimide	RFM	Blocks voltage-dependent Na^+ channels
Sodium Valproate	VPA	Blocks voltage-dependent Na^+ channels, facilitates the effects of GABA and reduces T-type Ca^{2+} currents
Tiagabine	TGB	Blocks GAT1,GABA transporter thus inhibits neuronal and glial reuptake of GABA to increase the availability of GABA and inhibit postsynaptic neurons
Topiramate	TPM	Blocks voltage-gated Na^+ channels and Ca^{2+} channels, enhances GABAergic neurotransmission and inhibits carbonic anhydrase
Vigabatrin	VGB	Enhancing biosynthesis and preventing degradation of GABA by inhibiting GABA transaminase, resulting in elevated brain levels of GABA
Zonisamide	ZNS	Blocks voltage-gated Na^+ channels

Data taken from Loscher and Schmidt 1999, Schachter et al 2007 and Kwan et al 2001

1.3.8 Principles of treatment

The licensing of many new AEDs in the last 20 years has presented a greater choice of AEDs for physicians and accordingly patients with epilepsy now have a better chance of achieving optimum treatment than in the past (Perucca, 2002a). Due to differences in individual efficacy, drug PK, side-effects and drug interactions between the different AEDs (Brodie and Kwan, 2001, Schachter, 2007), it is often difficult to predict which drug will be the best tolerated and most likely to provide the best seizure control in a given individual (Perucca, 2001a, 2002a).

Several patient characteristics are currently used to divide patients into subpopulations to aid drug selection. The effectiveness of the newer AEDs was determined with regulatory trials and post marketing studies in patients with defined seizure types (Schachter, 2007). Based on clinical trial data, mechanisms of action, and clinical experience, certain AEDs are generally preferred for focal epilepsy and other AEDs are preferred for generalised epilepsy (Perucca, 1999, Vajda, 2007). Therefore, the first step in selecting among the currently available AEDs for a particular patient is to determine his or her seizure type(s) and/or epilepsy syndrome (Schachter, 2007) (http://guidance.nice.org.uk /CG20/Guidance). Additional subpopulations are based on age, gender, medical comorbidities, concomitant medications, individual lifestyle, individual preference, childbearing potential, likelihood of AEs, and the licensed indication of the drug (http://gui dance.nice.org.uk/CG20/Guidance).

1.3.9 Clinical use

First-line AEDs are generally started at a low dosage and drug dose is titrated up gradually to a target dose. If seizures continue, titration is continued until seizures are controlled or up to the maximum tolerated dose (dose at which AEs appear in a given individual)(Brodie and French, 2000). If the drug is ineffective at this maximum tolerated dose, it is discontinued, but slowly (tapered off through dose reduction) and replaced by an alternatively appropriate AED; again selected based on seizure type, epilepsy type, specific syndrome etc. (Perucca, 2001a, Brodie and Kwan, 2002, Schachter, 2007).

Existing National Institute for Health and Clinical Excellence (NICE) guidelines on the management of epilepsy indicate that individuals should preferably be treated with a single AED, where possible. Clinicians are also recommended to maintain treatment with a single AED to avoid the complications that arise with the use of multiple drug combinations (http://guidance.nice.org.uk/CG20/Guidance). With AED monotherapy compliance is often enhanced, overall medication costs are usually less and there are generally fewer idiosyncratic reactions, teratogenic effects, and other dosage and/or drug interaction related side effects (Brodie and Kwan, 2001). AED monotherapy successfully controls seizures in the majority of

patients (Leppik, 2000). A significant number of patients with more severe forms of epilepsy do however require multiple drug treatment or polytherapy (Brodie and French, 2000, Leppik, 2000, Sander, 2004). Many patients however still require seizure management with a combination of drugs (Brodie and Kwan, 2001). Such 'combination therapy' (also known as adjunctive or 'add-on' therapy) is usually considered when all attempts at monotherapy have not resulted in seizure freedom (http://guidance.nice.org.uk /CG20/Guidance)(Duncan et al., 2006).

Typically if monotherapy is poorly tolerated or ineffective at the maximum tolerated dose, the strategy is to switch to another first-line drug. Second-line options are used when all first-line drugs fail (http://guidance.nice.org.uk/CG20/Guidance) (Mattson et al., 1985, Kwan and Brodie, 2000a, 2001a, Sander, 2004). With combination therapy, an additional drug is usually titrated to a tolerable and effective dosage before the first AED is tapered (Duncan et al., 2006). There is a lack of clarity with the dose at which a drug should be deemed ineffective and when alternative AEDs or combination therapy should be considered (Kwan and Brodie, 2000b, Brodie and Kwan, 2002, Kwan and Brodie, 2004). For combination therapy however there is now suggestion that the mechanism of action of each AED should be taken into consideration (Brodie and Kwan, 2001, Sander, 2004).

1.3.10 Effectiveness of pharmacotherapy in clinical practice

Differences in treatment response with particular AEDs, in terms of variation in clinical efficacy, dosing and tolerability, between individuals with seemingly similar disorders is a very common phenomenon among all pharmacotherapeutics (Dlugos et al., 2006). Prognosis with AED treatment thus varies considerably among the different types of epilepsy (Semah et al., 1998) and may also differ even between patients with the same epilepsy syndrome (Schmidt and Loscher, 2005). Better remission is generally evident for secondary generalised attacks when compared to partial seizures (Mattson et al., 1985). In addition most studies have reported prognosis to be poor in patients with multiple seizure types, associated neurological deficit and behavioral or psychiatric disturbance (Sander, 1993). The longer patients continue to have seizures after their initial diagnosis, the lower the probability of achieving remission (Annegers et al., 1979, French, 2002). Of the 70% of individuals responsive to AED therapy, nearly 50% are seizure free with initial treatment, and a further 16% of patients who find the first drug to be ineffective in suppressing seizure activity can expect freedom from seizures with additional AED treatment (Kwan and Brodie, 2000a). Those who fail treatment with a second AED are however thought to have a small chance of ever obtaining seizure freedom (Kwan and Brodie, 2000b, McCorry et al., 2004).

The use of AEDs in clinical treatment is often complicated and in some cases

problematic, even for responsive individuals (Perucca, 2001a) and this can sometimes be attributed to AED pharmacology. Some older and newer generation AEDs present different activity spectra and a narrow therapeutic index (Sander, 2004), some with highly variable and nonlinear PK, sub-optimal response rates and a propensity of many AEDs (particularly the older generation) to cause drug interactions. Consequently even though there are over 20 AEDs (old and new generation) available, which continues to grow steadily, the 60-70% response rate in epilepsy remains. The failure to achieve seizure freedom in a substantial minority of individuals is perhaps the most important issue with current AED therapy (Kwan and Brodie, 2000a). Additional serious issues in AED utilisation include the occurrence of AEs, as virtually all AEDs show common side effects and/or idiosyncratic reactions (Sander, 2004, Perucca and Meador, 2005, Schachter, 2007, Zaccara et al., 2007). There is also the challenge of identifying the most effective and best tolerated dose of a specific drug(s) for individual patients, which can vary greatly among individuals and is currently impossible to predict (Kwan and Brodie, 2001a, Perucca, 2002b).

1.4 Treatment failure in epilepsy

Individuals that fail to achieve remission with long term AED treatment are referred to as having drug-resistant or refractory epilepsy (also referred to as a pharmaco-resistant phenotype) (Kwan and Brodie, 2000a). Drug resistant individuals are usually treated with multiple AEDs that in combination can often have adverse sedative, behavioral and/or psychiatric effects (Kwan and Brodie, 2002, Elger, 2003). There is also a greater risk of cognitive impairment with prolonged drug use, the likelihood of which increases with seizure frequency, duration and severity (Cramer, 1994, Vermeulen and Aldenkamp, 1995, Kwan and Brodie, 2001b).

What causes epilepsy to become ‘refractory’ has so far remained elusive. Several clinical/environmental prognostic factors have been implicated however as previously discussed for AED treatment response the biological basis of ‘refractoriness’ is a multifactorial and variable phenotype with a genetic and clinical basis. As AEDs are principally required to traverse the blood–brain barrier (BBB) and then secondly bind to one or more target molecules to exert their particular therapeutic effect two theories have emerged from these pharmacological pathways in an attempt to explain treatment failure in epilepsy (Remy and Beck, 2006). These are the drug transporter and drug target hypotheses. The transporter theory proposes an overexpression of efflux transporters at the BBB (i.e. Pgp efflux protein), that leads to a surge in active efflux of AEDs, thus decreasing their concentration in the CNS, despite adequate AED exposure and/or measured serum levels (Loscher and Potschka, 2002). The target theory on the other hand proposes a reduction in AED target sensitivity (i.e.

molecular targets of AEDs such as the neuronal channels) in epileptogenic brain (Remy et al., 2003). This altered channel sensitivity would then cause reduced efficacy of a given AED at its molecular target (Remy et al., 2003).

1.4.1 Indications of resistance to antiepileptic drugs

The identification of refractory epilepsy is not only important for clinical decision making i.e. for doctors to consider alternative non-pharmacological treatment options for patients, but is a vital step towards understanding disease pathophysiology, determinants of natural history, predictors of prognosis and it can also benefit the development of novel treatment strategies (Kwan and Brodie, 2010). Despite many research studies investigating this clinical phenomenon, a precise definition to identify people with treatment resistance has remained elusive for many years and this has resulted in use of diverse criteria by different clinicians and researchers (Annegers et al., 1979, Kwan and Brodie, 2000a, Berg and Kelly, 2006).

The most recent description proposed by the ILAE defines refractory epilepsy as "failure of adequate trials of two tolerated and appropriately chosen and used AED schedules (whether as monotherapies or in combination) to achieve sustained seizure freedom" (Kwan et al., 2010). Complete failure with two or more AED monotherapies characterises individuals with intractable epilepsy. Combining a wide range of two or perhaps three different AEDs was effective in some of these difficult to treat individuals (Stephen et al., 2001, Stephen and Brodie, 2002), although robust data evaluating the effectiveness of AED combination therapy is currently scarce (Pearce et al., 2008).

1.4.2 Management of drug-resistance in epilepsy

New AEDs are being developed to address the issue of treatment-resistance in the epilepsy population (French et al., 2004, Bialer and White, 2010). However there is emerging evidence that better results can be obtained by more appropriately combining modern AEDs (that offer novel mechanisms of action and fewer drug interactions) with complementary modes of action (Moeller et al., 2009). This suggestion however remains to be proven with robust drug efficacy data (Pearce et al., 2008, Poolos et al., 2012). Efficacy of new AEDs used as adjunctive therapy in patients unresponsive to established AEDs has been investigated in several controlled trials. However, none of the new AEDs have produced a high percentage of seizure freedom in these studies (Cramer et al., 1999, Barcs et al., 2000, Cramer et al., 2001, French et al., 2003, Callaghan et al., 2011) thus treatment outcome in refractory epilepsy remains poor, with responder rates (50% seizure reduction) found to range from 15-50% in these studies. More recent evidence has also shown the probability of remission in people with

intractable epilepsy as around 5% per year (Callaghan et al., 2007, Choi et al., 2008).

1.4.3 Predicting drug response: clinical markers for drug resistance

Phenotypic markers for distinguishing those patients who appear unresponsive to AEDs from those able to successfully achieve seizure control (Loscher, 2005a, Tate and Sisodiya, 2007) would allow early consideration of non-drug therapies (Cockerell et al., 1995, MacDonald et al., 2000, Dlugos et al., 2001, Mohanraj et al., 2006). This, for the most severe cases at least (where individuals are inevitably likely to require surgery due to severity of seizures) would potentially prevent some of the devastating consequences associated with intractable epilepsy (Brodie, 2005a).

Pretreatment seizures have long been identified as a predictor of outcome, with a larger number of pre-treatment seizures universally associated with a poorer response to early AED treatment (Camfield et al., 1996, Kwan and Brodie, 2000a). High seizure frequency evident during the early stages of AED treatment is also considered to indicate poor seizure control and is similarly used to envisage the likelihood of developing pharmacoresistance (Brodie, 2005a). In addition there is good evidence that seizure type or syndromic diagnosis is associated with likelihood of seizure control (Semah et al., 1998). Individuals presenting with idiopathic generalised seizures more likely to become seizure free than those with symptomatic or cryptogenic epilepsies (Koster et al., 2009). Further potential risk factors for refractory epilepsy include seizure clusters, family history, febrile convulsions, environmental factors such as traumatic brain injury, and psychiatric comorbidity (Petsche et al., 1972, Hitiris and Brodie, 2006, Mohanraj et al., 2006).

1.4.4 Inherent role in drug resistance

Despite identification of several clinical and environmental factors that appear to contribute to the biological basis of drug-resistance in epilepsy, the prognostic value of most of these factors is rather limited, and none can explain multidrug resistance alone (Brodie, 2005a, Loscher, 2005a). Other influential features beyond those contributing to the aetiology of epilepsy have therefore been implicated in the multifaceted basis of AED response (Petsche et al., 1972, Depondt, 2006b, Koster et al., 2009, Johnson et al., 2011b). The interindividual variation often seen in response to AEDs between individuals who appear to have the same epilepsy phenotype suggests that genetic factors could also contribute to the AED responsive and resistance phenotypes (Sisodiya, 2003, 2005). The influence of genes on outcome of drug treatment is a rapidly evolving field (Weinshilboum, 2003) and may help to explain this variability in clinical outcome as well as enlighten the general unpredictability of epilepsy treatment (Spear, 2001, Sisodiya, 2005).

1.5 Pharmacogenetics

Even with the medical advances of the 20th Century, optimal drug treatment remains elusive for many of the world's common and high impact diseases including hypertension, cancer and diabetes (McLeod and Evans, 2001, Spear et al., 2001). The efficacy and toxicity of many major therapeutic agents is substantially heterogeneous when viewed across the globe, thus treatment response and failure in patient groups is hugely unpredictable (Mancinelli et al., 2000, McLeod and Evans, 2001, Shastry, 2006). Any given drug can be effective in some patient groups but ineffective in others, and some individuals experience AEs and/or ADRs whereas others are unaffected. This interindividual variability in response to most, if not all, drugs is well known and poses a serious problem in current medical treatment (Wolf et al., 2000, McLeod and Evans, 2001).

Potential causes for differences in drug response include the nature and severity of the disease being treated, the individual's age and race, organ function, concomitant therapy, drug interactions, and concomitant illnesses (Evans and Johnson, 2001). However, even when these factors are taken into account, considerable variation remains unexplained (Vinken et al., 1999). Over the past decades, much evidence has emerged indicating that a significant portion of this variability is genetically determined (Vinken et al., 1999, Mancinelli et al., 2000).

1.5.1 The history of pharmacogenetics

Although genetic differences among people have long been recognised to influence drug response phenotypes of individuals, research efforts have only begun to focus on the genetics of drug response in the last few decades, with previous focus being largely orientated towards disease predisposition (Goldstein, 2005). Leveraging the knowledge of an individual's genetic makeup initially gave the possibility to predict susceptibility to monogenic diseases and later proved particularly valuable for common diseases with a complex multifactorial nature (McLeod and Evans, 2001). A similar approach has since been taken for predicting the complexity of response to particular treatments, in order to match patients with the right medications given at the right doses (Goldstein, 2005). The relatively new field of studying a patient's response to a specific drug alongside their genetics is known as pharmacogenetics (PGx) (Weinshilboum, 2003).

PGx is characterised by the profiling of differences between individuals' DNA to identify definitive relationships between the structure and function of pharmacologically relevant genes and the drug response phenotypes observed in patients (McLeod and Evans, 2001). The overall goal of PGx is to better understand the genetic variation that determines heterogeneity in drug effects, to predict how individuals may respond to drugs, and to ultimately translate this to clinical practice (Evans and Relling, 2004). In the last ten years the

remarkable progress in human genomics and molecular genetics has initiated a surge in PGx research (Ferraro and Buono, 2005) and the discovery of genetic markers of phenotypes for response to medications has become one of the fastest growing fields in clinical and translational biomedical research (Wang et al., 2011).

PGx originated in the 1950's where early clinical observations identified patients with large differences in their response to "standard" drug doses, often with individual variations in plasma or urinary drug concentrations. This was followed by the discovery of drug-metabolizing enzymes (DMEs) and the realisation that such variation was largely due to inherited differences in metabolism (Weber, 2001, Weinshilboum and Wang, 2006). The genes of DMEs formed the first candidate genes for variable drug response. DME genes encode enzymes that metabolise drugs and their products and have evolved to neutralise toxins and/or to control concentrations of signaling molecules in endogenous pathways (Nebert and Dieter, 2000, Goldstein et al., 2003).

Researchers first described the role of genetic differences in determining biochemical traits through the disposition of succinylcholine, isoniazid, and antimalarial drugs such as primaquine (Weinshilboum, 2003). They were able to distinguish that prolonged paralysis following the use of succinylcholine was the result of a variant of the butyryl-cholinesterase enzyme, that hemolytic anemia due to the antimalarial drug primaquine resulted from a variant form of the enzyme glucose-6-phosphate dehydrogenase, and that peripheral neuropathies caused by the anti-tuberculosis drug isoniazid were a consequence of genetic differences in the enzyme N-acetyltransferase (Weber, 2001, Johnson, 2003, Weinshilboum, 2003).

The field of PGx has since developed and expanded to cover more complex drug response phenotypes (Goldstein, Need et al. 2007). Today PGx is considered a rational and systematic genetic approach to identifying specific genetic sources of drug response variability (Evans and McLeod, 2003, Weinshilboum, 2003). Understanding the molecular basis and functional consequences of these genetic variants on drug response phenotypes has the potential to enlighten the use of many medications, optimising their efficacy and preventing toxic effects during routine drug therapy (Evans and Relling, 1999).

1.5.2 The potential for tailored drug therapy

Decisions about the choice of drug and appropriate dosage are at present largely based on information derived from population averages (Mancinelli et al., 2000). As polymorphisms with functional consequences are identified, the potential for clinicians to utilise interindividual variation in a patient's genetics in the clinical setting for a more personalised approach to treatment becomes markedly greater (Vinken et al., 1999, Feero et al., 2010, Wang et al., 2011). This involves classifying patients with the same phenotypic disease profile into

smaller sub-populations, defined by genetic variations associated with disease, drug response, or both, with the assumption that drug therapy in these genetically defined sub-populations may be more efficacious than treating a broad population (Mancinelli et al., 2000).

Advances in genetic testing, and their transference to the clinic, generates the possibility of more effective dosing of medications across various therapeutic areas (Meisel et al., 2000, Johnson, 2003, Bhathena and Spear, 2008). An individualised approach to treatment decisions may also lead to improved tolerability to medications thus can enhance regimen adherence, improve drug safety and ensure optimum drug therapy across patient groups (Spear et al., 2001). The identification of genomic predictors for treatment response may additionally help with the discovery and development of new medications (Evans and Relling, 1999). Overall there is increasing evidence that PGx will continue to be extremely important in health care in the near future (Wolf et al., 2000) (Roses, 2000).

1.5.3 Principles of pharmacogenetics

Pathways controlling drug disposition or drug PK describe the course of a drug and/or metabolite through the body (Rang, 2003), whilst pathways controlling the efficacy of a drug, or drug pharmacodynamics (PD), refers to the relationship between the drug and its effect at target sites (Evans and McLeod, 2003, Rang, 2003). Drug PK pathways encompass the combined processes of drug uptake or absorption, drug distribution, protein binding, drug metabolism, and drug excretion (Evans and McLeod, 2003, Rang, 2003, Weinshilboum, 2003). PD pathways involve processes of drug interaction with therapeutic targets at the cellular level and the effect of drugs on the body, i.e. any resulting biological or therapeutic outcomes (Evans and McLeod 2003).

Genetic variation can affect the genes encoding proteins influencing both drug PK and PD processes (Goldstein et al., 2003, Goldstein, 2005, Nebert, 2008). This mainly consists of i) genes that encode DMEs and transporters which function in drug elimination and distribution/excretion respectively and ii) genes encoding channels, receptors and/or enzymes on which drugs act or modulate to elicit their effects (Roden and George, 2002, Goldstein et al., 2003, Bhathena and Spear, 2008). Some genetic variation within these genes can potentially affect protein function or expression thus can influence a drugs normal disposition or action. Many of the PK and PD proteins for a particular drug thus form key components to PGx research (Goldstein et al., 2003, Johnson, 2003).

1.5.4 Genetic variation

Any two individuals are greater than 99% identical in their DNA sequence (www.hapmap.org), however much variation exists between individuals (Nebert, 2008).

Genetic variation refers to the differences in DNA sequences between individuals, of which there are many types (Jazwinska, 2001), with single nucleotide polymorphisms (SNPs) being among the most common sources of naturally occurring variation in the human genome (McLeod, 2005, Orr and Chanock, 2008). Alternative classes of DNA variation include microsatellites, copy number variants (CNVs), insertion/deletion polymorphisms and mutations (Roses, 2000, Jazwinska, 2001, McLeod, 2005, Orr and Chanock, 2008). For common diseases, genome-wide linkage studies have had limited success, due to their complex genetic architecture (Sachidanandam et al., 2001). In the human population most variant sites are rare, but common polymorphisms can explain most of the heterozygosity (Inaba et al., 1995, Beckmann et al., 2007). There is clear evidence that common gene variation contributes to complex traits including drug response phenotypes and with the ease of studying these, common variants have dominated PGx thus far (Jazwinska, 2001, Johnson, 2003, Goldstein, 2005, Ferraro et al., 2012).

1.5.5 Single Nucleotide Polymorphisms

A significant effort towards large-scale characterisation of human SNPs has been initiated in the last decade (Brookes, 1999). The Human Genome Project (HGP) launched in the USA in the 1990s, was a multi-country effort to sequence the entire human genome in order to identify and catalog genetic similarities and differences in human beings (www. http://hapmap.ncbi.nlm.nih.gov)(Sachidanandam et al., 2001). Advances in technologies have since allowed genetic association studies in complex diseases/traits to take advantage of results of the HGP (McPherson et al., 2001).

SNPs comprise a large set of bi-allelic genomic variants (single base pair changes) of which there are an estimated 10 million in the human genome and they appear approximately every 300 base pairs (bp) on average and most commonly, these variations are found in the DNA between genes (http://ghr.nlm.nih.gov/handbook/genomic research/ snp). SNPs account for at least 90% of human sequence variation in the human genome (http://ghr.nlm.nih.gov/ handbook/genomicresearch/snp)(Pang et al., 2009) with the rest attributable to insertions or deletions of one or more bases, repeat length polymorphisms and rearrangements (Sachidanandam et al., 2001). As SNPs are extraordinarily abundant they offer a powerful means of assessing genetic association, allowing essentially any gene to be explored for variants that may associate with a disease or traits (Ferraro et al., 2012).

1.5.6 Genetic markers

The majority of genetic variation in the human genome is not pathogenic or of any biological significance (McCarthy and Hilfiker, 2000). The challenge for association studies is to identify the most influential polymorphic alleles. To sequence variants across whole genomes or all SNPs in a pathway of candidate genes is impractical. Most association studies genotype only a small proportion of marker SNPs in a target region (be that the whole genome, or within a set of candidate genes). Because alleles at different loci are sometimes found together more (or less) often than expected by chance based on their frequencies, non-random association can exist between allelic variants or SNPs in proximity to each other (Wall and Pritchard, 2003). These SNPs also tend to travel together in blocks through evolutionary time, a phenomenon known as linkage disequilibrium (LD)(Hirschhorn and Daly, 2005). Genomic patterns of LD are used to select a set of marker SNPs, known as tagging SNPs (tSNPs) that are statistically associated with other SNPs in the genome. Tagging SNPs alone are then typed to economically represent genomic variation across the entire region of interest (Goldstein et al., 2003, Wall and Pritchard, 2003, Hirschhorn and Daly, 2005).

Association studies are greatly facilitated by LD-based methods (Hirschhorn and Daly, 2005) that systematically represent variation in candidate genes (or the whole genome) (Goldstein et al., 2003). The more recent possibility of determining LD patterns on a genome-wide scale through the HapMap project has allowed the economical representation of genomic variation as a whole and enabled more efficient genome-wide research (Goldstein et al., 2003, Hirschhorn and Daly, 2005). If a risk polymorphism exists it will either be genotyped directly (as a selected marker or tSNP) or be in strong LD with a genotyped tSNP (Collins et al., 1997, Kruglyak, 1999, Servin and Stephens, 2007). Genetic variants found to be associated with a disease or trait using LD based studies may thus not be directly causal or influential, but may be statistically correlated (in LD) with an another important variant (McCarthy and Hilfiker, 2000, Goldstein and Weale, 2001, Mullen et al., 2009).

1.5.7 Genomic location of polymorphisms and functional affect

Not all polymorphisms are functional i.e. have the potential to cause a biological change (Harley and Narod, 2009). The position and type of SNP usually defines most of their biological effect: SNPs may occur in the coding portion of the genes (exons), intervening sequences (introns) or between two genes (intergenic regions)(Shastry, 2002). Most SNPs (around 75%) occur in non-coding regions and are of unclear consequence (Sachidanandam et al., 2001, Harley and Narod, 2009). These include introns found within genes as well as intergenic regions between genes, and form the majority of the human genome (Tabor et al., 2002). Although like intronic variants intergenic region SNPs also have no known function

(Tabor et al., 2002), some are also thought to impact gene expression or splicing (McLeod, 2005). Most recent research by the Encode Project (www. concerning non-coding regions has confirmed that much non-coding DNA has a regulatory role (Birney et al., 2007). The remaining 25% of variants occur in gene coding regions (exons) (Harley and Narod, 2009).

Only 50% of exonic SNPs result in an amino acid change (non-synonymous SNPs) that can potentially alter protein function (Shastry, 2002). The remaining 50% are synonymous SNPs and also result in a nucleotide change, but because of redundancy in codon usage, these have a neutral substitution that may not affect protein function (Shastry, 2002, Shastry, 2004, Harley and Narod, 2009). Non-coding SNPs can also alter protein function by altering the regulation of gene expression (Shastry, 2003, Harley and Narod, 2009, Pang et al., 2009). SNPs in the promoter region may alter promoter activity thus affecting gene transcription and SNPs close to binding sites for splicing machinery may alter RNA splicing, and subsequently affect amino-acid transcription (Sauna et al., 2007, Hunt et al., 2008, Harley and Narod, 2009)(Gray et al., 2000, Harley and Narod, 2009).

SNPs provide a powerful tool for association of loci at specific sites in the genome with complex traits (Bentley, 2000, Risch, 2000b) however most SNPs are not directly associated with causing disease instead they represent useful biological markers for analysing a particular disease or trait (Judson et al., 2000). It has been estimated that there are 50,000–250,000 SNPs that have a biological effect on one or more of the estimated 30,000 genes (Risch, 2000a) and in some cases, the biological effect may increase susceptibility to one or more diseases (Gray et al., 2000).

Due to their prevalence SNPs have been the variant type of choice for association studies in common diseases and complex traits (Beckmann et al., 2007). In addition to SNPs as a common source of genetic variation, the rare variant hypothesis has emerged. It is thought that multiple rare SNPs or additional variants of low genomic frequency may be the drivers of disease, and this has recently emerged in genomic studies for several complex traits (Tabor et al., 2002, Ferraro et al., 2012), with the theory that variants with very severe functional consequences are usually more infrequent. In addition to SNPs the human genome also contains another abundant source of polymorphism resulting in larger insertions, deletions or duplications known as copy number variations (CNVs). At least 10% of the genome is subject to copy number variation. CNVs are far less numerous to SNPs but can affect from one kb to several mega bases of DNA per event, adding up to a significant fraction of the genome, and so more likely to have a functional role in the aetiology of a trait (Beckmann et al., 2007). Several complex disorders have already been associated with CNVs including susceptibility to HIV-1, lupus and Crohn disease and are expected to potentially impact other complex traits including the inter-individual drug response (Ouahchi et al., 2006) as well as susceptibility to infection or cancer (Beckmann et al., 2007).

1.5.8 Candidate gene approach

The majority of research studies for complex disease traits have focused on the use of *a priori* hypotheses generated from knowledge of the pathways underlying disease traits for the selection of genomic variants for investigation (Tabor et al., 2002). This approach provides a narrow spectrum of candidate genes that are selected due to their potential role in the aetiology of the disease and are used for investigating the genetic influence on a complex trait (Tabor et al., 2002).

Rather than rely on markers that are evenly spaced throughout the genome without regard to their function or context in a specific gene the candidate approach focuses on the biological understanding of the phenotype, tissues, genes and proteins that are likely to be involved in the disease/trait (Tabor et al., 2002). Thus far the candidate gene approach has been widely and frequently used as a design strategy in PGx studies, using knowledge of pharmacological action, drug disposition and/or disease pathogenesis for gene selection and so adopting *a priori* hypotheses about the origin of the inherited variability in drug response (Evans and Johnson, 2001, Roden and George, 2002, Goldstein et al., 2003, Daly, 2010b).

Gene-association studies usually determine whether there are differences between case and control groups (i.e. phenotype groups; drug responders versus non-responders) with regards to the prevalence of a potentially functional and phenotypically influential gene variant (Tabor et al., 2002). SNP genotyping is a tool for genetic analysis that is used for uncovering the association of an allele(s) at specific locus in the genome with the potential to cause a change in protein expression or function, with diseases or phenotypic traits such as drug response (Shi et al., 1999, Bentley, 2000). This typically involves genotyping candidate gene variants in clinically relevant populations and comparing the frequencies of the alleles or genotypes at the site of interest in both patient groups (Shi et al., 1999, Tabor et al., 2002). Due to the complexity of drug response, PGx studies can involve multiple candidate gene-association studies and these can involve the use of hundreds of genes and 1000's of SNPs (Wang, 2010).

1.5.9 Polygenetics in drug response and the emergence of pharmacogenomics

The genetic basis of many monogenic (single gene mutation) rare inherited disorders is known (over 6000 rare monogenic disorders and their genes successfully identified thus far), (http://www.ncbi.nlm.nih.gov/omim). The attention of human genetics has now shifted towards the basis of more common complex diseases or traits with multiple genetic (polygenic) and environmental components contributing to susceptibility (Hirschhorn et al., 2002). Causal alleles for monogenic disorders are highly penetrant and often lead to severe phenotypes (Hirschhorn and Daly, 2005). By contrast, the alleles that underlie complex traits

usually have more subtle effects on disease risk and may involve non-coding regulatory variant alleles that are likely to have a modest impact on protein expression (Hirschhorn and Daly, 2005). The phenotype for complex traits is determined by the sum total of, and/or interactions between, multiple genetic and environmental factors. Each of the many genetic determinants are expected to make only a small contribution to overall heritability (Hirschhorn and Daly, 2005) owing to their multifactorial nature (McCarthy and Hilfiker, 2000) and subsequently have a relatively small individual effect on disease risk (Jazwinska, 2001, Reich and Lander, 2001, Weinshilboum, 2003). This is referred to as the common disease/common variant (CDCV) hypothesis (Carlson et al., 2004). The inherited basis of drug response has similarly been difficult to elucidate especially for drugs whose PD and/or PK pathways are poorly defined (Nebert, 2008, Nebert et al., 2008a, Goldstein, 2009).

Because most drug effects are determined by several gene products that influence the metabolism, disposition and efficacy of medications, inherited differences in these PK and PD genes have increasingly been shown to alter drug response and there is a growing perspective that the inherited basis of drug action is polygenic in nature (Johnson and Evans, 2002, Evans and McLeod, 2003). As drug response is equally as complex as disease genetics (McCarthy and Hilfiker, 2000), relative risk estimates for genetic influences on drug action are also expected to be low, owing to its multifactorial nature (McCarthy and Hilfiker, 2000). In support of this, several pharmacogenomic markers have been identified to date, each of which confers only about a two-fold increased likelihood of response due to common allelic variants (Poirier et al., 1995, Drazen et al., 1999, McCarthy and Hilfiker, 2000).

Polygenic determinants of drug effects have accordingly become increasingly important for PGx (Evans and Relling, 1999) and PGx research has recently transformed into a genomics-based field, (Evans and Johnson, 2001) leading to a new term, pharmacogenomics. The field of "pharmacogenomics" aims to utilise a genome-wide approach to identify the network of genes that govern an individual's response to drug therapy (Goldstein et al., 2003, Daly, 2010b). With current advances in genomic technology providing more sophisticated molecular tools for the detection of genetic polymorphisms and the wealth of new data emerging from the HGP, scanning the whole genome to identify and directly examine numerous common gene variants for any association with clinical response phenotypes (drug efficacy and toxicity) has rapidly become viable, and forms the basis of pharmacogenomics research (Nebert, 1999, Evans and Johnson, 2001).

1.5.10 Whole genome association approach

In spite of their success in the identification of genes with important contributions to drug response (Grant and Hakonarson, 2007, Daly, 2010a), candidate-gene studies have been

subject to several criticisms, the most important being that i) many significant findings of association in candidate-gene studies have not been replicated when followed up in subsequent association studies and ii) though candidate-gene studies are based on the ability to predict plausible candidate genes and variants through assumed functional potential, current knowledge is not always sufficient to make these predictions. This is evident in AED PGx, where several AEDs have broad mechanism of action and for several AEDs pharmacological pathways are not completely characterised (Mann and Pons, 2007, White et al., 2007). As only a small number of genes can be studied at a time it is difficult to isolate the numerous genetic variants that are suspected to influence complex traits (Goldstein et al., 2003) and these may therefore only represent a fraction of the possible genetic risk factors that are actually involved in drug effects (Hirschhorn and Daly, 2005).

Because of this progression in genomic research, a whole genome approach has rapidly become an alternative methodology to identify novel associations with common diseases (Hirschhorn and Daly, 2005) and genome-wide association studies (GWAS) have increasingly been applied to pharmacogenomics (Grant and Hakonarson, 2007, Gurwitz and McLeod, 2009, Daly, 2010a). The conventional genome-wide association (GWA) approach is a hypothesis-free, systematic search of SNPs (to function as genetic markers) across the genome (Guessous et al., 2009)(www.genome.gov/GWAStudies).

As no assumptions are made in GWAS with regards to the genomic location of potentially causal or influential variants, this approach represents an unbiased yet comprehensive option that can be attempted even in the absence of convincing evidence regarding the function or location of the causal genes (Hirschhorn and Daly, 2005). GWA approaches enable the detection of novel and less obvious genes, and this may be particularly useful for pharmacogenomics research into drug-target genetics, which is less well understood than that of drug metabolism (Daly, 2010a). Figure 1.1 summarises the main advantages and disadvantages of both the candidate gene and GWA approach to genetic association studies.

Since 2007, a range of pharmacogenomics GWA studies have been published (di Iulio and Rotger, 2012). These have either identified novel associations between drug responses and clinically relevant loci, or have confirmed previous associations (Daly, 2010a) (Table 1.4).

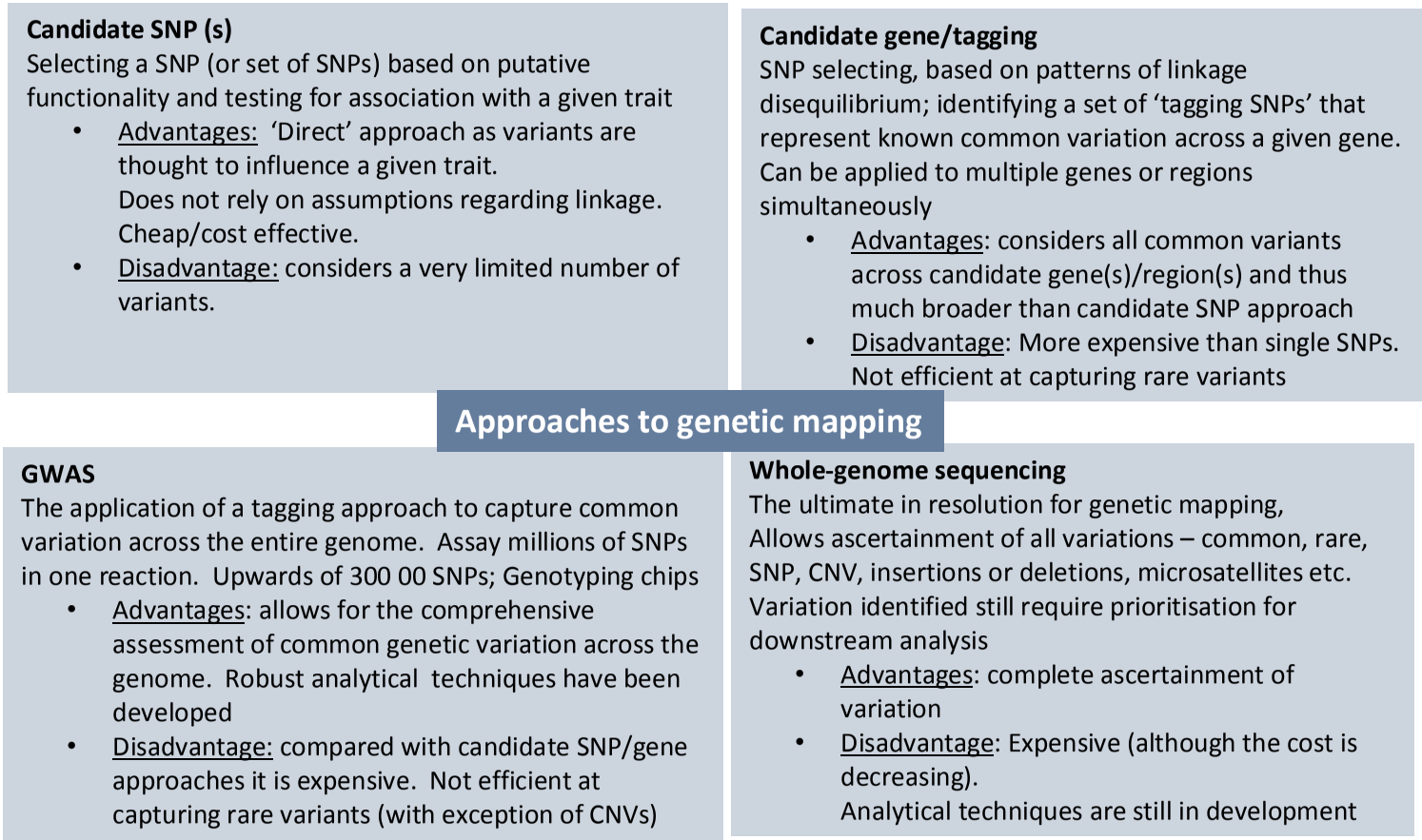


Figure 1.1 Approaches to genetic mapping adapted from Cavalleri *et* al. 2011
Brief description of methods the four main methods for studying the genetic basis of common, complex disease and/or traits. Studies to date have for the majority employed a candidate gene approach, however GWAS and whole genome sequencing have increased exponentially over the last decade

Table 1.4 Pharmacogenetic tests integrated into drug labels

Examples of genome wide association studies for some commonly used drugs. Data taken from Daly *et* al 2010, Wang *et* al 2011, di lulio and Rotger, 2012.

Drug type	**Drug Name**	**Pharmacogenetic effect**	**Gene Association**	**Study reference**
Anticoagulant Drug	Warfarin	Dose adjustment according to SNP genotype	*CYP2C9*,*VKORC1*, and *CYP4F2 genotypes*	Cooper et al, 2008 Takeuchi et al, 2009 Cha et al, 2010
Anti-platelet drug	Clopidogrel	Diminished anti-platelet effect with clopidogrel treatment and poorer cardiovascular	CYP2C19 Variant	Shuldiner et al, 2009
New Associations				
Antibiotic	Flucloxacillin	Flucoloxacillin-induced Liver injury	Allele HLA*B5701	Daly et al, 2009
Statin	Simvastatin	Simvastatin-induced myopathy	Allele SLCO1B1*5	Link et al, 2008
Anti-viral	Ribavirin	Hepatitis C treatment-induced clearance	Genetic variants in IL28B in hepatitis C virus (HCV) genotype 1 patients	Ge et al, 2009 Suppiah et al, 2009 Rauch et al, 2010 Tanaka et al, 2009

1.5.11 Clinical application

In spite of the significant association in drug response and disease that have been described over the years (Wang et al., 2011) the translation of these pharmacogenomics/PGx discoveries to the clinic has not been as rapid as was hoped. Only a few commercial tests are currently available (Evans and Relling, 2004, Weinshilboum and Wang, 2006, Swen et al., 2007). Recent instances of PGx information for individualised treatment (Table 1.5) include, trastuzumab treatment for HER2 overexpressing breast cancer, (Zanger, 2010) and more recently HLA-B*5701 testing to avoid abacavir hypersensitivity (di Iulio and Rotger, 2012).

One of the main challenges faced by PGx research is that the influence of genetic markers on therapeutic outcome is often lacking and/or non-reproducible (Colhoun et al., 2003, Swen et al., 2007). Past research indicates that of the 166 putative associations that have been studied three or more times, only 6 have been consistently replicated (Hirschhorn et al., 2002). Several reasons can be attributed to this irreproducibility that characterises genetic association studies whether they have employed a candidate gene or whole genome approach (Hirschhorn et al., 2002, McCarthy and Hirschhorn, 2008) and these include; i) the heterogeneity of disease phenotype, ii) the underestimation of the complexity of common complex traits iii) the small effect sizes of alleles of common risk variants, and iv) relatively small numbers of patients in PGx studies (Johnson, 2003, Evans and Relling, 2004).

1.6 Epilepsy pharmacogenetics

Inadequate seizure control (Kwan and Brodie, 2000a), AEs, ADRs (Depondt, 2006b, Zaccara et al., 2007) and variability in individual responses to the same AED doses (Loscher, 2002) encapsulate pharmacotherapy for a number of people with epilepsy (Depondt, 2006b) and represent global barriers to optimal AED treatment (Sisodiya 2005). Furthermore, the optimal doses of AEDs may differ four-fold among individuals (Loscher et al., 2009). The recent expansion of the field of PGx has allowed the study of drug response in a number of common complex traits across several disease domains including epilepsy (Depondt, 2006a). Epilepsy represents an ideal disease for PGx study due to its high prevalence, wide variety of phenotypes, variable treatment outcomes and at least some knowledge of the main pathways of drug action and drug distribution (Depondt, 2006b).

In current clinical practice AEDs are primarily used according to existing guidelines for the management of epilepsy in the general population and they are selected on the basis of known drug response profiles as well as patient and disease characteristics (Brodie and French, 2000, Perucca et al., 2001, Sander, 2004, Schachter, 2007). In the UK the NICE guidelines (http://guidance.nice.org.uk/CG20/Guidance) are referred to for general AED prescribing. Basing AED choice on anticipated efficacy is somewhat empirical and initial AED selection

using this approach is currently effective in around 50% of patients (Kwan and Brodie, 2000a, Depondt, 2006b).

In parallel to the difficulties inherent with basing AED choice on anticipated efficacy, dosing decisions are also largely reliant upon trial and error (Dlugos, Buono et al. 2006). For most AEDs a broad range of doses is used in clinical therapy and final maintenance doses are reliant on individual response. PGx thus additionally offers the potential to influence AED dosing regimens by perhaps using patient genotype to predict a patient's optimal dose for seizure control without causing ADRs and also how quickly drugs can be titrated up (Depondt, 2006b, Dlugos et al., 2006, Duncan et al., 2006). Furthermore, as a significant number of individuals continue to experience seizures despite multiple drug treatment, an increase in the understanding of epilepsy and drug-action mechanisms, may shed light on the genetic factors contributing to refractory epilepsy (Kasperaviciute and Sisodiya, 2009). PGx research might similarly predict which patients are likely to become refractory to drug treatment early on during the course of disease and this will encourage early surgical consideration to improve the overall outcome for these difficult to treat individuals (Depondt and Shorvon, 2006). PGx research could likewise provide a more rational basis for selecting AEDs at the outset of therapy (Dlugos et al., 2006).

1.6.1 Candidate genes for epilepsy pharmacogenetics

In line with PGx research in other disorders, most association studies in epilepsy to date have concerned candidate genes and focused on genetic variation across PK and PD pathways (Depondt, 2006a). Much of the PGx data in the field of epilepsy deals particularly with the PK of AEDs and PGx knowledge beyond PK genes (i.e. DMEs and transporters) remains limited (Depondt, 2006a, c, Kasperaviciute and Sisodiya, 2009). Since drug response to most AEDs is multifactorial, whole genome screening is expected to be more fruitful in epilepsy PGx than selecting potential candidate genes for AED response (Depondt, 2006b). The present direction of epilepsy PGx is therefore moving rapidly towards whole genome strategies to investigate common genetic polymorphisms (Kasperaviciute and Sisodiya, 2009).

There are three important categories of candidate genes with a potential influence on AED response that have been studied in epilepsy PGx to date; i) genes encoding drug transporters of which AEDs proposed substrates; ii) genes encoding DMEs involved in the breakdown of AEDs; and iii) genes encoding AED targets and their related pathways (Depondt, 2006b, Klotz, 2007).

Table 1.5 Pharmacogenetic tests integrated into drug labels

Success stories for pharmacogenetics; examples of drugs for which biomarkers of clinical relevance have been identified. Table taken and adapted from (Tauser, 2012).

Drug indication	Pharmacogenetic biomarker	Comments
Mandatory, required predictive pharmacogenetic tests in drug label		
Trastuzumab HERCEPTIN® Metastatic BC	HER2/neu over-expression	Improve drug efficacy: clinical benefit is limited to the responsive patients, whose tumors overexpress the drug-target HER2/neu
Lapatinib TYKERB® Metastatic BC	HER2/neu over-expression	Improve drug efficacy: clinical benefit limited to tumors overexpressing HER2/neu
Cetuximab ERBITUX® Metastatic CRC	EGFR expression	Improve drug efficacy: clinical benefit limited to patients with EGFR-positive tumors
Dasatinib SPRYCEL®; Imatinib GLEEVEC® ALL (adults)	Philadelphia chromosome positive	Disease confirmation and patients' selection: BCR-ABL translocation
Maraviroc SELZENTRY® HIV (adults)	CCR-5 C-Cmotif receptor	Disease confirmation: infection with CCR-5-tropic HIV-1 and resistance to other antiretrovirals
Recommended predictive pharmacogenetic tests in drug label		
Warfarin COUMADIN® Thrombo-embolism	CYP2C9 and VKORC1 (-1639G>A)	Improve drug efficacy and safety: avoid increased risk of bleeding to patients homozygous or heterozygous for CYP2C9*2 or CYP2C9*3 alleles by prescribing differentiated doses Pharmacogenetic test: "Nanosphere Verigene Warfarin Metabolism Nucleic Acid Test; therapeutic algorithm based on genotype and clinical factors (http://www.WarfarinDosing.org.)
Carbamazepine TEGRETOL® Epilepsy	HLA-B*1502 allele	Improve drug safety: avoid serious dermatologic reactions (Stevens–Johnson syndrome and/or toxic epidermal necrolysis).

Available from: http://www.intechopen.com/books/clinical-applications-of pharmacogenetics/ pharmacogenomics-matching-the-right-foundation-at-personalized-medicine-in-the-right-genomic-era-

1.6.2 Novel computational methods; an approach to solving issues in complex data analysis

Linkage analysis has been successfully used by statisticians to locate genes responsible for simple monogenic diseases (Rodin et al., 2011). Unlike rare Mendelian diseases however, multiple genes are likely to influence or confer susceptibility to common complex diseases and traits. Interactions between these genes and between genes and the environment also exist (Ritchie and Motsinger, 2005, Rodin et al., 2011), the cumulative effect of which is thought to contribute to complex disease phenotypes. This multifactorial basis of complex disease has led to several difficulties in data analysis, mostly because of statistical and computational issues (Ritchie and Motsinger, 2005). In complex disease analysis a relatively large number of genetic variants are investigated for disease association in a bigger sample size of individuals, leading to significant statistical concerns (Zhang and Rajapakse, 2009, Motsinger-Reif et al., 2010). Such analytical issues have thus far made the identification of definitive influential factors for many disease traits difficult (Moore et al., 2004, McCarthy et al., 2008). With complex diseases and traits having potential genetic contributions from thousands of variants and with current genotyping technology reporting millions of polymorphisms, the statistical challenge of detecting small polygenic effects using large volumes of genetic data whilst also controlling for false positive signals has become apparent (Baksh and Kelly, 2007).

Traditional parametric statistical approaches for gene discovery and genetic analysis, such as logistic regression, typically evaluate the effects of individual SNPs in isolation, thus assuming independence between variants (Risch and Merikangas, 1996, Hoppe, 2005). Such marker-by-marker approaches ignore the multigenic nature of complex disease and also fail to account for the interplay of many genes that is likely to contribute to the genetic composition of complex traits (Hirschhorn et al., 2002, Pander et al., 2010, Rodin et al., 2011, Vanneschi et al., 2011). (Ritchie and Motsinger, 2005). This illustrates the challenge of analysing data for complex traits and the need for accurate classification and prediction algorithms.

1.6.3 Pharmacokinetic variation and metabolising enzymes

More than 30 families of DMEs can be found in humans (Evans and Relling, 1999, Ingelman-Sundberg et al., 1999) and nearly all of these have genetic variants, many of which translate into functional changes in the proteins they encode (Weinshilboum, 2003). Hepatic metabolism consists of two established phases; phase I reactions (oxidation, reduction and hydrolysis) and phase II reactions (conjugation reactions between an endogenous molecule such as glucuronic acid and a drug metabolite) and both pathways function to produce

metabolites that are usually more water soluble than the parent compound, thus enhancing their excretion from the body (Nagasawa and Nakahara, 1992).

Metabolism in the liver by the cytochrome P450 (CYP) metabolising enzymes represents the most common route of drug turnover, and it has long been known that fast- and slow-metabolising variants in the genes encoding these enzymes can lead to under- and over-dosing of drugs (Evans and Johnson, 2001, Wilkinson, 2005). The CYP super-family is thus considered the most important class of DMEs and up to 80% of all prescribed drugs undergo initial metabolism (Phase I reactions) through oxidation reactions catalysed by these enzymes (Eichelbaum et al., 2006).

Inherited differences in individual DMEs are typically monogenic traits, and the clinical importance of enzyme variants, i.e. their influence on the therapeutic effects of medicinal drugs, depends on allele-frequency, the effects of the polymorphisms on protein function (i.e. whether the biological activity of the enzymes are altered), and the importance of the enzyme for the activation or inactivation of drug metabolites (Evans and Johnson, 2001, Kirchheiner and Seeringer, 2007). Polymorphic CYP450 enzymes can either; reduce enzymatic activity to slow down metabolism and cause an over-accumulation of a drug or its metabolites resulting in drug toxicity (Park et al., 1995, Kitteringham et al., 1998) or reduce efficacy of medications that require a polymorphic enzyme for activation as this can reduce it's function (Kitteringham et al., 1998, Evans and Johnson, 2001).

The metabolic pathways involved in the elimination of most AEDs have largely been defined (Ramachandran and Shorvon, 2003, Saruwatari et al., 2010). Functional polymorphisms in the genes of AED metabolising enzymes are expected to give rise to interindividual differences in metabolic profile and to influence drug levels in the plasma. This can also lead to differences in AED efficacy and/or toxicity (Ramachandran and Shorvon, 2003). Detoxification of AEDs occurs via hepatic (metabolism) and/or renal (excretion) routes (Klotz, 2007, Anderson, 2008). Most AEDs are eliminated from the body initially through biotransformation in the liver by several different DMEs before their elimination via the kidneys (Klotz, 2007, Anderson, 2008), Due to the major role of CYP450 genes as Phase I metabolising enzymes for many AEDs, these may influence interindividual variability in the PK of AEDs and have been the focus of several candidate gene studies in AED PGx to date (Loscher et al., 2009, Saruwatari et al., 2010).

Functional polymorphisms underlying alleles with variable metabolism rates are known for several CYP450 genes and these variants have the potential to result in interindividual differences in AED concentration, and in their effectiveness and/or the occurrence of ADRs (Daly, 2003). Of the various polymorphic CYP species however, only CYP2D6, CYP2C9 and CYP2C19 have shown any clinical significance to date (Ingelman-

Sundberg, 2004b, Klotz, 2007) and only CYP2C9 and CYP2C19 variants are relevant to AED metabolism (Kirchheiner and Seeringer, 2007, Klotz, 2007).

Polymorphisms are also known to exist in all the major phase II enzyme systems including N-acetyltransferases (NAT1 and NAT2), uridine glucoronyltransferases (UGTs), sulfatases and glutathione-s-transferases (GSTs) (Ferraro and Buono, 2005), however the contribution of phase II enzymes to AED metabolism is currently less-well characterised than that of phase I (Ferraro and Buono, 2005, Depondt, 2006b). The UGT superfamily of conjugating enzymes are one of only a few phase II DMEs that are known to contribute to AED metabolism (Ferraro and Buono, 2005, Depondt, 2006b). Members of the UGT1 superfamily act upon approximately 35% of all drugs metabolised by phase II DMEs, including several AEDs such as CBZ, VPA, LTG, OXC, TPM and ZNS (Ferraro and Buono, 2005, Szoeke et al., 2006, Saruwatari et al., 2010). UGTs conjugate their substrates with a glycosyl group (glucuronidation), a major conjugation pathway responsible for increasing the water solubility and enhancing the elimination of a variety of drugs (Nagar and Blanchard, 2006, Saruwatari et al., 2010). Genetic polymorphisms in UGT enzymes may modify their glucuronidation capacity, a phenomenon seen in an increasing number of studies of a variety of substrates (Inaba et al., 1995, Miners et al., 2002, Guillemette, 2003). Most of the metabolism of VPA and LTG occurs via this glucuronidation pathway, rather than via the CYP450 enzymes (Nagasawa and Nakahara, 1992). Knowledge of the genetic mechanisms underlying variability in glucuronidation capacity is however currently limited and only a few clinically relevant genetic polymorphisms in UGTs have been described thus far (Guillemette, 2003).

Additional phase II enzymes with a major role in AED metabolism include GSTs (Guillemette, 2003, Hayes et al., 2005) and microsomal epoxide hydrolase (mEH) (Depondt, 2006a, Klotz, 2007, Saruwatari et al., 2010). GST is an essential enzyme of defense and detoxification and another hepatic conjugating phase II enzyme. GSTs catalyse the conjugation of glutathione (GSH) for detoxifying and aiding the elimination of a wide range of therapeutic agents (Whalen and Boyer, 1998, Hayes et al., 2005, Saruwatari et al., 2010) and play an important role in metabolising AEDs (Tang et al., 1996, Bu et al., 2007, Shang et al., 2008). The mEH enzyme encoded by the EPHX1 gene is a biotransformation enzyme that also metabolises reactive epoxide intermediates (often formed during phase I metabolism) to more water-soluble derivatives and is a candidate for variation in response to CBZ, PB and PHT (Nagasawa and Nakahara, 1992, Depondt, 2006b).

1.6.4 Pharmacokinetic variation and transporter proteins

Drug transporters function to regulate the absorption, distribution, and excretion of many medications (Evans and McLeod, 2003) through regulating both inward and outward transport of drugs and their metabolites (Daly 2010). These proteins also show considerable genetic variation, including many potentially functional polymorphisms (Goldstein et al., 2003, Leabman et al., 2003). The membrane bound efflux transporter super-families form a category of major transport proteins and include ATP-binding cassette (ABC) proteins, and the solute carrier proteins (SLC), with the ABC proteins being among the most extensively studied transporters involved in drug disposition and effects (Borst et al., 2000, Evans and McLeod, 2003). Genetic variation in the genes encoding these proteins are expected to alter the rate of drug uptake, distribution or efflux and can result in variable drug concentrations, effectiveness and/or occurrence of side effects (Goldstein et al., 2003, Cox, 2010).

Functional polymorphisms in genes encoding drug transporters, of which AEDs are proposed substrates, may alter the cerebral uptake, distribution or efflux of AEDs, and thus can result in interindividual differences in the concentration of AEDs in the brain, thereby impacting on effectiveness and/or AEs (Depondt and Shorvon, 2006). The blood brain barrier (BBB) is a physical and metabolic barrier between the CNS and the systemic circulation, which serves to regulate and protect the microenvironment of the brain (Gillham et al., 1990, Scherrmann, 2002). Any therapeutic agents required to target neurological pathways are required to penetrate the BBB to achieve efficacy (Gillham et al., 1990, Scherrmann, 2002).

The ATP-binding cassette or ABC transporter super-family function as active pumps facilitating the efflux of foreign substances from cells across luminal membrane borders (Abbott et al., 2002). Within the ABC superfamily are the multidrug transporter proteins (MDRs), encoded by the ABCB genes, multidrug resistance-associated proteins (MRPs) encoded by the ABCC genes and breast cancer-resistance protein (BCRP) encoded by the ABCG2 gene (Robey et al., 2008). These are expressed in endothelial cells of the BBB and in choroid plexus epithelial cells of the blood-cerebrospinal fluid (CSF) barrier, where they limit brain accumulation of many lipophilic drugs and so appear to provide a defense mechanism to the brain (Fromm, 2000, Loscher and Potschka, 2002).

ABC transporters were initially found to influence clinical refractoriness to the effects of several drugs, including chemotherapeutics for the treatment of cancer (Schinkel, 1997). Although most AEDs are quite lipophilic, allowing penetration into the brain, such multidrug efflux transporters may similarly limit the brain uptake of AEDs by mediating their extrusion (Kwan and Brodie, 2005) and could prevent AEDs from reaching sufficient concentration (Elger, 2003, Kwan and Brodie, 2005, Loscher and Potschka, 2005a, c). Reports of Pgp overexpression in epileptogenic brain tissue promoted it as a candidate gene for refractory epilepsy. Genetic variations in ABC multidrug transporters are thought to play a role in drug-

resistant epilepsy by determining the expression of efflux transporters regulating the levels of AEDs in predisposed individuals (Ramachandran and Shorvon, 2003).

1.6.5 Pharmacodynamic variation and drug target genes

Progress on the PGx of drug target proteins has been slower than studies on drug metabolism and transport. However, the revolution in human genomics has provided new insights into this area (Daly, 2010b). There has been a recent increasing focus on genetic polymorphisms in drug targets, with an interest in defining their impact on drug efficacy and/or toxicity. The main candidate protein categories comprise of receptors, transporters, channel proteins and enzymes and include genes encoding i) direct targets of a drug such as a receptor or enzyme, ii) signal transduction proteins, downstream proteins and other proteins involved in the pharmacological response of a drug and iii) proteins associated with disease risk or pathogenesis that is altered by the drug (Evans and Johnson, 2001).

PD gene variants are often considered as likely causes of variability when drug response appears independent of dose i.e. when PK influences can be ruled out (Vinken et al., 1999). Studies have revealed that the genetic polymorphisms in many PD genes can alter their sensitivity to specific medications (Evans and Relling, 1999). Functional polymorphisms in these genes thus may have a profound effect on drug efficacy and/or toxicity (Johnson, 2001, Roden and George, 2002, Evans and McLeod, 2003).

Review of literature concerning drug target PGx studies reveal that although numerous single gene/single variant associations have been identified, providing 'proof of concept' that genetic variability in PD factors contributes to the variability in drug response, has so far proved unsuccessful. Inconsistencies are evident across studies and the data are not as yet clinically useful in most cases (Evans and Johnson, 2001, Johnson, 2001). Such apparent discordance among studies also suggests the inability of a single polymorphism is highly predictive of response, and thus it seems unlikely that a single polymorphism in a single gene would explain a high degree of drug response variability across drug therapy (Evans and Johnson, 2001, Johnson, 2001). Given that most drug responses involve a large number of proteins, a polygenic, or genomic approach to PGx study may provide more reproducible results (Evans and Johnson, 2001, Johnson, 2001, Goldstein et al., 2003).

Genetic variation in AED target proteins affects the PD of specific AEDs and could potentially contribute to interindividual variation in AED response (Depondt and Shorvon, 2006, Ferraro et al., 2006). AED targets or PD candidate genes, as sources of genetic variation, have only recently been the focus of AED PGx (Depondt and Shorvon, 2006). The main candidates in this category are the genes encoding the targets of currently utilised AEDs, namely neuronal ion channel subunits and elements of neurotransmitter pathways (Kwan et

al., 2001, Depondt, 2006b). Several first-line AEDs including CBZ, LTG and PHT are thought to primarily act through binding to and modulation of voltage-gated Na^+ (Na_V) channel subunits (Ragsdale and Avoli, 1998), therefore the genes encoding neuronal Na_V channels have been prime candidates for PGx study (Depondt and Shorvon, 2006, Loscher et al., 2009).

Mutations in the α-subunit of the Na_V channel were first associated with familial and sporadic epilepsies (Wallace et al., 2001b) and early observations indicated that these Na_V channel mutations could also affect the clinical response to AEDs in genetic epilepsies (Guerrini et al., 1998). Dravet syndrome, caused by de novo mutations in the SCN1A gene, is characterised by a marked aggravation of seizures upon treatment with LTG (Guerrini et al., 1998). Other major targets for PGx study include subunits for potassium channels, calcium channels, GABA and glutamate receptors, GABA transporters, GABA transaminase and synaptic vesicle protein 2A (SV2A) (Lynch et al., 2004). Additional PD molecules of potential significance to clinical response include genes for effector components of the downstream pathways associated with AED target binding and action (Gillham et al., 1990, Ferraro et al., 2006).

1.6.6 Epilepsy or disease related candidate genes

Prognosis studies of epilepsy suggest that the underlying molecular disease pathogenesis is an important determinant of outcome or response to AED treatment (Depondt and Shorvon, 2006). Differences in AED response can be seen between types of epilepsy, seizure types and particular seizure syndromes (Semah et al., 1998). Any genes causing epilepsy are thus potential candidates for genetic variation that may also influence differences in AED response (Spear, 2001, Depondt, 2006a). In recent years at least a dozen genes have been identified in rare forms of monogenic epilepsies (Graves, 2006). Whether by design or coincidence, drugs often act upon gene products that play a role in the molecular pathology of a particular disease, and so this class of epilepsy-causing genes (Na^+, Ca^{2+} and GABA receptor subunit genes) not surprisingly overlaps with common AED targets (Ferraro and Buono, 2005, Depondt, 2006a). Animal models with mutations in these epilepsy associated genes, have accordingly demonstrated changes in sensitivity to several AEDs (Picard et al., 1999, Lucas et al., 2005). Disease susceptibility genes that do not encode actual AED targets are can also predispose to drug response (Depondt, 2006a, Depondt and Shorvon, 2006).

1.6.7 Current epilepsy pharmacogenetic research effort

Over the last two decades, a considerable effort has been made to unravel the genetic basis of variable response to AEDs (Nakajima et al., 2005, Loscher et al., 2009). Clinical efficacy for AEDs involves preventing seizure occurrence through identifying optimum drug

dosing regimens during AED administration, whilst also avoiding issues of tolerability. Substantial evidence from studies of the PK pathways (metabolism and transport) of AEDs also indicate that genetic variation may additionally affect clinically effective drug dose (Klotz, 2007).

The majority of PGx studies have until recently aimed at identifying PK variation in the multidrug resistance phenotype of epilepsy (Loscher and Delanty, 2009, Johnson et al., 2011b) and this has largely focused on drug transporter candidate genes (Depondt and Shorvon, 2006). The biological basis of 'refractoriness' is however thought to most likely be multifactorial and variable (Tate and Sisodiya, 2007). As AEDs are required to traverse the BBB and bind to one or more target molecules to exert their particular therapeutic effect, two PGx theories have emerged in an attempt to explain treatment failure in epilepsy (Remy et al., 2003, Remy and Beck, 2006). Firstly the drug transporter PK hypothesis, which is almost entirely focused around the Pgp efflux protein that was overexpressed in epileptic brain tissue from patients with drug resistant epilepsy and secondly the more recent drug target PD hypothesis that proposes ion channel and/or neurotransmitter dysfunction in AED resistance (Sills et al., 2002, Sills, 2004, Tate and Sisodiya, 2007). Experimental evidence from animal studies of drug resistance originally associated altered Na_V channel pharmacological sensitivity and electrophysiological properties between responsive and pharmacoresistant models of refractory epilepsy (Remy et al., 2003, Loscher, 2005c, Remy and Beck, 2006).

The genes studied in regard to this phenotype of response include those encoding the efflux proteins; *ABCB1*, *ABCC2*, *ABCG2* and BCRP, encoding MDR, MRP and BCRP respectively and the RLIP transporter gene; *RLIP76* (these form the prominent transporter hypothesis for drug resistance). The drug target proteins studied include the Na_V channel encoding genes *SCN1A*, *2A*, *3A* and a number of $GABA_A$ receptor genes (drug target hypothesis)(Tate and Sisodiya, 2007). Recent research has additionally implicated the astrocytic GABA transporter GAT-3 (encoded by *GAT3*) with drug responsiveness (Meldrum and Rogawski, 2007, Kim et al., 2011a). *GAT3* variation was associated with the pharmacoresistance phenotype in a recent candidate gene association study (Kim et al., 2011a) and the GAT-3 protein has also been proposed as a potential drug target for new AED development (Meldrum and Rogawski, 2007, Kim et al., 2011a).

The drug transporter hypothesis of multidrug resistance (Loscher and Delanty, 2009) proposed that increased brain expression of efflux transporters could either be a result of prolonged or frequent seizures, as demonstrated in rodent models of epilepsy (Loscher and Potschka, 2005b, Loscher and Brandt, 2009), and/or be due to a genetic contribution, such as polymorphisms in the encoding genes (Loscher and Potschka, 2005a, Loscher and Brandt, 2009). Numerous PGx studies have implicated *ABCB1* polymorphisms in multidrug resistant epilepsy, with several indicating that *ABCB1* polymorphisms that affect the expression or

functionality of Pgp such as the well-known 3435C>T polymorphism (Siddiqui et al., 2003), are more frequent in AED non-responders than responders (Loscher et al., 2009, Schmidt and Loscher, 2009). This finding could not be reproduced in many other studies for AED responsiveness (Loscher et al., 2009) including a meta-analysis effort using over 3000 refractory epilepsy patients and controls across multiple populations (Bournissen et al., 2009, Haerian et al., 2011). There is also a matter of debate on which AEDs are transported by the human Pgp transporter (Luna-Tortos et al., 2008, Loscher et al., 2011). The association between *ABCB1* 3435C>T and pharmacoresistance in epilepsy is thus unclear with only three of the published studies showing positive associations (Tate and Sisodiya, 2007, Robey et al., 2008, Loscher et al., 2009).

A number of associations have additionally been reported for DME genes, the most prominent being; PHT dose-related toxicity and *CYP2C9* polymorphisms. This was one of the first positive associations of genetic variation altering the metabolism of AEDs (Mamiya et al., 2000). Several *in vitro* studies have illustrated that *CYP2C9**2 and *3 genotypes have a decreased capacity for PHT metabolism (Saruwatari et al., 2010, Depondt et al., 2011). Numerous different reports of PHT toxicity have demonstrated low activity with the homozygote *CYP2C9*3* genotype and heterozygote genotype for both *CYP2C9* and *CYP2C19* enzymes (Kasperaviciute and Sisodiya, 2009). CYP2C9/19 genotyping may in theory influence AED dosing (Gardiner and Begg, 2006), however is not routinely used as a clinical guide, in part because it only explains a limited proportion of dosing variation (Anderson, 2008, Kasperaviciute and Sisodiya, 2009, Loscher et al., 2009, Depondt et al., 2011). Genotype associations were also demonstrated for *EPHX1* and more recently *UGT1A4* and *UGT2B7*, however these were again of limited proven strength for clinical application (Saruwatari et al., 2010).

The PD hypothesis for drug-resistance in epilepsy (Remy et al., 2003, Remy and Beck, 2006) proposes altered pharmacological sensitivity of the Na_V channel (Remy, Gabriel et al. 2003) leading to reduction in AED sensitivity (Remy, Gabriel et al. 2003; Remy and Beck 2006). The hypothesis suggests Na_V channel polymorphisms can alter the subunit composition or structure (Remy and Beck, 2006) through modifications in the transcription of channel subunits as a result of persistent seizures, (Remy and Beck, 2006). Most experimental studies investigating the molecular basis of altered drug target sensitivity have focused on transcriptional changes of ion channel subunits in response to seizures (Remy and Beck, 2006, Volk et al., 2006, Bethmann et al., 2008, Loup et al., 2009). Whether these structural changes are indeed influenced by polymorphisms in drug target genes has not been as widely described in literature (Nakajima et al., 2005).

Since the 2003 implication by Remy *et* al, Na_V gene variants, particularly the Na_V α1 subunit gene (*SCN1A*) have received further attention (Gillham et al., 1990, Tate et al., 2005,

Tate et al., 2006, Abe et al., 2008, Kwan et al., 2008). A functionally validated polymorphism located in *SCN1A* in a retrospective dosing study provided the first evidence that drug target (PD pathway) polymorphisms may additionally be influential in the responsiveness to AED treatment (Tate et al., 2005). This proof of concept study prompted a handful of additional investigations concerning the relevance of Na_V gene variation to both AED efficacy and AED resistance (Abe et al., 2008, Kwan et al., 2008). As of yet no definite evidence confirming a major role of *SCN1A* in AED responsiveness can be found (Loscher et al., 2009, Manna et al., 2011). Table 1.6 summarises some of the main research studies carried in epilepsy PGx to date. This includes recent data from studies on SCN1A (2013).

1.6.8 Pharmacogenomics and AEDs

Despite numerous studies conducted in epilepsy PGx, there is an absence of conclusive data to explain drug responsiveness and to inform treatment decisions (Nakajima et al., 2005, Johnston et al., 2009, Kasperaviciute and Sisodiya, 2009). As discussed in sections 1.4.9, 1.4.10, overall, advancements in understanding of human genetics and improvements in genomic technology have shed some light on the response to pharmacological treatment, to aid the clinical management of several common complex traits and disorders (Ritchie, 2012). Pharmacogenomics studies have observed a number of successes in recent years, most of which concern pharmacotherapy for cancer i.e. *EGFR* tyrosine kinase inhibitors (TKIs) in the treatment of lung cancer (Yi et al., 2009) and *HER2*-directed therapies in the treatment of *HER2*-positive early-stage breast cancer (Grant and Hakonarson, 2007, Arteaga et al., 2012, Ritchie, 2012). Additional PGx successes include the use of the analgesic codeine (Crews et al., 2012), anticoagulant therapy with warfarin (Johnson et al., 2011a) and abacavir therapy for HIV.

The first impact of pharmacogenomics in clinical epilepsy was the discovery of the HLA-B*1502 polymorphism as a strong predictor of CBZ induced Stevens–Johnson syndrome (SJS), in people of Chinese and south Asian ancestry (Ferrell and McLeod, 2008). Testing for HLA-B*1502 in at-risk ethnic populations is now recommended by regulators globally, including in the United States, United Kingdom, and Canada (Ferrell and McLeod, 2008, Kasperaviciute and Sisodiya, 2009, Johnson et al., 2011b). Moreover since this clinically proven association in Chinese and south-Asian patients, the HLA-A*3101 variant (Alfirevic et al., 2006) was identified in European patients and demonstrated to significantly associate with CBZ hypersensitivity (McCormack et al., 2011, Yip et al., 2012). The HLA-A*3101 variant has thus also been proposed as a clinically relevant marker to predict hypersensitivity reactions (McCormack et al., 2011, Yip et al., 2012). Screening for the *HLA-B*1502* allele in patients of Asian descent in order to prevent CBZ and PHT-induced SJS

(Ferrell and McLeod, 2008) has now been incorporated into standard medical practice (Chen et al., 2011). Though impressive, this remains the only epilepsy PGx finding that has resulted in clinical application so far (Kasperaviciute and Sisodiya, 2009).

To summarise, advances have been made in identifying genetic markers of AEs in terms of severe cutaneous reactions but there has been little progress in predicting AED efficacy (Kasperaviciute and Sisodiya, 2009, Johnson et al., 2011b). Progress for epilepsy PGx is thus lagging behind when compared to many polygenic neurological conditions and PGx data generated to date has had little impact on current AED treatment guidelines (Ferraro and Buono, 2005, Loscher et al., 2009). No definite predictors of drug efficacy are known and current treatment for unresponsive individuals remains largely based on trial and error of existing medications (Kasperaviciute and Sisodiya, 2009).

1.6.9 Epilepsy pharmacogenetic studies: research limitations and design issues

Several possible reasons have been proposed for the limited success in discovering susceptibility loci for AED response as well as the numerous failed attempts to replicate the few potential risk alleles identified (Nakajima et al., 2005, Baksh and Kelly, 2007). Among the reasons for the lack of success in general, are the overall lack of clarity in study findings due to the retrospective design and analysis and small cohort size and/or short duration of follow-up (Johnson et al., 2011b). Additional problems include general methodological limitations associated with complex outcomes (Cardon and Bell, 2001, Hirschhorn et al., 2002, Colhoun et al., 2003, Depondt, 2006b, McCarthy et al., 2008). Of particular concern with epilepsy PGx studies are i) the diversity in the definition of AED resistance, ii) the heterogeneity of the study populations and ii) the lack of a multigenic approach associated with candidate gene based research (Ferraro et al., 2006, Kasperaviciute and Sisodiya, 2009, Johnson et al., 2011b). The latter of these has produced disappointing results in many genetic studies (Colhoun et al., 2003, Goldstein et al., 2003, Grant and Hakonarson, 2007, di Iulio and Rotger, 2012).

In terms of the issue of heterogeneity of study cohorts, this is often due to i) differences in clinical phenotype definitions used for selecting participants between study sites (i.e. definitions used to classify epilepsy syndromes and epilepsy severity: chronic long-term epilepsy versus newly treated epilepsy patient populations, with the former predominantly used in the majority of PGx reports), and ii) differences in the clinical treatment regimens used between studies (i.e. dosing strategies and drug selection decisions used by neurologists), (Johnson et al., 2011b).

Another likely reason for the lack of progress in identifying genetic contributions to drug efficacy in epilepsy is the small effect size of variants detected to date (Cavalleri et al.,

2011). Even when a study has been successfully replicated, the effect of any given polymorphism on drug response is often lower than initially described (Ioannidis, 2003). Such studies have commonly tested for associations between single candidate genes or single SNPs, and therefore only explain a fraction of the variability in drug response, with accordingly limited plausibility and clinical utility (Goldstein, 2009, Cavalleri et al., 2011). This approach and its inherent limitations explains why only a handful of PGx markers are actually useful for individualising treatment in clinical practice (Ikediobi et al., 2009). The fact that the classic candidate gene approach does not take into account the full complexity underlying drug response is another possible explanation for the lack of positive findings to date (Ritchie and Motsinger, 2005). Drug response is now widely considered to be the joint effect of multiple polymorphisms, gene-gene interactions (epistasis)(Hardy and Singleton, 2009), and the interplay with environmental factors (gene-environment interactions) (Hirschhorn et al., 2002).

SNPs with small effect sizes in combination are more likely to underpin the multifactorial basis of drug efficacy and tolerability (CDCV hypothesis) (Gillham et al., 1990, Evans and McLeod, 2003, Iyengar and Elston, 2007) and a genomic approach for the identification of these genetic variants is more appropriate (Depondt and Shorvon, 2006, Baksh and Kelly, 2007). Study design and data analysis, employing wide-scale mapping of biologically relevant loci in much larger cohorts (through large collaborations and consortia meta-analysis) (Cavalleri et al., 2011), and/or employing GWAS may help nullify some of the previous failures of candidate gene studies in AED response and enable replication studies to validate any previously reported true effects (Baksh and Kelly, 2007). Wide-scale mapping of large sections of the genome in PGx (www.genome.gov/GWAStudies) (Nagasawa and Nakahara, 1992, Hardy and Singleton, 2009, Daly, 2010a, Wang, 2010, Wang et al., 2011), is growing, and genome-wide efforts for epilepsy PGx, are likewise expanding. The first examples of the application of GWAS have however only just emerged and both of these concern cutaneous drug reactions and were prompted by the discovery of the clinically important *HLA-B*1502* variant (McCormack et al., 2011, Ozeki et al., 2011b).

Despite limited clinical significance and difficulties in replication across epilepsy PGx findings, many of the SNPs and gene associations found to date appear relevant and warrant further consideration. For future PGx studies in epilepsy to have sufficient power to detect genetic variants with small effect sizes, much larger sample sizes are required (Crowley et al., 2009, Johnson et al., 2011b). Recent developments in genetic technology do however hold great promise for the field of epilepsy. With an increasing number of robust associations found in different diseases to date and the rise in GWAS being applied to neurological conditions, the status of pharmacogenomics/PGx for epilepsy is likely to change rapidly (Kasperaviciute and Sisodiya, 2009, Mullen et al., 2009, Rees, 2010).

1.6.10 Machine learning

To deal with these issues many researchers have begun to explore more powerful statistical methodologies capable of dealing with both the problem of detecting small, multiple associations and the analysis of high-dimensional data (Moore et al., 2004, Hoppe, 2005, Rodin et al., 2011) and this includes the machine learning (ML) data mining method (Hastie et al., 2001, Koster et al., 2009). The ML approach to data analysis of pharmacogenomics data typically involves three-steps, i) selection of variables (SNPs), ranked in order of effect on drug response phenotype, ii) modelling step involving generation of a predictive model using SNPs and any other relevant factors, iii) evaluation of generated models using conventional statistical analysis methods (Koster et al., 2009). Typical ML approaches applied to genomic studies model data using Bayesian networks, which allow the inferential exploration of previously undetermined relationships among genetic and clinical variables, and describe these relationships, once identified, using a hypothesis or model-free approach (Hoppe, 2005, Zhang and Rajapakse, 2009, Rodin et al., 2011). Data mining methods generally involve the development of disease association models that allow integration of the interactions between multiple SNPs in addition to clinical variables and disease phenotype, and so overcome the main limitation of traditional statistical approaches through their ability to model high-dimensional data (Hoppe, 2005, Wilke et al., 2005). Additional advantages of ML algorithms include robustness of parametric assumptions, high power and accuracy, ability to model non-linear effects, and the availability of numerous well-developed algorithms (Moore and Ritchie, 2004, McKinney et al., 2006).

Table 1.6 Summary of genes and SNPs associated with antiepileptic drugs so far

Gene	Genetic polymorphisms	AED	Associated drug parameter	Reference
DRUG DOSING STUDIES				
PHASE I DRUG METABOLISING ENZYMES				
CYP2C9	*CYP2C9**1/*2	PHT	Maintenance dose	(van der Weide et al., 2001)
	*CYP2C9**2	PHT	Altered metabolism	(Odani et al., 1997)
CYP2C9/ C19	*CYP2C9/C19*	PHT	PHT clearance	(Lee et al., 2007)
		PB	PB clearance	(Goto et al., 2007)
		VPA	VPA clearance	(Wu et al., 2010)
		PHT	PHT dosage	(Hung et al., 2004)
CYP2C19	*CYP2C19*2/*3*	PHT	PHT clearance	(Seo et al., 2008b)
		ZNS	ZNS clearance	(Okada et al., 2008)
		CLB	Efficacy to CLB	(Yukawa and Mamiya, 2006)
		PHT/PB	Pharmacokinetics	
CYP3A5	*CYP3A5**3 genotype	CBZ	Concentration	(Park et al., 2009)
			Pharmacokinetics	(Seo et al., 2006)
GSTM1/ GSTT1	null genotype	VPA	Hepatotoxicity	(Fukushima et al., 2008a)
GSTM1	null genotype	CBZ		(Ueda et al., 2007)
PHASE II DRUG METABOLISING ENZYMES				
UGT2B7	*UGT2B7* -161C>T	LTG	Concentration to daily dose ratio	(Blanca Sanchez et al., 2010)
EPHX1	Try113His and His139Arg	CBZ	Maintenance dose	(Makmor-Bakry et al., 2009)
			Metabolism; increased and decreased CBZ diol:CBZ epoxide ratios	(Nakajima et al., 2005)
DRUG TARGET PROTEINS				
SCN1A	IVS5-91G>A	CBZ/PHT	Maximum dose	(Tate et al., 2005)
		PHT	Maintenance dose	(Tate et al., 2006)
		CBZ	No association with CBZ dosage	(Zimprich et al., 2008)
DRUG EFFICACY STUDIES				
DRUG TRANSPORTER PROTEINS				
ABCB1	C3435T	Multiple	Association with refractory epilepsy	(Siddiqui et al., 2003)
ABCB1	C3435T	Multiple	No association with resistance to AEDs in a meta-analysis	(Bournissen et al., 2009)
ABCB1, ABCC2,	Multiple variants	Multiple	No association with drug resistance	(Kim et al., 2009)

ABCG2	Including C3435T			
RLIP76	Multiple	Multiple	No association with AED treatment response	(Soranzo et al., 2007)
DRUG TARGET PROTEINS				
GAT-3	Multiple	Multiple	AED resistance	(Kim et al., 2011a)
SCN1A	IVS5-91G>A	CBZ	CBZ resistance	(Abe et al., 2008)
SCN1A *SCN2A*	rs229877 rs17183814	Multiple	Association with epilepsy Association with multidrug resistance	(Lakhan et al., 2009)
SCN1A	IVS5-91G>A	Multiple	No association with responsiveness	(Manna et al., 2011)
SCN1A *SCN2A* *SCN3A*	Several SNPs including; IVS5-91G>A rs2298771 rs17183814	Multiple	No association with responsiveness: Multi-centre meta-analysis	(Haerian et al., 2011)
SCN1A *SCN2A* *SCN3A*	Multiple SNPs including *SCN1A*IVS5-91G>A and *SCN2A*IVS7 -32A>G		Association with AED non-responsiveness No association with with *SCN1A/3A*	(Kwan et al., 2008)
SCN1A *GABA* *GABRA1*	Multiple including *SCN1A*IVS5-91G>A, rs2290732 rs2298771 *GABRA1* rs2290732		Association with CBZ tolerability and or efficacy and *SCN1A* SNPs and *GABRA1;* rs2290732	(Zhou et al., 2012)
Multiple genes and SNPs	Five SNPs: *SCN4B, SCN4B* *KCNQ4,* *GBBR2,* *SLC1A3*		Prediction of AED treatment response by five SNP model	(Petrovski et al., 2009)

AED= antiepileptic drug, PHT= phenytoin, PB= phenobarbital, VPA= valproate, ZNS= zonisamide, CLB= clobazam, CBZ= carbamazepine, LTG= lamotrigine, SNP= single nucleotide polymorphism

1.7 Research justification and aims of the thesis

Elucidating the basis of AED response would both aid the understanding of pathogenic mechanisms underlying drug resistance in epilepsy and additionally allow the development of innovative rational treatments for people with refractory epilepsy (Sisodiya, 2005). The optimum drug therapy in epilepsy however continues to trail behind that of many other common, complex disorders (Kasperaviciute and Sisodiya, 2009).

The majority of candidate association studies are characterised by irreproducibility often attributed to a lack of statistical power and are most likely due to weak genetic effects and/or population specific gene-gene and/or gene-environment interactions (Hirschhorn et al., 2002). The confirmed signals emerging from GWA scans and subsequent replication efforts similarly remain only signals (McCarthy, Abecasis et al. 2008). A substantial body of experimental evidence now supports a multifactorial, polygenic basis for common traits (Ferraro et al., 2012). Research indicates the significance of capturing the interactions between genetic factors and other variables including phenotypes, environment and drugs (Baksh and Kelly, 2007, Kim et al., 2011a, Rodin et al., 2011). A number of studies have already applied powerful statistical methods that use data mining approaches such as ML classification methods and ML based methods for detecting epistatic and additional genetic interactions to PGx and pharmacogenomics data (Ritchie and Motsinger, 2005, Wilke et al., 2005, Rodin et al., 2011) and studies are now emerging for the characterisation of the genetic variables underlying refractory epilepsy (Petrovski et al., 2009, Johnson et al., 2011b). The promising results indicated by initial studies for refractory epilepsy moreover advocate further consideration of ML approaches and interaction data analysis methodologies for investigating AED efficacy (Cavalleri et al., 2011, Johnson et al., 2011b).

1.7.1 Research goals

The intention of this PhD thesis was to explore genetic contribution to drug response phenotypes in epilepsy, with the purpose of identifying and/or validating genetic markers influencing drug efficacy and optimal dosing. To achieve these research goals several lines of previous PGx and pharmacogenomic evidence for the genetic contribution to individual responsiveness in epilepsy treatment were followed. ML methodologies have recently been applied to PGx data from patients with epilepsy and such approaches were additionally explored and assessed for utility using epilepsy phenotype data from UK patients.

Data currently available for epilepsy PGx is limited by the use of heterogeneous populations consisting of different epilepsy phenotypes, varying definitions of drug responsiveness, retrospective data and mainly individuals with long standing epilepsy (often exposed to multiple drugs and thus a greater chance of the existence of uncontrollable genetic

and environmental influences from continuous drug exposure and or seizure related neurological damage). The majority of AEDs differ in their pathways of drug action and distribution and also dosage and titration. Including patients with multiple AED treatment can also confound the effective detection of a genetic influence on response to a single drug. The studies presented in this thesis thus also aimed to provide a set of genetic investigations using a more homogenous epilepsy population.

1.7.2 Specific aims and thesis outline

The specific research aims are listed below and each one is tackled in the individual research chapters that follow.

Aim 1: To characterise genetic variation across DMEs responsible for the metabolism of CBZ to identify markers for optimal AED dosing in newly treated epilepsy (Chapter 3)

Aim 2: To establish the contribution of a single functional SNP in the SCN1A gene to optimal dosing of AEDs in individuals with newly treated epilepsy (Chapter 4)

Aim 3: To assess the validity and predictive value of a ML-based multi-genetic model for classifying treatment outcome with AEDs, using an independent cohort of patients with newly treated epilepsy (Chapter 5)

Aim 4: To explore the utility of ML approaches for the identification of influential markers for classifying primary generalised epilepsy (Chapter 6)

Aim 5: To validate the findings of a recent GWAS (Speed et al., 2013) that reported significant genetic influences on the likelihood of achieving 12 months seizure freedom, using an independent cohort of newly treated epilepsy (Chapter 7)

The following chapter (Chapter two) presents experimental methods common to two or more research chapters. Methodologies specific to each research aim are described in the respective results chapter.

CHAPTER TWO

RECURRENT METHODS

CONTENTS

2.1 Materials

2.1.1 Consumables

100% ethanol was used for sterilisation and cleaning of the nanodispenser for both PCR and extension reactions. PCR reagents included Hot Star Taq® enzyme, deoxyribonucleotide triphosphate (dNTP) mix, magnesium chloride ($MgCl_2$) and polymerase chain reaction (PCR) buffer (PCR reagent set purchased directly from Sequenom (Hamburg, Germany). The following consumables for MassARRAY genotyping (extension reaction) were also purchased from Sequenom: the iPLEX® Gold Reagent Kit (iPLEX® Gold SNP genotyping assay for single base primer extend) containing; shrimp alkaline phosphatase (SAP) buffer, SAP enzyme, iPLEX Termination mix, iPLEX buffer, iPLEX enzyme, Clean Resin kit (for removing PCR impurities) consisting of resin (28g) and a 384-well dimple plate for resin application, and 10 x 384 SpectroCHIP® Arrays (for allele detection).

All primers were supplied by Metabion (Martinsried, Germany). 96-channel tips with a volume of 30µL were required for the liquid handler Matrix as were Matrix reagent reservoirs; these were also purchased from Qiagen. The Quant-iT PicoGreen® dsDNA reagent kit v 1.0 was purchased from Invitrogen Ltd, (Paisley, UK). Ultrapure agrose powder and 20x Tris-HCl-EDTA was obtained from Invitrogen Ltd (Paisley, UK). Ethidium bromide and 0.5x Tris-EDTA (TE) buffer were purchased from Sigma-Aldrich Ltd (Gillingham, UK). The 100 base pair (bp) DNA molecular weight marker XIV used for all electrophoresis was purchased from Roche Applied Science (Burgess Hill, UK). 384-well micro-plate adhesive polymerase chain reaction (PCR) films and films for general plate sealing were purchased from ABgene (Loughborough, UK). 96-well polystyrene plates were purchased from Sarstedt (Leicester, UK). Costar 96-well solid, flat bottom plates for Picogreen® DNA quantification were purchased from VWR International Ltd (Lutterworth, UK). All other generic reagents and consumables were available as standard and were obtained from University stores; and these were purchased from standard supply companies such as Sigma.

2.1.2 Equipment

Pipettes: Small volumes were dispensed using single channel pipettes (with 2.5µl, 20µl, 200µl and 1000µl volumes) and multichannel pipettes (with 10µl and 50µl volumes) from Eppendorf (Cambridge, UK). Ultra-purified laboratory water: Molecular biology grade purified water for all experimental procedures was obtained using the ELGA PURELAB water system (minimum 18.2 MΩ/cm resistivity) (ELGA, Marlow, UK).

DNA quantification: Genomic DNA samples were quantified either using the NanoDrop 1000 Spectrophotometer (Thermo Scientific Inc., Hemel Hempstead, UK) and/or the Beckman Coulter DTX880 multimode detector (Beckman Coulter Ltd., High Wycombe,

UK). DNA amplification and primer extension: For all MassARRAY genotyping projects G-Storm Thermal Cycler GS-4 Kappa from G-storm (Ringmer, UK) was used for both genomic amplification and primer extend cycles.

Genotyping: Equipment required for MassARRAY® genotyping was mainly specialised Sequenom® technology and purchased directly from Sequenom®. Dispensing post-PCR samples was carried out using a Sequenom® Matrix Liquid handler (a 96-channel pipetting robot that provides pre-programmed optimised pipetting schemas for all MassARRAY applications), and a Sequenom® MassARRAY nanodispenser was used to transfer iPLEX® Gold reaction products on to a Sequenom® SpectroCHIP. A Sequenom® MassARRAY READER real-time (RT) Matrix-assisted laser desorption/ionisation-time of flight mass spectrometry (MALDI-TOF) instrument (specifically designed for genomic applications) was used to read SpectroCHIPs containing experimental samples and Sequenom® SpectroAcquire computer software was used to visualise all genotype data.

2.2 Patient cohorts

Patients used in the various studies were principally from two distinct UK cohorts; the SANAD cohort and the Glasgow cohort. An Australian cohort was also used in some analyses, comprising patients from both the Department of Medicine, University of Melbourne Hospital, Melbourne and the Department of Medicine and Epilepsy Research Centre, Austin Health, Heidelberg, Victoria. All patients were identified as having a diagnosis of epilepsy (as defined by the ILAE) and were treated with AEDs for seizure control. Patients of non-European ancestry were excluded. Clinical information for each cohort was contained in electronic databases generated from clinical trial data or hospital notes.

2.2.1 Glasgow cohort

The Glasgow cohort consisted of 893 patients attending the epilepsy outpatient clinic at the Western Infirmary in Glasgow. Individuals had newly treated epilepsy (n=462) or long term/chronic epilepsy (n=427), and had been treated with a wide range of AEDs, as monotherapy or polytherapy. DNA was extracted from venous blood samples using a standard phenol-chlorophorm method (Szoeke et al., 2009) and all individuals provided informed consent for the collection and pharmacogenetic analysis of DNA (approved by the West Ethics Committee; North Glasgow University Hospitals NHS Trust in September 2002 (ref: 02/119(2)). All samples were aliquots of original DNA stored at the Western Infirmary in Glasgow. Aliquots were stored in 1mL cryovials at the Wolfson Centre for Personalised Medicine, Liverpool.

2.2.2 Glasgow cohort clinical data

Although a cross-sectional outpatient clinic cohort, a large proportion of these individuals had participated in randomised monotherapy trials (Stephen et al., 2007). Drug response phenotypes in the Glasgow cohort were identified by retrospectively reviewing the prospectively collected clinical data from a database generated from trial and/or hospital notes. Patient phenotype data that was available in the clinical database included general demographic details (i.e. date of birth (DOB), gender) and the following clinical information and phenotype data; previous drug treatment, epilepsy type, date of first ever seizure, pre-treatment seizures, EEG and imaging results, and AED treatment history including initial AED and subsequent AEDs until last follow up, with dates of withdrawal and dosage for each AED. Of these patients, individuals were not considered for genetic analysis in this thesis if they had long-standing epilepsy (i.e. not newly treated with AEDs), were not monotherapy patients, and their ethnic origin was non-European.

2.2.3 SANAD cohort

The SANAD cohort was drawn from patients who had participated in the Standard and New Antiepileptic Drug study, an un-blinded, multicentre, randomised trial comparing the efficacy, tolerability and cost-effectiveness of established and newer AEDs in patients with newly-diagnosed epilepsy from epilepsy centres across the UK (Marson et al., 2007a, 2007b). More than 2,400 patients were recruited in the trial and followed-up prospectively for a minimum period of two years from initiation of the first ever AED.

985 SANAD participants gave informed consent to the collection and analysis of DNA, approved by the North-West Multicentre Research Ethics Committee in August 2002 (ref: MREC 02/8/45). DNA was extracted from blood and or saliva samples using a standard phenol-chloroform extraction method and purity and concentration confirmed by spectrophotometry. SANAD DNA samples were stored at the Welcome Trust Sanger Centre, Cambridge and subsamples of these were stored and available for experimental use in Liverpool. All 985 samples were considered for genetic analysis except those of non-European ancestry and those without epilepsy.

2.2.4 SANAD cohort clinical data

Neurological history and seizure history was recorded at recruitment. Seizures and epilepsy syndromes were classified by ILAE classifications. Patients were seen for follow up at 3 months, 6 months, 1 year, and at successive yearly intervals from the date of randomisation and details of drug treatment and effectiveness were recorded. Where patients ceased attending hospital clinics, follow-up information was obtained from general practitioners, or directly

from the patient via a telephone interview.

Due to the nature of the SANAD trial, a large amount of clinical and phenotype data was collected. The following clinical data was available for patient selection and subsequent data analysis:

i) Neurological disease history i.e. the presence of a learning disability, neurological deficit, neurological disorder, head Injury, meningitis/encephalitis, intracranial surgery, acute symptomatic seizures or family history of epilepsy
ii) Epilepsy type i.e. LRE, IGE, UNC,
iii) Seizure type and syndromic diagnosis (where available)
iv) EEG and CT/MRI results
v) Treatment history (untreated or monotherapy)
vi) Seizure history i.e. recent seizure occurrence
vii) AED history; initial or randomised drug, dosage history and withdrawal details if applicable (including dates of treatment and each study visit).

Data collected for the SANAD trial included dates for each follow up visit and drug treatment history during follow up period. Due to the outcomes of interest of the trial, (time to first seizure, time to 12-month remission or treatment failure, time to withdrawal due to inadequate seizure control, and time to withdrawal for unacceptable AEs or ADRs), seizure history i.e. dates of occurrence, number, type was also recorded as were reasons for drug withdrawal, and 12 month remission status.

2.2.5 Additional Australian cohort

Clinical data was provided from a population of Australian individuals for the purposes of research studies 6 and 7 of this thesis. The Australian cohort consisted of patients prospectively recruited from clinics in Victoria, Australia; Royal Melbourne Hospital and the Austin Hospital on the basis of being newly treated with AEDs (Petrovski et al., 2009), as part of a multicentre collaboration study of epilepsy aeiteolgy and seizure types (the Epilepsy Genetics Consortium; EPIGEN) (Cavalleri et al., 2007). All patients were of self-identified European Australian ethnicity and were recruited after obtaining written informed consent. Individuals were all diagnosed with epilepsy (ILAE) and were followed up prospectively. Patient treatment response was phenotyped once individuals reached their 1-year follow-up with their initial AED treatment. In total clinical and genetic data was provided for n=427 patients on the basis of having primary generalised epilepsy (PGE) or LRE. All DNA genotyping was done using the Illumina GoldenGate platform at Duke University, Durham,

NC, USA (Cavalleri et al., 2007).

2.2.6 Australian cohort clinical data

Patient phenotype data that was provided included age of onset, initial drug treatment, epilepsy syndrome type and seizure type. Patients with PGE were classified with the following syndromes, subsyndromes and seizure types; juvenile myoclonic epilepsy (JME), juvenile absence epilepsy (JAE), childhood absence epilepsy (CAE), CAE generalising to JME (CAE -> JME), CAE -> JAE, CAE/JAE, idiopathic generalised epilepsy (IGE) excluding JME (non-JME IGE), mesial temporal lobe epilepsy associated with hippocampal sclerosis and all other focal neocortical epilepsies. The seizure classifications used were; GTCS; occurring only in the context of a syndromic diagnosis of an IGE, myoclonic seizures, absence seizures, secondarily GTCS and partial seizures (either simple or complex). Patients with any epilepsy type who also had a history of febrile convulsions (FS) were also included (Cavalleri et al., 2007).

2.3 DNA preparation and storage

All available DNA samples (985 SANAD and 893 Glasgow) were quantified and aliquots of between 50-200µL were prepared for later experimental use.

2.3.1 DNA quantification using spectrophotometry

All Glasgow samples were quantified using the NanoDrop 1000 spectrophotometer. The NanoDrop 1000 Spectrophotometer accurately measures double-stranded DNA (dsDNA) up to 3700 ng/µL without dilution. To do this, the instrument automatically detects the high concentration and utilises the 0.2 mm path length to calculate the absorbance. The machine was blanked using a 1.5µL volume of 0.1 x TE buffer and the recommended volume of 1.5µL of each stock DNA sample was placed on the spectrophotometer for a measurement of DNA concentration. The Nanodrop spectrophotometer additionally provides two DNA purity readings based on sample absorbance. The 260/280 ratio of a sample is the ratio of absorbance at 260 and 280 nm and is used to assess the purity of DNA. A ratio of ~1.8 is generally accepted as “pure” for DNA, and if the ratio is substantially lower, it may indicate the presence of protein, phenol or other contaminants that also absorb strongly at or near 280nm. The 260/230 ratio of a sample is the ratio of absorbance at 260 and 230nm. This is a secondary measure of nucleic acid purity with values for “pure” nucleic acid often being higher than 260/280 values. These ratios are commonly in the range of 1.8-2.2 and if appreciably lower, this may indicate the presence of co-purified contaminants.

Table 2.1 Quantification method using Nanodrop Spectrophotometer.

Table taken manufacturer's instructions (http://nanodrop.com) (© 2008 Thermo Fisher Scientific)

Detection limit (ng/µL)	**Approximate upper limit (ng/µL)**	**Typical reproducibility (minimum 5 replicates) (SD=ng/µL; CV= %)**
2	3700 ng/µL (dsDNA)	sample range 2-100 ng/µL: ± 2 ng/µL sample range >100 ng/µL: ± 2%

2.3.2 DNA quantification using Picogreen

All SANAD samples were quantified using Picogreen®. PicoGreen® is an ultrasensitive fluorescent nucleic acid stain for quantifying double-stranded DNA (dsDNA) in solution. Free dye does not fluoresce, but upon binding to dsDNA it exhibits a >1000-fold fluorescence enhancement. This allows the quantification of as little as 25 pg/mL of dsDNA. The PicoGreen® DNA quantification method was better suited to the SANAD samples, which were of low concentration in comparison to the Glasgow samples and additionally assumed to be of less purity. The major disadvantages of using the 260 nm absorbance method is the large relative contribution of nucleotides and single-stranded nucleic acids to the absorbance signal, the interference caused by contaminants commonly found in nucleic acid preparations, the inability to distinguish between DNA and RNA, and the relative insensitivity of the assay.

2.3.3 Picogreen methodology

The Quant-iT^{TM} Picogreen® manufacturer's instructions were followed for quantifying all available SANAD DNA samples. For this a standard curve was first generated using diluted Picogreen® and standardised DNA (lambda stock DNA at 100 ug/mL provided in the kit), from which the unknown concentration of all SANAD DNA samples was calculated. The Picogreen® reagent stock provided was first diluted to a working solution using 1 x TE (100uL Picogreen® reagent added to 19.9 mL TE solution). A standard curve was then generated using lambda standard DNA diluted 50 fold (1.47µL of 1 x TE solution added to 30µL of DNA); see Table 2.2 for dilutions used to generate a high-range standard curve.

Table 2.2 Dilution factors used for Picogreen® quantification.

Volume of Standard DNA (µL)	Volume of 1xTE (µL)	Volume of picogreen working solution (µL)	Final concentration
100	0	100	1 ug/mL
10	90	100	100 ng/mL
1	99	100	10 ng/mL
0.1	99.9	100	1 ng/mL
0	100	100	Blank

These were added to the first row of a 96-well plate. The remaining wells were filled with 99µL of diluted Picogreen® solution and 1µL of DNA of unknown concentration. After mixing and incubation (as detailed in the protocol), the fluorescence of each well was measured using the Beckman Coulter DTX880 multimode detector and the concentration of each DNA sample calculated from the standard curve.

2.3.4 DNA sample dilution and storage

All DNA stock (Glasgow and SANAD) was stored at 80 °C. All working stock solutions were diluted to a concentration of 20 ng/µL for experimental use and stored at -20°C.

2.4 Single nucleotide polymorphism selection

2.4.1 Resources for selection of genes and single nucleotide polymorphisms

Several freely accessible online genomic databases are available as a resource for the investigation of common genetic variation in human genes including the HapMap website (www.HapMap.org), database SNP (dbSNP) function of the National Centre for Biotechnology Information (NCBI) website (www.ncbi.nlm.nih.gov/projects/SNP/), the UCSC Genome Browser Human Genome Browser Gateway website (genome.ucsc.edu/cgi-bin/hgGateway) and the Ensemble Human Genome Browser (www.ensembl.org/Homo_sapiens/Info/Index).

These resources were used to locate information concerning human genes and or single SNPs to be investigated and allowed visualisation of gene regions and genetic variation within loci. Information that was extracted was mainly dependent on the requirements of the research

project, but included i) all known polymorphic sites for a particular gene, ii) their population minor allele frequency (MAF), iii) chromosome location, and iv) type or location in the corresponding gene (synonymous, non-synonymous, 5'UTR, 3'UTR). Information on those in LD and on clinical relevance, i.e. previous association with a particular condition, or previously demonstrated to associate with a therapeutic drug was also recorded (ww.pharmgkb.org; (www.ncbi.nlm.nih.gov/pubmed)).

Any potential transcriptional or regulation changes due to a SNP (i.e. location in a transcription factor binding site (TFBS), DNA methylation region, histone and polymerase binding region) was also investigated through functions available on the UCSC Genome Browser and the Ensemble Human Genome Browser.

2.5 Genotyping using Sequenom MassARRAY

2.5.1 Sequenom MassARRAY platform and reaction overview

The main method used for high-throughput SNP genotyping was the Sequenom® MassARRAY matrix-assisted laser desorption/ionisation-time of flight mass spectrometer (MALTI-TOF) platform. Although several high-throughput SNP genotyping technologies are available, Sequenom Mass ARRAY provides affordable and accurate custom genotyping assays, with a modest multiplexing methodology (Gabriel et al., 2009). The MassARRAY® MALTI-TOF platform uses a single base homogenous reaction format that can throughput >100,000 genotypes per day. This utilises multi-plex PCR reactions, a single termination mix, provides universal reaction conditions for all SNPs, requires small reagent volumes, and generates allele-specific products with distinct masses for subsequent mass spectroscopy detection. The iPLEX® reaction allows the design of assays at a multiplexing level of 36-plex.

There are several key tasks involved in genotyping using the Sequenom® MassARRAY system:

- Primer and multiplex design
- DNA amplification
- Preparation of iPLEX® Gold reaction products
- Transfer of processed iPLEX® Gold reaction products to SpectroChip® arrays
- Assay design, plate design, and setup using Sequenom® design software
- Use of Sequenom® mass spectrometer for the acquisition of reaction spectra
- Use of TyperAnalyzer software for the analysis of spectral data

The Sequenom® SNP assay is based on a locus-specific PCR reaction followed by a locus-specific primer extend reaction (Tang et al., 1999). During the primer extension or

iPLEX® reaction, an oligonucleotide primer anneals immediately upstream of the polymorphic site being genotyped. The primers and amplified target DNA are then incubated with mass-modified dideoxynucleotide terminators. The primer is extended dependent upon the template sequence and results in allele specific differences in mass between extension products, the mass of which is determined by the use of MALDI-TOF mass spectrometry. The molecular mass of extension products is used to indicate which alleles are present at the polymorphic site of interest. The primer mass for each reaction is translated into a genotype using Sequenom® software (SpectroTYPER) (Gabriel et al., 2009). Genotyping using Sequenom® MassARRAY genotyping was performed as stated in the manufacturer's instructions, an overview of which is provided below.

2.5.2 MassARRAY required components and consumables

The MassARRAY reaction can be performed in both 96- and 384-well plate format and with automated liquid handling process (for 384-well format). All experimental work performed required single wells (per sample and SNP assay) ranging from 150 up to 2000 reactions, thus a 384-well format using the automated liquid handler matrix was undertaken for all genotyping studies. All reagents and equipment for MassARRAY are listed in sections 2.1.1 and 2.1.2, respectively.

2.5.3 Primer and assay design

The online Sequenom MassARRAY Designer software was used to electronically design PCR and extension primers for a SNP of interest (https://mysequenom.com/Tools). The software also provided a plex design function to manually balance multiplex levels of SNP groups to minimise the number of reactions (Gabriel et al., 2009). The primer design process involved the automatic checking and avoidance of primer combinations and non-template extension products that could result in non-specific extension and has a proven design efficiency of >95% for all confirmed SNPs (Gabriel et al., 2009).

The reference sequence (rs) number for all SNPs of interest was typed into the rs sequence retriever function, which then retrieved genomic sequences for each specified SNP from the NCBI dbSNP database in FASTA format. The input FASTA file typically includes 500 base pairs of specific genomic sequences, 250 upstream and downstream from the SNP of interest. The SNP sequences were then passed through four additional functions involved in the SNP design and checking process, with SNPs resulting in non-specific extension excluded at each stage of the process (Gabriel et al., 2009):

i) The ProxSNP sequence-mapping step compared input sequences against the NCBI database and looked for registered SNPs that were proximal to the SNP of interest. The genome assembly version selected for all MassARRAY genotyping experiments was always Genome build 36 and the formatted sequences selected were SNP/MNP sequences.

ii) The PreXTEND SNP validation / amplicon design step then aligned the input sequences from the ProxSNP mapping step against the genome build to determine the best location for PCR primers that would result in a unique amplification product containing the target for the extension primer. Genome build 36 was again employed and the designable sequences chosen were uniquely mapped.

iii) The Multiplexed iPLEX Assay Design step then designed multiplexed genotyping assays and the multiplexed design was used for ordering PCR & extension primers as well as importing into the SEQUENOM assay editor of the Typer software. For this function, the stop mix selected was iPLEX. The multiplex level allowed a maximum of 36 SNPs and a minimum of one SNP per plex. The maximum plex level used was never more than 25 and the minimum plex level used was never less than five. Although using a plex level of 36 would maximise productivity, a lower plex level was chosen as this was the most reliable in routine laboratory genotyping.

iv) The PleXTEND Multiplexed Assay Validation step was the final step that validated all designed primers in the entire multiplex by comparing sequences using the Basic Local Alignment Search Tool (BLAST; a set of algorithms designed to perform similarity searches on all available biological sequence data) to check for potential cross-hybridisation.

2.5.4 Primer pooling

All primers for PCR and iPLEX reactions were ordered unmodified and unmixed, with standard purification. All primer plexes thus required pooling before use. Prior to any pipetting, primers were centrifuged (1200 rpm for 3 minutes) and pipetted up and down to ensure sufficient mixing. All pipetting during primer pooling was performed using extended sterilised tips. Because the Sequenom MassARRAY platform requires a multistep process, checks were put in place for each 384-well plate used. A 384-well assay plate was designed to contain approximately 10% negative control wells and duplicate DNA samples. These were also placed in unique positions on the plate to be checked for expected results at the end of the experiment (i.e., no extension in the negative controls and genotype concordance among duplicates).

2.5.5 PCR

The PCR assay pool of plexes included the multiplexed forward and reverse PCR primers for each reaction in one multiplexed assay pool. PCR primers for each plex were pooled and diluted to a working concentration of 0.5 μM for each primer in eppendorf tubes. For one 384-well reaction plate and a 24-plex reaction, 5uL of each forward primer was mixed with 5uL of each reverse primer and 260uL of Nanopure water to give a total volume of 500uL, with adjustment in the volume of water for lower or higher plex levels. All diluted working stocks and concentrated stocks of primers were stored at −20°C.

2.5.6 DNA preparation and experimental design for PCR

For Sequenom genotyping assays, 10-20 ng/μL of genomic DNA per reaction is recommended. All working stock was previously prepared at 20ng/μL and this concentration was used for all genotyping reactions. A 1uL volume of DNA was transferred from the working DNA stock into 384-well PCR reaction plates using a four-channel 10uL pipette. Plated DNA was then evaporated, sealed, and stored at −20°C prior to experimental use.

2.5.7 Mix preparation for PCR

PCR mix was prepared for each plex in 1mL eppendorf tubes using the reagents described in Table 2.3 below (separate tubes for each plex). To account for possible pipetting loss 25% extra volume was added (Table 2.3). Volumes were adjusted for a maximum plex level of 24 and for half of one 384-well reaction plate, assuming dry DNA is used. All reagents were thawed at room temperature then mixed gently before centrifuging (1500 rpm for 1 minute) prior to use. Due to its unstable nature, the Hot Star Taq enzyme was kept at a low temperature and so all reagents were placed on ice throughout preparation.

Table 2.3 PCR reaction cocktail solution preparation

(25% excess volume prepared to allow for pipetting loss)

Reagent	**Volume for single reaction**	**Volume for one 24-plex ($_{1/2}$ 384 well)**
Nanopure water	2.85µl	684.00µL
PCR Buffer (10x)	0.625µl	150.00µL
$MgCl_2$ (25mM)	0.325µl	78.00µL
dNTP mix (25mM)	0.10µl	24.00µL
Primer mix (0.5µM)	1.00µl	240.00µL
Hot Star Taq (5 U/µl)	0.10µl	24.00µL
Total volume	5.00µl	1200µl

2.5.8 Transfer of PCR mix and PCR reaction per plex

A 5uL aliquot of PCR mix was added to each well of the 384-well reaction plate containing dry DNA. Separate DNA plates were prepared for each plex. One 384-well plate generally consisted of two plexes, with half a plate dedicated to each plex. For one plex 1200uL PCR mix was first transferred into one row of a 96 well plate (total volume of mix divided into 8 columns), and 5uL was then transferred to the 384-well reaction plate using a multichannel 10uL pipette. The reaction plate was sealed with adhesive PCR film (AB-0558) to prevent evaporation and to allow plate to be spun down in a centrifuge (2000rpm for 2 minute). This ensured the solutions were at the bottom of the wells and any air bubbles were removed.

The pre-programmed PCR reaction on the G-Storm thermal cycler was then executed. The reaction volume was set to 7uL (additional 2uL accounted for air bubbles) and had a running time of approximately two hours and thirty minutes. The details of the cycling program are summarised below:

Cycle program

1 cycle:	5 min	94°C	(initial denaturation)
45 cycles:	20 sec	94°C	(denaturation)
	30 sec	56°C	(annealing)
	1 min	72°C	(extension)
1 cycle:	3 min	72°C	(final extension)
Final step:	indefinite	5°C	(hold)

Once completed, the 384-well plate was removed, sealed with the AB-0558 film, centrifuged (2000 rpm for 2 minutes) and stored at 4°C until required.

2.5.9 Post PCR cleanup

Treatment with SAP is performed after a PCR reaction in order to remove any remaining, non-incorporated dNTPs from the amplification products. SAP dephosphorylates unincorporated dNTPs by cleaving the phosphate groups from the 5′ termini, thereby rendering them inactive for future reactions. The SAP enzyme solution was prepared for each 384-well plate, according to Table 2.4. All reagents were defrosted at room temperature then mixed gently before centrifuging (1500 rpm for 1 minute). The reagents were placed on ice throughout the solution preparation.

Table 2.4 SAP solution preparation (38% excess volume for any pipetting loss)

Reagent	**Volume for single reaction**	**Volume for 384-well plate**
Nanopure Water	2.85µl	1368.00µL
hME Buffer (10x)	0.625µl	300.00µL
Shrimp alkaline phosphatase (SAP)	0.325µl	156.00µL

This procedure was performed on a post-PCR automated Matrix Liquid Handler robot using the SAP addition program. Prior to running any program on the Matrix Liquid Handler robot, a weekly maintenance protocol was performed as stated in the manufacturer's guidelines. A new tip magazine was inserted into the robot for each new program that was performed. A liquid handler tip wash program was also initiated prior to running any post-

PCR program. To transfer the SAP enzyme solution into the 384-well PCR plate, firstly a 96 well plate was prepared with 10uL of SAP in each well. The Liquid Handler robot SAP program was then used to dispense 2µl of the SAP cocktail from the 96-well microplate into each individual well of the 384-well post-PCR reaction plate. After SAP cocktail addition, the plate was removed from the robot, sealed using the AB-0558 PCR film and centrifuged (2000 rpm for 2 minutes). SAP treated plates were then placed on the G-Storm thermal cycler for a 50 min incubation as detailed below. The final reaction volume used was 9µL (additional 2µL added to account for air bubbles).

1 cycle:	**40 min**	**37°C**
1 cycle:	**10 min**	**85°C**
Final step:	**indefinite**	**4°C**

Once the SAP enzyme reaction was completed, the 384-well PCR reaction plate was sealed using AB-0558 film, centrifuged (2000 rpm for 2 minutes), then stored at 4°C until ready to process for the iPLEX Gold primer extend procedure.

2.5.10 iPLEX Gold primer extend reaction

The iPLEX primer extend reaction is a method for detecting single base polymorphisms in amplified DNA. Extension primers, buffer, enzyme, and mass-modified dNTPs are added to the amplification products. Each extension primer anneals directly 5' to the SNP locus and is extended by one mass-modified nucleotide (present in the iPLEX termination mix) based on the alleles present. This results in single base elongation with a corresponding mass increase that is measured using the MALDI-TOF MassARRAY platform and SNP genotype assigned accordingly.

2.5.11 Primer pooling and dilution

A four-step adjustment method based on primer concentration was used for pooling the extension primers into multiplexed pools (Table 2.5). Due to the inverse relationship between peak intensity and analyte mass, the iPLEX extension primers required adjustment by concentration in order to ensure that they were as equal in intensity as possible. For this, the primers were adjusted by dividing each plex into four concentration groups based on primer mass. The highest mass group was diluted to 7µM, the next groups to 9.3 µM and 11.66 µM, and the lowest mass group to 14 µM, as shown in the Table 2.5 below. All diluted working

stocks and concentrated stocks of primers were stored at −20°C.

Table 2.5. Preparation of iPLEX Gold Extend primers (total volume 500µL)

Extension primer group	**Final concentration/ primer**	**Volume / primer**	**24-plex (6 primers) (µl)**
1	7µM	8.75µL	52.50
2	9.3µM	11.63µL	69.78
3	11.66µM	14.58µL	87.48
4	14µM	17.5µL	105.00
Total volume of nanopure water to add (final volume500µL)			185.24µL

2.5.12 Cocktail preparation

The iPLEX reaction cocktail was prepared for each plex as described in Table 2.6 below. Volumes shown are for half of one 384-well plate. All reagents were defrosted at room temperature, mixed, centrifuged gently (1500 rpm for 1 minute) and kept on ice throughout the cocktail preparation procedure.

Table 2.6. iPLEX Gold extend reaction cocktail solution preparation (38% excess volume for any pipetting loss)

Reagent	**Volume for single reaction**	**Volume for one plex**
Nanopure Water	0.755µl	200.05µl
iPLEX Buffer (10x)	0.2µl	52.99µl
Primer mix (0.5µM)	0.2µl	52.99µl
iPLEX Termination mix	0.804µl	213.03µl
iPLEX Enzyme	0.04µl	10.6µl
Total volume	2.0µl	529.67µl

2.5.13 Cocktail transfer

10uL of the iPLEX cocktail was first added to each well of a 96-well plate and 2 µl of the cocktail was subsequently added to each well of the 384-well post-SAP reaction plate using the liquid handling robot. After cocktail addition, the plate was sealed using AB-0558 PCR film and centrifuged (2000 rpm for 2 minutes) to bring the solution to the bottom of the wells and remove any air bubbles before running the extend reaction.

2.5.14 Primer extend

The 384-well reaction plate containing the iPLEX cocktail was then placed in the G-Storm Thermocycler and the iPLEX extend reaction was executed. The reaction volume was set to 11uL (additional 2µL added to account for air bubbles). The details of the iPLEX Gold extend reaction cycle are summarised below.

Thermal cycling primer extend reaction

Number of cycles	**Time**	**Temperature**	**process**
1 cycle:	30 sec	94°C	(initial denaturation)
40 cycles:	5 sec	94°C	(denaturation)
***5 cycles*:**	5 sec	52°C	(annealing)
	5 sec	80◦ C	(extension)
1 cycle:	3 min	72°C	(final extension)
Final step:	indefinitely	4°C	(hold)

(Note that the 5 cycles sit within the 40 cycles)

Upon completion of the extend reaction cycle, the plate was centrifuged for 2 minutes at 2000 rpm and stored at 4°C.

2.5.15 Post iPLEX reaction conditioning

The conditioning or clean-up of iPLEX Gold reaction products is a crucial step for optimising the mass spectrometry analysis. SpectroCLEAN is a cationic resin pretreated with acid reagents that is added to primer-extend reaction products to remove salts such as Na^+, K^+, and Mg^{2+} from unincorporated products from the reaction. If not removed, these ions can result in high background noise in the mass spectra, thus increasing the likelihood of false data.

The SpectroCLEAN resin was transferred from its container to a 384-well dimple plate using an elongated spoon and then spread across the whole plate using a plastic scraper. Excess resin was then scraped away from the plate and placed back into the original container. The resin was then allowed to dry for 15 minutes. Whilst the resin was left to stand, the Matrix liquid handler robot was used to add 16µl of Nanopure water to each well of the 384-well post-iPLEX reaction plate. Once the water addition was complete, the 384-well PCR plate was sealed and again centrifuged (2000 rpm for 2 minutes).

After removing the plate seal, the 384-well reaction plate was turned upside-down and gently placed on top of the resin dimple plate. Holding the sample plate and the dimple plate together, they were then both gently flipped over to allow the resin to fall out of the dimple plate into the wells of the 384-well reaction plate. The dimple plate was then tapped gently until all the resin fell out into the wells of the 384-well reaction plate. Each well was manually checked for resin addition. The plate was sealed using AB-0558 PCR film and secured between two polystyrene blocks of the Heidolph®-Reax 2 rotator, and rotated for 10 minutes on the lowest setting (level 1), to allow the resin to mix thoroughly with the reaction plate products. Once completed, the 384-well PCR plate was centrifuged for five minutes at 3000 rpm to allow the SpectroCLEAN resin to settle down into the wells.

2.5.16 Mass spectrometry

The manufacturer's protocol was followed for arraying the extended products from the 384-well reaction plate on to a 384-sample SpectroCHIP using the MassARRAY Nanodispenser instrument. A small volume (~25nl) was arrayed by the dispenser onto the existing matrix spots on the SpectroCHIP for MALDI-TOF analysis. This process involved the capillary action of slotted pins and contact dispensing for nano-volumes (Gabriel et al., 2009).

2.5.17 MassARRAY spectroscopy methodology

The Sequenom MassARRAY MALDI-TOF platform and Sequenom real-time software was used in order to detect the extended products. The spotted SpectroCHIP was placed in the scout plate (chip holder) of the mass spectrometer, introduced to the MassARRAY reader, and

then placed on vacuum and the analyser software started (FlexControl, ServerControl, MassARRAY Spectro Caller 3.4, SpectroAcquire 3.4, Typer ChipLinker).

A virtual experimental plate was created using Plate and Assay editor on the MassARRAY® Spectra Typer software. Typer produces spectral data acquired from SpectroCHIPs and analyses each spectrum based on the assay or assays applied to it. An assay establishes where mass peaks are expected in a spectrum and how to interpret each peak. Typer automatically identifies the genotype in genotyping experiments based on the peaks present in a spectrum. Individual samples and assays appropriate for a particular experiment were assigned to each well on the virtual plate.

The Chiplinker software was used to connect the virtual chip layout created to the chip being analysed. Once files were created on ChipLinker software, this was linked to the SpectroAcquire software that controls the mass spectrometer and acquires spectral data. The total time for detection of one SpectroCHIP is 30 to 60 minutes. Spectral data is automatically sent to the MassARRAY Typer Server. These are then analysed by Typer, which combines the base caller with a clustering algorithm.

2.5.18 MassARRAY reaction

The general principal of the MassARRAY platform is to use MALDI-TOF mass spectrometry to determine differences in primer masses due to changes in sequence, i.e. the incorporation of different terminator nucleotides at the 3' end of a primer bound adjacent to a variant site (Gabriel and Ziaugra, 2004, Gabriel et al., 2009). The mass spectrometry system involves the laser treatment of the spotted sample under vacuum by the MALDI-TOF method. This method is a modified version of a standard mass spectrometry technique that involves the absorption of most of the incident laser energy, allowing the de-absorption and the ionisation of large biomolecules such nucleic acids with minimal damage and ion fragmentation. High transmission and sensitivity, along with theoretically unlimited mass range, are some of the main advantages of TOF instruments. The theory behind the stages of the MALDI-TOF process is described briefly below:

Sample irradiation and ionisation: The spotted samples (embedded in crystalline structure or matrix of small organic compounds) are irradiated with a nanosecond of ultraviolet laser (wavelength 337 nm). The laser energy causes structural decomposition of the irradiated crystal (ionisation) and generates a rapidly expanding matrix cloud.

Electrostatic acceleration: Once the sample molecules are vaporised and ionised, they are transferred into a time-of-flight mass spectrometer, where they are separated from the matrix

ions by an electric field that results in the disintegration of the crystal molecules. Following acceleration through an electric field, the ions drift through a field-free path and finally reach the detector in the form of a secondary electron multiplier.

Detection of ions using TOF: The ions are individually detected based on their mass-to-charge (m/z) ratios and analysed. Ion masses (m/z ratios) are calculated by measuring their flight time, which is longer for larger molecules and shorter for smaller molecules.

2.5.19 Genotyping quality control

In addition to the quality control (QC) procedures applied to the sample processing it was also necessary to apply separate quality checks on the outputted genotype data, as described below. Any samples that failed these QC measures were re-genotyped on a single 384-well plate, where practicable.

Negative control wells: The control wells (no DNA added) that were included in each 384-well reaction plate were first inspected to check for contamination. This would indicate false positive results and unreliability in the calls assigned to the surrounding samples.

Positive control wells: The duplicate samples that were included in each reaction plate were checked for consistency of genotype calls. Any duplicate samples for which there were inconsistencies in the assigned genotypes were marked for exclusion from the data analysis. The spectra for these samples were also checked to assess the quality of the peaks from which bases are called, prior to exclusion.

Spectra check: The Typer software provides a genotype call and spectrum for each sample. Each sample spectrum is annotated with the expected location of allele peaks and the un-extended primer peak. In some cases, contaminant peaks are also indicated. The spectra of samples that i) failed genotyping, ii) were either negative or positive controls, and iii) required repeating were all checked to assess sample genotyping quality.

Cluster graph check: The cluster graphs that are produced for each assay were examined carefully to assess the quality of genotyping of a particular assay. Cluster graphs are useful as they provide a visual description of genotype calls for an assay on a SpectroCHIP, thus they can help to determine if an assay is reliable. If there were chemistry problems with an assay, they usually appear in these cluster graphs. The cluster graphs were also checked before any manual calling decisions.

Histogram check: the Typer software produces a single histogram summarising the success of all included assays. For each experimental run this summary histogram was checked for a quick overview of all problematic assays (assays with a large failure percentage).

Manual calling for all failed samples and inconsistencies: For any samples that i) failed genotyping (i.e. those for which the software was unable to assign a genotype) or ii) samples for which the assigned genotype was questionable, (i.e. negative and positive control samples) the spectra were reviewed and genotypes manually assigned where possible.

2.5.20 Pre-analysis quality control

Before the genotype data for an experiment was analysed, all data that survived the above genotype quality control checks was subject to the following data quality checks:

Patient success: Typically, patients with less than 90% call rate for all genotyped SNPs (i.e. with genotype data at fewer than 90% of typed loci) were excluded.

SNP success: Typically, individual SNP assays with a call rate of less than 90% (i.e. successfully typed in fewer than 90% of patients) were excluded.

Hardy-Weinburg equilibrium and minor allele frequency: Each of the remaining SNPs were tested for deviation from Hardy-Weinberg equilibrium (HWE) using Haploview software version 4.1 (Barrett et al., 2005). In general a p-value of less than 0.001 was assumed to indicate deviation (a significant difference between observed and expected genotype frequencies), and such SNPs were excluded from data analysis. SNPs with a MAF of less than 0.001 as calculated by the Haploview software, were too low for the reliable detection of any genetic association and were also excluded from the analysis.

Comparison of minor allele frequencies to the general population: The frequency of the polymorphic allele for each assay was compared to that of the general population (manual comparison using frequencies from HapMap), in order to confirm the reliability of the genotyping and that the sample population was representative of the general population in terms of their genetic structure.

2.6 Data analysis

2.6.1 Statistical analysis

The majority of statistical analyses were performed using SPSS software (version 18.0; SPSS Inc., Chicago, IL, USA). Specific tests performed for each study varied and details of these, including the use of additional statistical software can be found within each results chapter. Correction for multiple testing was undertaken by calculation of the false discovery rate (FDR) for each test (Benjamini et al., 2001) using the 'p.adjust' function in the statistical package R, with an FDR <0.05 deemed statistically significant (R Development Core Team (2010).)

2.6.2 Bioinformatics analysis

Several freely available online tools were used for the purpose of predicting the potential biological significance of any associations with genetic variants identified from the statistical analysis for each of the studies. This included the use of online genomic databases described previously (section 2.4.1) and additional databases specifically allowing the search for TFBSs and regulatory regions and/or predicting functional changes in protein coding regions (Pang et al., 2009).

2.6.2.1 Fast SNP and PupaSuite

Several tools exist to predict regulatory regions and then cross check them with databases of known SNPs to highlight which SNPs fall in these regions. These include the freely available and widely used Function Analysis and Selection Tool for Single Nucleotide Polymorphisms (Fast SNP) (Yuan et al., 2006) and Pupa Suite (Conde et al., 2006), which were the main two tools used in the present studies. Each tool uses different means to predict the regulatory regions. These tools run programs such as splicing site enhancers (ESE)-Finder and Transfac (for locating TFBS) for both the wild type and the variant sequences and check whether they differ in their results, i.e. whether one has a predicted ESE within it and the other not.

PupaSuite (http://pupasuite.bioinfo.cipf.es/) retrieval of the location of SNPs in TFBS, ESE, splicing site silencers (ESS), and splice sites (SS) using both Transfac and JASPAR, ESE-Finder3.0, ExonScan and GeneID respectively. Fast SNP is a web server that allows users to efficiently identify SNPs of potential biological significance according to twelve phenotypic risks and putative functional effects, such as changes to the transcriptional level, pre-mRNA splicing, protein structure, etc. Fast SNP can be used to find SNPs in genomic and mRNA sequences using the following tools; ESS (FAS-ESS), ESE (both Rescue-ESE and ESEfinder),

TFBS (TFSearch), and Polymorphisms Phenotyping (PolyPhen) to look at non-synonymous SNPs in protein sequences.

2.6.2.2 Predicting presence and functional consequences of variants; coding, promoter region, intronic and synonymous

TFBSs can be found within both promoter and intronic regions of DNA. All non-coding variants of potential interest were evaluated for the presence of putative binding sites of known transcription factors (TFs) using the following search databases: Transcription Element Search System (TESS, http://www.cbil.upenn.edu/tess) and Fast SNP (http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp). Fast SNP identifies and predicts changes in TFBS regions using the TF search tool.

The effect of non-synonymous or coding variants on protein function were predicted using Sorting Intolerant From Tolerant (SIFT) and Fast SNP. SIFT (http://sift.jcvi.org/) uses a sequence alignment method to measure conservation of each amino acid, predicting whether a coding SNP will affect protein function (by calculating a scaled probability for the amino acid substitution using sequence homology and the physical properties of amino acids). Fast SNP utilises the PolyPhen tool for predicting protein structural changes and these predictions are based on physical and comparative considerations that estimate the impact of the amino acid change on the 3D structure and function of the protein. Fast SNP was also the main tool used for analysis of all intronic and synonymous variants in order to assess their potential effect on regulatory regions.

2.6.2.3 Machine learning and SAS Enterprise Miner

For two of the research studies presented in this thesis (those consisting of a large number of SNPs; over 1000), in addition to standard parametric statistical analysis methods for detecting genetic association, a ML data-mining approach was adopted in order to i) build predictive models through extracting patterns from the large genomic data available, ii) as a more appropriate method for analysing high-dimensional and complex genomic data-sets. Several well-known ML models were utilised for this, each of which are described in the corresponding chapters for these studies (Chapter 5 and 6). In house-software was utilised for the ML approach used in Chapter 5 (Petrovski et al., 2009). The additional ML models used in Chapter 6 were generated and assessed using SAS® Enterprise Miner data-mining software.

CHAPTER THREE

CARBAMAZEPINE DOSE REQUIREMENT AND GENETIC VARIATION IN DRUG METABOLISING ENZYMES

CONTENTS

3.1. Introduction

CBZ is a widely used AED that has been employed as first-line treatment for partial and generalised tonic-clonic seizures for over 40 years (Brodie and Dichter, 1997). Like many older AEDs, CBZ undergoes predominantly hepatic metabolism and has a recognised therapeutic concentration range (Kwan and Brodie, 2001a). It also demonstrates considerable inter-individual variability in terms of PK and dosing requirement for effective seizure control, with maintenance doses in clinical practice often ranging from 200 to 2000 mg/day (Kwan and Brodie, 2001a).

3.1.1 Antiepileptic drug dosing

Therapeutic doses of AEDs are less well defined than those of drugs prescribed in many other disease areas and are typically influenced by titration regimen (Shorvon, 2004). Current monotherapy treatment with CBZ involves slow titration of the drug over a six-week period to a modest target dose (usually 600mg/day), with subsequent dosage adjustment according to clinical response (Shneker and Fountain, 2003). This approach is, however, sub-optimal for many patients. Those with a low CBZ dose requirement may develop early adverse effects including possible hypersensitivity reactions, whereas those with a high dose requirement are likely to be under-dosed for a significant period and subject to ongoing seizure activity. Sub-optimal dosing may also lead to patients switching to alternative AEDs to achieve an adequate response without the complete dosage range being fully explored (Perucca, 2001a). Thus, determining the dose of CBZ that provides maximal seizure control with minimal adverse effects for individual patients can be challenging and quality of life is often compromised until this is achieved (Depondt, 2006b, Depondt and Shorvon, 2006).

There is increasing awareness that dose requirements of AEDs vary greatly from one patient to another. This variability has led to the rejection of a standard dose approach to treatment and requires consideration of tailored drug therapy (Shorvon, 2004). Development of individualised dosing strategies for AEDs such as CBZ has the potential to improve the treatment of epilepsy by providing more prompt seizure control and safer drug initiation.

3.1.2 Variability in carbamazepine pharmacokinetics

For most AEDs, the serum concentration at any given dose can vary up to 50-fold between individuals (Perucca et al., 2001). Inter-individual variability in dose requirement results, at least in part, from variability in PK factors (Perucca, 2001a) that can be monitored through measurements of serum drug concentration (Perucca et al., 2001). Therapeutic drug

monitoring (TDM) has been used widely in epilepsy for improving drug efficacy and tolerability and promoting individualization of therapy (Nagasawa and Nakahara, 1992). However, not all current AEDs are candidates for TDM; the newer-generation of AEDs are thought to possess more linear PK and are less likely to cause drug-drug interactions (Clarke and McMillin, 2006, Anderson, 2008). Moreover, there remains disagreement regarding the value of TDM in routine AED treatment (Perucca et al., 2001, Johannessen et al., 2003).

Numerous factors are known to influence the serum concentrations and dose requirements of therapeutic agents, including age, gender, body weight and co-medications (Levy, 2002, Battino et al., 2003, Engel and Pedley, 2008). These patient-specific influences on drug PK are reasonably well characterised for AEDs, although their clinical utility is generally limited (Perucca et al., 2001, Perucca, 2002b). Variability in PGx genes, i.e. those encoding drug transport proteins, metabolic enzymes, and drug targets, are also increasingly recognised as contributors to PK heterogeneity (Kirchheiner and Seeringer, 2007).

3.1.3 Variation in metabolising enzymes as determinants of dosing

Genetic polymorphisms are known to affect the metabolism of many drugs (Weinshilboum, 2003, Kirchheiner and Seeringer, 2007, Bhathena and Spear, 2008). This contribution to variability in drug metabolism may be reflected in differences in clearance, half-life and maximal plasma concentrations and can be corrected by genotype-based dose adjustments (Ma et al., 2002, Kirchheiner and Brockmoller, 2005, Crowley et al., 2009). The effect of polymorphic metabolism is particularly evident for substrates of CYP isoform 2D6, an enzyme that displays a variety of genetically determined phenotypes, including poor, intermediate, extensive, and ultra-rapid metabolism (PM, IM, EM and UM, respectively) (Wilkinson, 2005, Kirchheiner and Seeringer, 2007). Although none of the current AEDs is a substrate for CYP2D6, many undergo extensive Phase I hepatic metabolism mediated by at least eight other members of the CYP superfamily (Klotz, 2007). Figure 3.1 below presents the main pathways of metabolism known for CBZ (Pearce et al., 2008).

CBZ mainly undergoes hepatic metabolism (Eichelbaum et al., 1985), predominantly mediated by CYP3A4 and CYP3A5 enzymes (Tomson et al., 1983, Saruwatari et al., 2010). Other CYP contributors include CYP1A2 and CYP2C8, with an additional role of the Phase II UGT2B7 enzyme, while its principal active metabolite, CBZ-10,11-epoxide (CBZ-E), undergoes biotransformation mediated by mEH (Tomson et al., 1983, Saruwatari et al., 2010). All of these enzymes have known polymorphisms that potentially influence their metabolic activity and could theoretically impact on the PK of CBZ (Kirchheiner and Seeringer, 2007, Saruwatari et al., 2010).

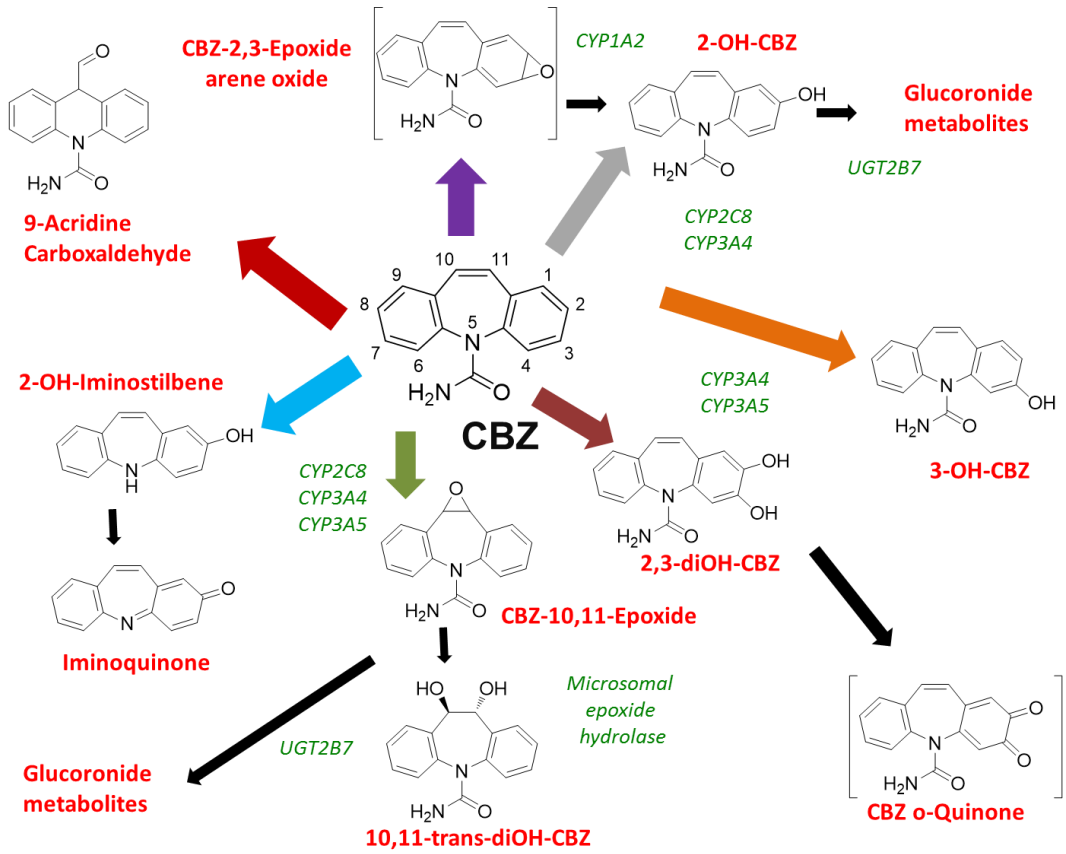


Figure 3.1 Metabolic pathways and proposed metabolites of carbamazepine

The main metabolic pathways of carbamazepine and the major metabolites formed during its metabolism are shown (separated by boxes). Some of the enzymes proposed to be involved in these biotransformation pathways are also highlighted. Figure adapted from Pearce, Lu *et* al 2008.

3.1.4 Effect of CYP450 variants on carbamazepine pharmacokinetics

A good example of the functional consequence of AED metabolism by polymorphic CYP enzymes is that of PHT and the CYP2C9 and CYP2C19 isoforms (Saruwatari et al., 2010, Cavalleri et al., 2011). The genes encoding CYP2C9 and CYP2C19 have well-characterised functional variants that exhibit different drug metabolism phenotypes, similar to that of *CYP2D6* (Klotz, 2007). Several studies have demonstrated that individuals with defective alleles for *CYP2C*9 or *CYP2C19* have reduced PHT metabolism, leading to both a lack of efficacy with PHT treatment and in many cases drug toxicity (Klotz, 2007, Anderson, 2008, Loscher et al., 2009). Other AEDs, including PB, diazepam, VPA and ZNS that are substrates for CYP2C9 and/or CYP2C19 have likewise shown reduced metabolism rates in individuals with *2/*3 alleles, when compared to those with the wild-type CYP allele (Klotz, 2007, Anderson, 2008, Seo et al., 2008a, Loscher et al., 2009, Saruwatari et al., 2010).

In addition to the CYP2C enzymes, recent PGx evidence has implicated a known functional polymorphism in *CYP3A5* with altered serum concentrations of CBZ (Park et al., 2009, Meng et al., 2011) and a lower dose requirement during CBZ treatment (Meng et al., 2011). *CYP3A5*3* SNP (rs776746) encodes a truncated non-functional protein causing a loss of CYP3A5 enzymatic activity (Kuehl et al., 2001, Lin et al., 2002, Yamaori et al., 2004) and has been associated with altered PK parameters of several CYP3A substrates (Huang et al., 2004).

3.1.5 Phase II metabolism of carbamazepine

Further evidence for potential genetic influences on the hepatic metabolism of AEDs has recently emerged for UGT2B7 (Chung et al., 2008, Blanca Sanchez et al., 2010). The UGT2B enzyme family is highly polymorphic, containing several well characterised functional polymorphisms (Burchell, 2003), and may be responsible for inter-individual variation in the detoxification of metabolites, including several carcinogens (Desai et al., 2003). In addition to CBZ, UGT2B7 also contributes to the glucuronidation of LTG, VPA, OXC and ZNS (Staines et al., 2004, Rowland et al., 2006, Chung et al., 2008). The functional *UGT2B7*2* variant is associated with enhanced metabolism of some opioids and has also been suggested to increase the area under the curve (AUC) of VPA (Chung et al., 2008, Blanca Sanchez et al., 2010). A *UGT2B7* promoter region variant (*UGT2B7 -161C>T*), believed to be in LD with the *UGT2B7*2* SNP, has also been reported to alter serum AED concentrations (Blanca Sanchez et al., 2010). In this report by Blanca Sanchez and colleagues, the *UGT2B7*2* variant was associated with LTG concentration/dose ratio in a multivariate model adjusted for potentially confounding factors such as age and co-medication with VPA and was found to explain 12% of the dose variation (Blanca Sanchez et al., 2010). Although this association was modest in

terms of effect size, it is the first study to implicate genetic variations in UGT enzymes with variability in AED PK (Chung et al., 2008, Blanca Sanchez et al., 2010).

In contrast to UGT2B7, the phase II enzyme mEH has been the focus of several AED gene-association studies (Cavalleri et al., 2011). Increasingly, research has shown that haplotypes in LD blocks are more precise for detecting un-observed phenotype–genotype links than individual SNPs (Zhang et al., 2002, Nakajima et al., 2005). Haplotypic variation within the *EPHX1* gene encoding mEH has been reported to correlate with plasma concentrations of the CBZ metabolites CBZ-diol and CBZ-E in a Japanese study. The CBZ-diol to CBZ-E ratio differed greatly depending on the number of variant alleles of two known *EPHX1* non-synonymous polymorphisms: *EPHX1*-Try113His (337T>G; rs1051740) and *EPHX1*-His139Arg (416A>G; rs2234922) (Nakajima et al., 2005). Ratios increased significantly with 337T>G-bearing haplotypes and decreased significantly with 416A>G-bearing haplotypes (Nakajima et al., 2005). These known functional polymorphisms have since been associated with maintenance dose in a CBZ monotherapy study when considered in a multivariate model with age (Makmor-Bakry et al., 2009).

The handful of association studies that have correlated AED PK with genetic variation in DMEs suggest that this is an important area that merits further investigation with regard to individualization of AED dosing. Relatively few drugs and their corresponding metabolic pathways have been explored to date. Those studies that have reported genetic associations with dose or PK require replication to verify those associations and provide more definitive evidence that the observed effect is real and of sufficient magnitude to be considered clinically useful and implementable in a genotype-based dosing strategy.

3.2 Aims

The principal aim of the study presented in this chapter was to assess the degree to which genetic variation in drug metabolism contributes to CBZ dose requirement when used as monotherapy in newly treated epilepsy. An association analysis of common variation across genes encoding CBZ metabolising enzymes was performed, capturing variation by applying a gene-wide tagging methodology and undertaking a haplotype analysis to determine whether multiple variants in combination can be used to more successfully identify associations. A secondary aim was to use this analysis to validate a previous study that reported a significant influence of two functional variants (rs1051740 and rs2234922) in the *EPHX1* gene on CBZ dosing (Makmor-Bakry *et* al, 2009).

3.3 Methods

3.3.1 Selection criteria for patient inclusion and study population

Individuals were selected for the study from both SANAD and Glasgow cohorts on the basis of strict inclusion criteria. Patients were required to have a new or recent (within 3 years at the time of CBZ initiation) diagnosis of epilepsy and to have achieved optimal seizure control (defined as no seizures for a period of at least 12 months) on a fixed dose of CBZ monotherapy. This was subsequently referred to as the CBZ maintenance dose. Maintenance dose was defined as the uppermost stable dose or unchanged dose over the 12-month seizure-free period. The study population comprised 77 patients from the SANAD cohort and 90 patients from the Glasgow cohort (Table 3.1).

3.3.2 Clinical data collection

Non-genetic information for each patient was extracted from clinical databases, hospital notes or clinical trial folders, as appropriate. This included age (at the start of the 12 month seizure-free period), sex, epilepsy type and CBZ maintenance dose. Epilepsy type was defined as IGE, LRE, or UNC.

3.3.3 Candidate SNP selection

The aim of candidate SNP selection was to find common genetic variation within DMEs relevant to CBZ metabolism that might potentially affect dose requirement. A total of six genes were targeted; *CYP1A2*, *CYP2C8*, *CYP3A4*, *CYP3A5* (encoding the corresponding CYP enzymes), *EPHX1* (encoding mEH) and *UGT2B7* (encoding the corresponding UGT enzyme).

3.3.4 The International HapMap project

The objective of the International HapMap Project (www.hapmap.org) was to identify and record all genetic differences and similarities within human subjects. This involved genotyping at least one common SNP every 5 kilobases (kb) across the genome in 270 individuals from geographically diverse populations, including the Yoruba people from Ibadan, Nigeria, Caucasians of north and west European descent from the Centre d'Etude du Polymorphisme Humain (CEPH) research in the USA, 45 unrelated individuals from Beijing, China, and 45 unrelated individuals from Tokyo, Japan. The results of the project are freely available to researchers for use in genetic association studies.

Table 3.1 Characteristics of the carbamazepine patient population

Clinical characteristics of patients forming the study population included in the analysis* (n=159) reported by source cohort and in combination.

		COHORT		
		SANAD (n=71)	Glasgow (n=88)	Combined (n=159)
Age (years)	Minimum	6	13	6
	Median	36	32	35
	Maximum	78	68	78
Gender (n)	Male	38	42	80
	Female	33	46	79
Epilepsy type (n)	IGE	1**	15	16
	LRE	65	66	131
	UNC	5	7	12
CBZ maintenance dose (mg/day)	Minimum	400	200	200
	Mean (± SEM)	663 (± 23)	798 (± 35)	738 (± 23)
	Maximum	1400	2000	2000

*IGE = idiopathic generalised epilepsy, LRE = localisation-related epilepsy, UNC = unclassified epilepsy, CBZ = carbamazepine, SEM = standard error on the mean,*8 patients from the study population failed minimum genotyping criteria and were excluded*

*** Difference in number of IGE patients between the two cohorts can be attributed to the design and purpose of the SANAD trial: (individuals with partial epilepsy forming larger Arm A; n=1721 and those with generalised and unclassified epilepsies forming smaller Arm B; n=716)*

3.3.5 SNP selection methodology

The CEPH population data were interrogated for variation in all six DME genes using HapMap release # 24 (phase II Nov 08; NCBI build 36 assembly) and dbSNP on the NCBI website (www.ncbi.nlm.nih.gov). The reference genotyping data from HapMap was used to identify all known SNPs across each of the six genes that were present in individuals of Caucasian/European ancestry and with a MAF of at least 1%. Chromosomal positions of each gene were identified and coordinates extended by 10 kilobases upstream and downstream to include the 5' and 3' flanking regions. A list of tSNPs and putatively functional variants for each DME gene was then prepared, as described below.

3.3.6 SNP Tagging SNP approach for representing gene-wide variation

Candidate gene SNP association studies for complex traits need to screen a large number of SNPs to capture the potentially influential variation across the whole gene (Zhang and Sun, 2005). However, this is an expensive and time-consuming option for what are often small-scale studies (Johnson et al., 2001). In contrast, using a haplotype-tagging strategy in which a subset of SNPs, each of which acts as a marker for a genomic region (haplotype block) in which all variants are thought to be inherited together, reduces the number of SNPs required for genotyping (Hirschhorn et al., 2002, Zhang et al., 2002, Chapman et al., 2003). This reduces time and cost by avoiding typing redundant SNPs whilst maintaining sufficient coverage of genetic variation (Sabbagh et al., 2008).

The human genome can be divided into regions with low haplotype diversity and high LD, interspersed with regions of high haplotype diversity and low LD (Zhao et al., 2003). In regions of low haplotype diversity, typing a smaller number of markers or tSNPs would capture most of the haplotypic diversity due to LD between variants, and therefore, could potentially capture an association between a human trait and causal loci in each haplotype block (Chapman et al., 2003, Zhao et al., 2003). A tSNP is often in strong LD with several other SNPs. The assumption is thus that the tSNP selected for genotyping will capture all the other SNPs it tags (Zhao et al., 2003). The pairwise tSNP approach thus represents an indirect approach to identifying genetic association SNPs by utilising the LD between SNPs in close proximity and so it is usually not necessary to genotype all SNPs of interest (Shastry, 2004). The pairwise correlation coefficient (r^2 statistic) is a commonly used measure to quantify the degree of association between two polymorphisms (Chapman et al., 2003, Zhao et al., 2003).

3.3.7 Using Haploview and Tagger to generate a list of tagging SNPs

Haploview (version 4.1) and Tagger (de Bakker, 2009) were used to select tSNPs that capture common variation and putative regulatory regions up and down stream (within 10kb) of the DME genes. Haploview is a program designed primarily for haplotype analysis and has several functions, including LD and haplotype block analyses, haplotype population frequency estimation, single SNP and haplotype association tests, and permutation testing for association significance (Barrett et al., 2005). The tagger function in Haploview contains an algorithm that performs tSNP selection using the pair-wise method (de Bakker, 2009).

3.3.8 tSNPs and supplementary SNPs selected for genotyping

For tSNP generation, Caucasian/European genotype data previously downloaded from HapMap release # 24 (www.hapmap.org; phase II Nov 2008) for each DME gene (±10kb) was

imported into Haploview (as to match the ethnicity of SANAD/GLASGOW study cohort). SNPs meeting the following criteria were used; i) MAF ≥ 5%, ii) HWE cut-off p-value > 0.001 (Barrett et al., 2005), iii) SNP coverage of 80% (using r^2=0.8 ensures at least 80% correlation between the tSNP and all of the SNPs it tags), iv) ≥80% HapMap genotyping data for each common polymorphism, and v) Mendelian inheritance errors in the HapMap CEPH population of no greater than 1. An additional set of SNPs with a particularly low MAF (≥ 0.1%) were chosen to allow the capture of more SNPs from coding regions and/or those reported in previous association studies. The pair-wise Tagger function was then executed.

A total of 52 tSNPs were identified across the six genes as follows: 1 from *CYP1A2*, 13 from *CYP2C8*, 8 from *CYP3A4*, 4 from *CYP3A5*, 18 from *EPHX1*, and 8 from *UGT2B7* (Table 3.2). These were then supplemented with a further 42 SNPs (12 *CYP1A2*, 6 *CYP2C8*, 3 *CYP3A4*, 5 *CYP3A5*, 12 *EPHX1*, 4 *UGT2B7*) reported as either being putatively functional in existing literature or located in gene regions with potential functional significance (i.e. exon, 3'-UTR, 5'-UTR, promoter region, splice site and enhancer site region SNPs) and possessing a MAF ≥1% according to the NCBI SNP database (build 126) (Table 3.2). This resulted in a list of 94 candidate SNPs across each of the six DME genes being chosen for genotyping. A full list of all 52 tSNPs for these SNPs is provided in Appendix 1.3.

Table 3.2 SNPs and tagging SNPs selected and genotyped for each candidate gene

DME Gene	Nº of tSNPs identified	Nº of supplementary SNPs	Nº of SNPs genotyped
CYP1A2	1	12	13
CYP2C8	13	6	19
CYP3A4	8	3	11
CYP3A5	4	5	9
EPHX1	18	12	30
UGT2B7	8	4	12

DME = drug metabolising enzyme, Nº = number, SNPs = single nucleotide polymorphisms, tSNPs = tagging single nucleotide polymorphisms

3.3.9 Genotyping methods

The online Sequenom MassARRAY iPLEX assay design software (https://mysequenom.com/Tools/genotyping/default.aspx)*(Gabriel et al., 2009)* was used for primer and assay design for all 94 SNPs, as detailed in Section 2.5.3. Three SNPs (rs7438284, rs11773597 and rs45540739 from *UGT2B7*, *CYP3A4* and *EPHX1*, respectively) were excluded during the assay design phase as a result of their predicted potential for cross binding and introduction of genotyping errors. These could not be accommodated in any of the five plexes or replaced with alternative tSNPs and were thus excluded from the analysis. In total, 91 SNPs within five multiplex assays (plexes 1-5), comprising panels of 23, 21, 21, 20 and 6 SNPs respectively were produced by the software (Table 3.3). DNA for all 167 patients was genotyped for the 91 SNPs. PCR conditions and extension primer sequences are listed in Appendix 1.1. Genotyping was performed on the Sequenom MassARRAY iPLEX platform (Sequenom, Hamburg, Germany) as described in Chapter 2 and in accordance with the manufacturer's instructions (Gabriel et al., 2009).

Table 3.3 Assay design output of Candidate SNPs

The 91 candidate SNPs selected for genotyping were placed into 5 SNP plexes by the Sequenom MassARRAY iPLEX assay design software

DME gene	**SNPs**					
	PLEX 1 (23)	PLEX 2 (21)	PLEX 3 (21)	PLEX 4 (20)	PLEX 5 (6)	Total
CYP1A2	2	3	3	4	1	13
CYP2C8	9	3	3	4	1	20
CYP3A4	2	3	5	-	-	10
CYP3A5	1	5	-	2	1	9
EPHX1	7	4	7	8	2	28
UGT2B7	2	3	3	2	1	11

DME = drug metabolising enzyme, SNPs = single nucleotide polymorphisms

3.3.10 Processing of genotype data for quality control purposes

A total of 10 positive control samples (duplicates) and two negative control samples (water blanks) were included per 384-well reaction plate for each experiment to improve reliability of genotype calls. Patient and sample QC measures, as described in section 2.5.20, were applied. There was also a purposeful reduction in data dimension prior to analysis in order to decrease the number of variables and limit the impact of correction for multiple testing. This was achieved by exploring LD structure across the SNP panel in the study population. For each pair of highly correlated SNPs ($r^2 \geq 0.9$), the variant with the fewest missing data was retained, whilst the other was excluded. A pair-wise correlation of $r^2 \geq 0.9$ allowed accurate model fit with retention of the majority of genetic variation.

3.3.11 Bioinformatics analysis

In addition to exonic SNPs that may directly influence amino-acid sequence, many SNPs are also found in splice sites, enhancer or silencer sites, and TFBS (Pagani and Baralle, 2004, Schug, 2008, Kasowski et al., 2010) and may still affect protein expression and the transcriptional efficiency of protein coding genes (Prokunina and Alarcon-Riquelme, 2004, Pang et al., 2009). Since functional studies are usually time-consuming, they tend to be initiated only when a statistically significant association with a phenotype is already established and has been replicated (Prokunina and Alarcon-Riquelme, 2004, Pang et al., 2009). Several online bioinformatics databases were used to predict potential functional and/or expression effects of SNPs (Pang et al., 2009). These included FASTSNP (Yuan et al., 2006), TESS (Schug, 2008), and SIFT (Ng and Henikoff, 2001, 2003, Ng et al., 2009). Ensemble Human Genome Browser and UCSC Genome Browser were also used to visualise and explore genetic variation within each of the six genes (see sections 2.4 and 2.7).

3.3.12 Statistical analysis

Statistical analysis was performed as described in section 2.6. Haploview (version 4.1) was used for haplotype analysis and PHASE software (version 2.1) to infer likely haplotype pairs (Stephens et al., 2001, Stephens and Donnelly, 2003, Scheet and Stephens, 2006).

CBZ maintenance dose (expressed as mg/day) was the phenotype of interest in this analysis. It showed a skewed distribution (Figure 3.2) and was log-transformed to achieve normality and to allow parametric statistical testing. Source cohort (SANAD or Glasgow) was included as a covariate in the analysis to account for any fundamental differences in patient characteristics.

3.3.13 Non-genetic univariate association with carbamazepine dosage

Testing for confounding non-genetic factors that may associate with dose was required to exclude their potential influence on inter-individual variability in dose. Initial analysis tested for association between CBZ maintenance dose and the following non-genetic variables; age at the start of the seizure-free period, sex, epilepsy type, and source cohort (SANAD and Glasgow). Univariate linear regression was used for testing age as a continuous variable, and analysis of variance (ANOVA) was used for analysing the categorical variables (sex, epilepsy type, source cohort and genotype).

3.3.14 Single variant analysis

All SNPs were analysed individually for association using regression statistics. A multiple linear regression model was built for non-genetic factors that proved significant ($p<0.05$) in the initial analysis. Thereafter, the regression model was re-fitted by the inclusion, as a covariate, of each of the individual candidate SNP genotypes included in the analysis in turn, assuming an additive mode of inheritance. The likelihood ratio test (LRT) was then used to compare the initial baseline model (containing non-genetic factors alone) with the genotype model (containing non-genetic factors and single SNP genotype), with adjustment for the potentially confounding non-genetic factors, to test for association between individual candidate SNPs and CBZ maintenance dose. A chi-square distribution test p-value was generated to assess the significance of any association.

3.3.15 Haplotype analysis

In addition to single SNP analysis, a gene-based haplotype analysis was also undertaken as an alternative means of detecting genetic associations with dose. As haplotype blocks define a region of a chromosome that is unlikely to undergo recombination, they provide greater power to detect likely causative alleles within large genetic data sets. In any DNA sample, there are two copies of each gene (one maternal and one paternal), which are typically different. Genotyping technologies, when applied to DNA from a diploid individual, are able to determine which two alleles are present at each locus but not which combinations of alleles are present on each of the two chromosomes. Such haplotype information or haplotype phase requires determination.

All genotype data was entered into Haploview, together with the chromosomal location for all SNPs (Barrett et al., 2005). The pattern of LD between each of the included SNPs and the haplotype blocks existing across all six genes was visualised using the solid spine of LD method for defining blocks of LD. Maximum likelihood estimates of haplotype

frequencies from unrelated individuals (the most likely haplotype pair at each block and its associated probability), were inferred for each individual using fastPHASE software (Stephens et al., 2001, Stephens and Donnelly, 2003, Scheet and Stephens, 2006). Quality control was subsequently performed on the generated data. Individuals in whom the fastPHASE assigned haplotype-pair had a probability of <90% were first excluded. A common haplotype occurs in a population with a frequency of at least 5%. All common haplotypes were included in the analysis and rare haplotypes (occurring at a frequency of less than 5%), were grouped together for analysis.

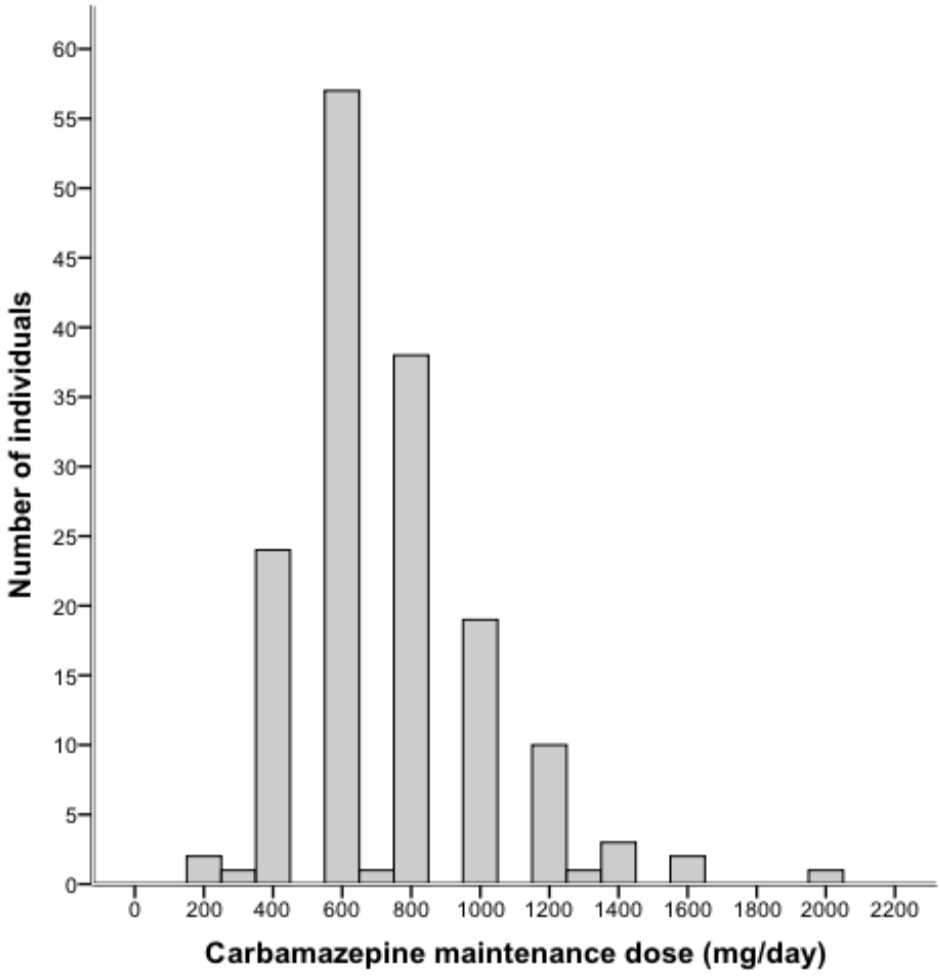


Figure 3.2 Carbamazepine dose distribution

Distribution of carbamazepine maintenance dose (mg/day) across the study population of n=159 individuals. Maintenance dose was defined as the uppermost stable or unchanged dose over a 12-month seizure-free period.

To test for association between CBZ maintenance dose and variation at each haplotype block, a regression-based approach was again employed, where the baseline model was compared to the haplotype model using the LRT and again assuming an additive mode of inheritance. The haplotype model was the same as the baseline model but in addition to non-genetic factors this included covariates to represent the haplotype pair assigned to the SNPs within the haplotype block for each individual. Further study of each of the phase-generated haplotypes across the gene-based haplotype block was only considered if a statistically significant association ($p<0.05$ after FDR) was identified in the initial regression analysis of the haplotype blocks.

3.4 Results

Of the 91 SNPs selected and genotyped, 15 failed genotyping due to an unsuccessful PCR and/or iPLEX reaction, 14 had a MAF <0.001, 3 deviated from HWE (HWP=<0.001), and 1 was monomorphic (Appendix 1.2). These were excluded from further analysis. With the remaining 58 SNPs, an additional effort was made to reduce data dimensionality, with 7 SNPs found to be in strong LD ($r^2 \geq 0.9$) with other genotyped variants and though not excluded, these were not included in the final data analysis. Of the 167 patients who underwent SNP genotyping, 8 had a genotype call-rate < 90% and were excluded from the analysis. This left 51 SNPs and 159 patients (71 SANAD, 88 Glasgow) for the association analysis. Basic demographic and clinical characteristics of the study population included in the analysis are reported in Table 3.1.

Of the remaining 51 candidate SNPs, 16 had previously been typed by the International HapMap project (NCBI build 36, dbSNP build 126) and had published MAFs that did not deviate from those observed in this study (Appendix 1.2). Several SNPs were selected on the basis of a previous report of association in literature. Of these n=3 were associated with AED serum concentration and/or dosing in epilepsy (rs776746, rs2234922, rs1051740)(Nakajima et al., 2005, Makmor-Bakry et al., 2009, Park et al., 2009, Meng et al., 2011). The putatively significant *CYP3A5*3* variant rs776746 C>T proposed to affect the metabolism of several drugs (Huang et al., 2004) and more recently reported to influence both CBZ serum concentrations and dosing (Park et al., 2009, Meng et al., 2011) however failed QC (HWP<0.1%), thus was not included in the final analysis.

3.4.1 Associations between genetic variants and maintenance dose

Of the four non-genetic factors considered in this study, only age (*P*= 0.014) and source cohort (*P*= 0.023) were significantly associated with CBZ maintenance dose, as shown in Table 3.4. Older ages and patients from the SANAD cohort appeared to have lower CBZ maintenance doses when analysed using a univariate regression model. When age and source cohort were included in individual regression models with each SNP genotype, eleven of 51 SNPs showed association with CBZ maintenance dose (Figure 3.3, Table 3.5a and Table 3.5b). Two of the SNPs were non-synonymous coding variants (rs4149229 in *EPHX1* and rs7439366 in *UGT2B7*), one was a synonymous coding variant (rs2234922 in *EPHX1*), and the remaining SNPs were located in non-coding or intronic regions. None of these associations survived FDR correction for multiple testing (Table 3.5).

Table 3.4 Univariate analysis of non-genetic factors

Regression analysis results for association between non-genetic factors associated with carbamazepine maintenance dose. A $p<0.05$ was considered significant.

Variable	Analysis	F-statistic	*P*-value
Age	Continuous	1.647	0.014
Gender	Categorical (male / female)	0.589	0.444
Epilepsy type	Categorical (IGE, LRE, UNC)	0.031	0.969
Source cohort	Categorical (SANAD, Glasgow)	5.248	0.023

IGE = idiopathic generalised epilepsy, LRE = localisation-related epilepsy, UNC = unclassified Epilepsy

3.4.2 Validation of previous *EPHX1* association with CBZ dose

In an effort to confirm the previously reported association between CBZ maintenance dose and two putatively functional SNPs in *EPHX1* (Makmor-Bakry et al, 2009), a further regression analysis incorporating age, source cohort, and the genotype of both SNPs was performed. Neither SNP in *EPHX1* was significantly associated with CBZ dose in isolation (uncorrected *P*= 0.494 for rs1051740, uncorrected *P*= 0.046 for rs2234922) and the regression analysis incorporating both loci was similarly unremarkable.

Table 3.5a Genotype associations with carbamazepine dose

Genotype associations identified by regression analysis; a null and alternative regression model was generated for each of the 51 single nucleotide polymorphisms and a chi-square test for statistical difference between the two models was performed.

SNP ID (rs)	Gene	*Uncorrected P*-value
rs4356975	*UGT2B7*	0.003
rs3924194	*UGT2B7*	0.003
rs4646450	*CYP3A5*	0.007
rs2292558	*TMEM63A*	0.007
rs4149229	*EPHX1*	0.010
rs7439366	*UGT2B7*	0.012
rs7375178	*UGT2B7*	0.014
rs1934956	*CYP2C8*	0.019
rs2246709	*CYP3A4*	0.026
rs12721617	*CYP3A4*	0.029
rs2234922	*EPHX1*	0.046
rs3738040	*EPHX1*	0.056
rs11572080	*CYP2C8*	0.061
rs11572126	*CYP2C8*	0.064
rs28365062	*UGT2B7*	0.065
rs2671272	*EPHX1*	0.088
rs2071426	*CYP2C8*	0.091
rs2275622	*CYP2C8*	0.120
rs2854461	*EPHX1*	0.124
rs1934980	*CYP2C8*	0.125
rs1536430	*CYP2C8*	0.125
rs3753660	*EPHX1*	0.128
rs762551	*CYP1A2*	0.160
rs2275620	*CYP2C8*	0.175
rs12333983	*CYP3A4*	0.185
rs4646437	*CYP3A4*	0.192

SNP = single nucleotide polymorphism, MAF = minor allele frequency

Table 3.5a Genotype associations with carbamazepine dose continued.

SNP ID (rs)	Gene	*Uncorrected P*-value
rs15524	*CYP3A5*	0.209
rs2470890	*CYP1A2*	0.236
rs1934952	*CYP2C8*	0.241
rs2069525	*CYP1A2*	0.270
rs10050146	*UGT2B7*	0.289
rs1419745	*CYP3A5*	0.317
rs2740574	*CYP3A4*	0.321
rs2740168	*EPHX1*	0.346
rs11572172	*CYP2C8*	0.366
rs2260863	*EPHX1*	0.390
rs6976017	*CYP3A5*	0.419
rs28365095	*CYP3A5*	0.429
rs28365083	*CYP3A5*	0.434
rs6600894	*UGT2B7*	0.448
rs1051740	*EPHX1*	0.495
rs2292566	*EPHX1*	0.495
rs1877724	*EPHX1*	0.515
rs2234698	*EPHX1*	0.543
rs11572079	*CYP2C8*	0.623
rs28371764	*CYP3A5*	0.642
rs34143170	*EPHX1*	0.665
rs1058930	*CYP2C8*	0.834
rs2740170	*EPHX1*	0.841
rs17861157	*CYP1A2*	0.938
rs4149230	*EPHX1*	0.964

SNP = single nucleotide polymorphism, MAF = minor allele frequency

Table 3.5b SNP genotypes associated with carbamazepine dose prior to correction

Of the 51 SNP analysed only variants with *P* <0.05 (before correction for multiple testing) are shown. The reference sequence (rs) numbers for each variant, their location in the respective gene and individual allele information is also provided.

Gene	SNP ID (rs)	Location	Amino acid change	MAF	Un-corrected *P*-value	FDR *P*-value
CYP2C8	rs1934956	Intron	-	0.116	0.019	0.124
CYP3A4	rs2246709	Intron	-	0.269	0.026	0.145
CYP3A4	rs12721617	Intron	-	0.006	0.029	0.145
CYP3A5	rs4646450	Intron	-	0.182	0.006	0.088
Flanking EPHX1	rs2292558	Intron	-	0.095	0.007	0.088
EPHX1	rs4149229	Exon	P.K416K	0.006	0.007	0.104
EPHX1	rs2234922	Exon	P.H139R	0.163	0.046	0.214
UGT2B7	rs4356975	Intron	-	0.229	0.003	0.069
UGT2B7	rs3924194	Intron	-	0.167	0.012	0.069
UGT2B7	rs7439366	Exon	P.H268Y	0.399	0.010	0.104
UGT2B7	rs7375178	Intron	-	0.396	0.014	0.104

FDR = false discovery rate, MAF = minor allele frequency, SNP = single nucleotide polymorphism

Figure 3.3a

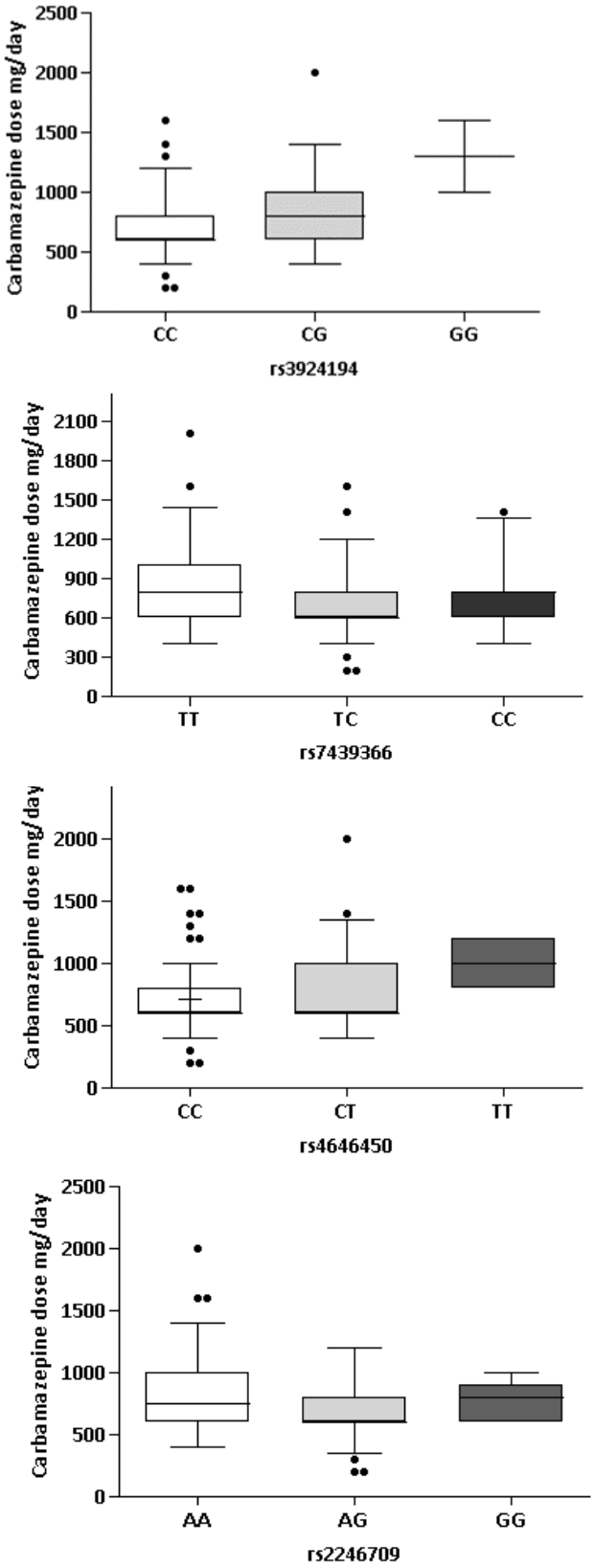

Carbamazepine dose mg/day
2500
2000
1500
1000
500
0
CC
CG
GG
rs3924194
Carbamazepine dose mg/day
2100
1800
1500
1200
900
600
300
0
TT
TC
CC
rs7439366
Carbamazepine dose mg/day
2000
1500
1000
500
0
CC
CT
TT
rs4646450
Carbamazepine dose mg/day
2500
2000
1500
1000
500
0
AA
AG
GG
rs2246709

Figure 3.3b

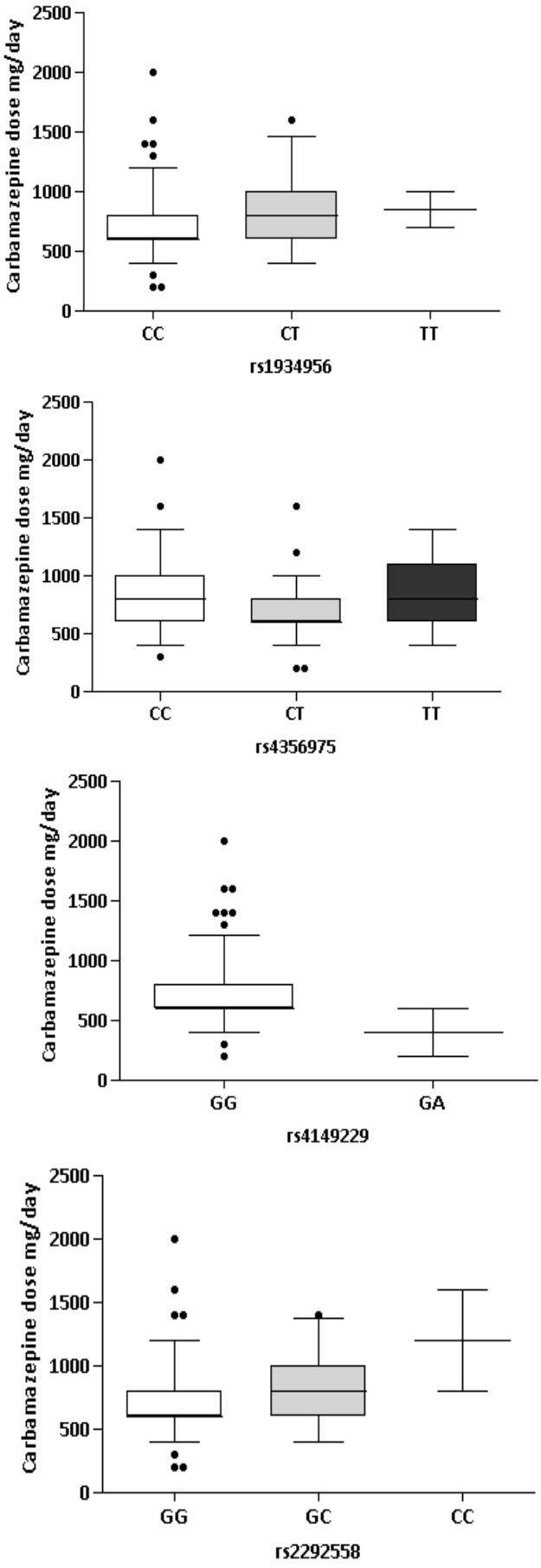

Carbamazepine dose mg/day
2500
2000
1500
1000
500
0
CC
CT
TT
rs1934956
Carbamazepine dose mg/day
2500
2000
1500
1000
500
0
CC
CT
TT
rs4356975
Carbamazepine dose mg/day
2500
2000
1500
1000
500
0
GG
GA
rs4149229
Carbamazepine dose mg/day
2500
2000
1500
1000
500
0
GG
GC
CC
rs2292558

Figure 3.4c

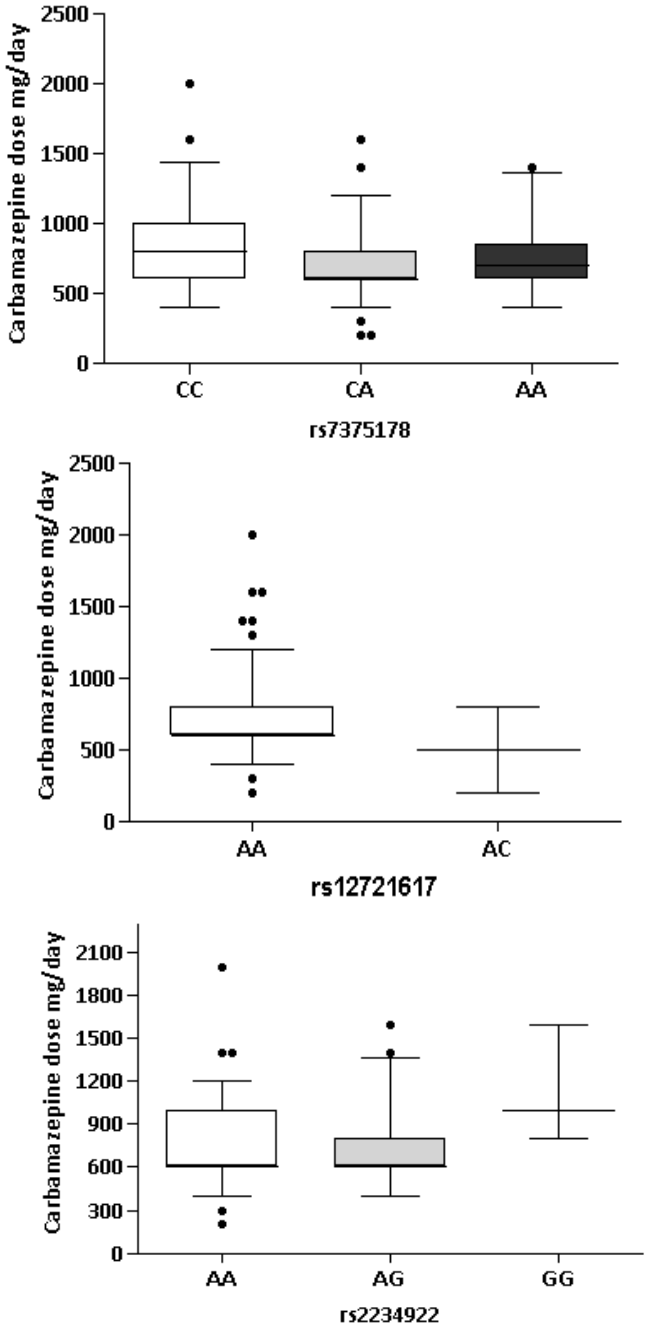


Figure 3.3 a, 3.3b, 3.3c Variant genotype and carbamazepine dose

Box and whisker plots of single nucleotide polymorphisms associated with carbamazepine maintenance dose (P<0.05, prior to correction for multiple testing). Dose is distributed according to individual genotype. Solid lines represent the median carbamazepine dose in each group, boxes represent the 25th and 75th percentile, whiskers represent 5th–95th percentiles, and dots represent outliers.

3.4.3 Gene haplotype identification and variability in dosing

Haplotypes within each of the six genes were next investigated to determine whether they explained a greater percentage of dose variability than single SNPs. In total eight distinct haplotype blocks were identified across the six DME genes. A single block spanned each of *UGT2B7*, *CYP1A2* and *CYP2C8*, two blocks overlapped *CYP3A4* and *CYP3A5*, and the remaining three blocks spanned *EPHX1* (Figure 3.4). Patients with a haplotype pair allocation probability <90% for each block were excluded from the analysis prior to performing a regression analysis (2 patients were excluded from Block 1, 2 from Block 2, 1 from Block 3, 1 from Block 4, 13 from Block 5, 2 from Block 6, 19 from Block 7 and 19 from Block 8). Results of the regression analysis for the PHASE generated haplotypes are presented in Table 3.6. Out of the eight identified haplotype blocks, two showed association with CBZ maintenance dose; Block 1 spanning *UGT2B7* (*P*= 0.023) and Block 5 overlapping both *CYP3A4* and *CYP3A5* (*P*= 0.011). Both blocks were only modestly associated with CBZ maintenance dose and failed to remain statistically significant following FDR analysis (Table 3.6). Individual haplotypes within each gene were therefore not examined.

3.4.4 Bioinformatics work

Bioinformatics analysis for the exploration of likely function of each of these 11 SNPs was performed despite their failure to remain significantly associated with CBZ dose after correction for multiple testing. Such investigations have the potential to identify subtle effects that may be lost in a candidate gene association analysis where statistical power is low and associations weakened by the need to correct for multiple comparisons. None of the 11 SNPs was predicted to have a significant influence on protein function and/or expression. Results for bioinformatics analyses can be found in Table 3.7. The SIFT and FastSNP webservers were used to evaluate the functional potential of the two non-synonymous variants, and predicted no effect of either polymorphism on protein function. FastSNP did, however, predict the presence of two 2 ESEs with the variant allele for both the *EPHX1* exonic SNPs rs2234922 and rs4149229 (non-synonymous and synonymous SNPs respectively). In addition to this, FastSNP predicted the loss of a TFBS for both the *UGT2B7* rs4356975 and *CYP3A4* rs12721617 intronic variants and the exonic *UGT2B7* rs7439366 variant (Table 3.7). These SNPs were also predicted to be located within a TFBS by TESS.

Figure 3.4a

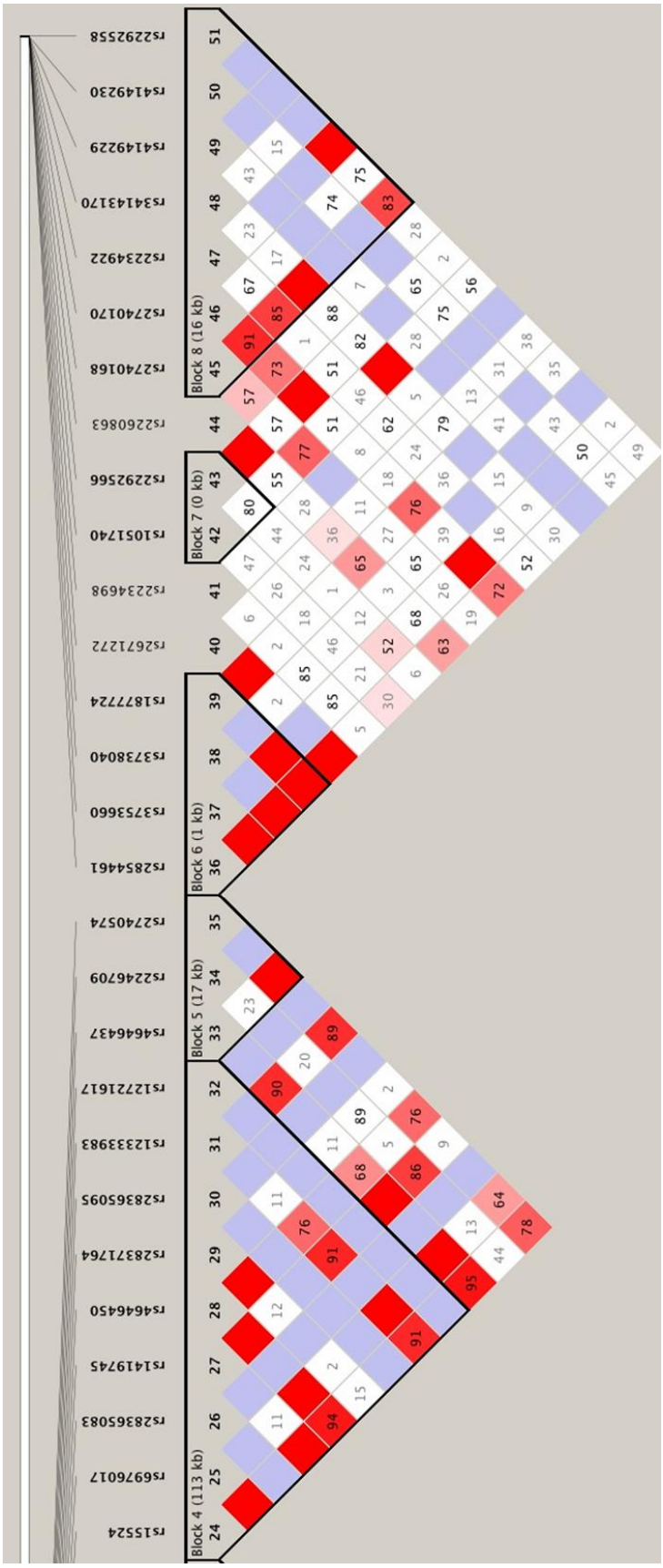

rs2292558
rs4149230
rs4149229
rs34143170
rs2234922
rs2740170
rs2740168
rs2260863
rs2292566
rs10517 40
rs2234698
rs26712272
rs18777 24
rs3738040
rs3753660
rs2854461
rs2740574
rs2246709
rs4646437
rs12721617
rs12333983
rs28365095
rs28371764
rs4646450
rs1419745
rs28365083
rs6976017
rs1552 4
Block 8 (16 kb)
Block 7 (0 kb)
Block 6 (1 kb)
Block 5 (17 kb)
Block 4 (113 kb)

Figure 3.4b

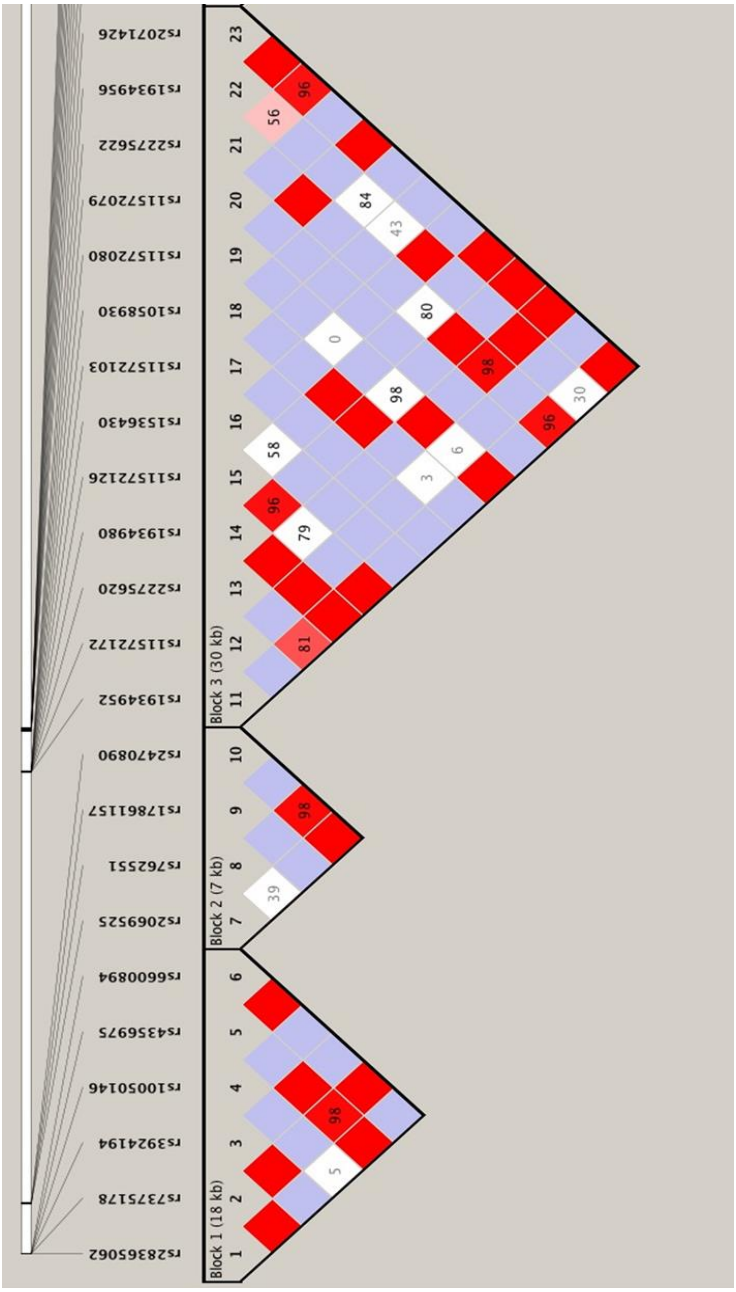


Figure 3.4 a and 3.4 b

Blocks of linkage disequilibrium and haplotypes identified across the six candidate drug metabolising enzyme genes. Linkage disequilibrium between each of the 51 SNPs across the six candidate genes that were included in the study, as visualised by Haploview v.4.2 (Barrett et al., 2005). Haplotype maps were generated using solid spine linkage disequilibrium method of block definition (Haploview v.4.2). A total of eight haplotype blocks were identified; (blocks 1-8 left to right) spanning the genes *UGT2B7*, *CYP1A2*, *CYP2C8* (Figure 3.4b), *CYP3A4* and *CYP3A5* and *EPHX1* (Figure 3.4a) respectively (3 across *EPHX1* and one each across the remaining 5 genes).

Table 3.6 Regression analysis of haplotype associations with carbamazepine dose

A chi-square test was used for testing for a statistical difference between a null and alternative regression model generated for each of the 8 blocks. A chi-square p-value of <0.05 after correction for multiple testing was considered statistically significant.

Haplotype block	Gene	Number of SNPs	Chi-squared value	Uncorrected *P*-value	FDR *P*-value
Block 1	*UGT2B7*	9	14.658	0.023	0.091
Block 2	*CYP1A2*	4	3.282	0.350	0.092
Block 3	*CYP2C8*	12	14.118	0.079	0.170
Block 4	*CYP3A4/ CYP3A5*	9	8.192	0.146	0.170
Block 5	*CYP3A4/ CYP3A5*	3	12.982	0.011	0.233
Block 6	*EPHX1*	4	8.192	0.085	0.416
Block 7	*EPHX1*	2	0.91	0.923	0.416
Block 8	*EPHX1*	7	7.652	0.364	0.923

FDR = false discovery rate, SNPs = single nucleotide polymorphisms

3.5 Discussion and summary

Despite the introduction of more than 12 new AEDs in the past two decades, drug therapy for epilepsy remains sub-optimal, with an estimated 50% of all treated patients experiencing ongoing seizure activity, significant AEs, or both. Newer AEDs have a more benign side effect profile than their established counterparts but none represents a significant advance in efficacy terms. As a result, there is a growing consensus in the epilepsy field that greater efforts should be directed at learning to use existing compounds in a more effective manner, rather than continually developing new agents of questionable additional benefit. Suggestions for the better use of existing AEDs include the investigation of rational polypharmacy and the individualisation of drug choice and dosing strategies through the identification and implementation of validated biomarkers. This study investigated the potential influence of polymorphic variants in DME genes on CBZ maintenance dose. Understanding an individual's dose requirement could help tailor titration schedules and target doses in order to minimise early withdrawals due to intolerable AEs and reduce the time to achievement of seizure control.

Hepatic metabolism represents the major elimination pathway for the majority of older AEDs and is the primary determinant of inter-individual variability in their PK. Many of these

compounds, including CBZ, have a relatively narrow therapeutic concentration range that renders them susceptible to clinically meaningful consequences of fluctuations in their serum levels. Multiple DMEs are involved in the biotransformation of CBZ (Kerr et al., 1994, Levy, 1995, Browne, 1998, Ketter et al., 1999, Huang et al., 2004, Staines et al., 2004, Ferraro and Buono, 2005, Klotz, 2007). CYP3A4 and CYP3A5 are the principle enzymes involved in the phase I metabolism of CBZ to its major and pharmacologically active metabolite CBZ-E, with CYP1A2 and CYP2C8 playing a more minor role. Phase II metabolism is predominantly mediated by mEH, which converts CBZ-E to CBZ-10,11-diol, while UGT2B7 is the major enzyme involved in glucuronide conjugation of the parent compound and its multiple metabolites. The genes encoding each of these enzymes harbour polymorphisms that are known to influence their catalytic function (Saruwatari, Ishitsu et al. 2010)(Ferraro and Buono, 2005).

This study employed a candidate gene approach using tSNPs plus putatively functional variants in an effort to identify genetic influences on CBZ maintenance dose across these six DMEs, with adjustment for known non-genetic influences on dose. Modest associations with dose were observed with eleven SNPs in five genes, four in *UGT2B7*, three in *EPHX1* region (one of which is located within the *TMEM63A;* a gene flanking *EPHX1*), two in *CYP3A4*, and one each in *CYP2C8* and *CYP3A5*, but none that survived correction for multiple testing. Reducing the dimension of the genetic data by haplotype analysis was similarly unsuccessful, with again only modest associations between CBZ maintenance dose and haplotype blocks spanning *UGT2B7* and *CYP3A4/CYP3A5* that failed to survive adjustment for multiple comparisons.

It is not possible to confidently implicate the predictive influence of genetic markers with drug dose without strong statistical evidence. This is however difficult when statistical power is limited by analysing a large number of variables in a limited number of patients. Although no SNP presented a strong correlation with maintenance dose in the independent SNP analysis, of significance is the potential influence of several of these SNPs on gene regulation as implicated by the bioinformatics analysis. Three SNPs were located in coding regions (1 *EPHX1* synonymous, 1 *EPHX1* non-synonymous and 1 *UGT2B7* non-synonymous SNP), but were not predicted to have a direct effect on the function of their respective enzymes. The predicted effect on splicing, TFB, and/or protein expression could however signify some importance of these coding SNPs. Alternatively the weak dose association with each of the 11 SNPs could be an indication that the variants are in LD with as yet unidentified genetic variants of stronger biological function. Alternatively, the single SNP results may point toward a potential role of the respective genes in CBZ response and/or dosing. The *EPHX1* non-synonymous rs2234922 variant that was originally associated with CBZ dose (Makmor-Bakry *et* al 2009) was additionally predicted to alter a TFBS in the bioinformatics analysis, thus again

implicating *EPHX1* with CBZ dosing. The other intronic SNP is located in the TMEM63A gene encoding a transmembrane protein close to *EPHX1*. Although no significant association has been reported for this gene with regards to epilepsy, or any other condition/ disease state, the rs2292558 variant has previously been associated with pulmonary arterial pressure in patients with chronic obstructive pulmonary disease (Castaldi et al., 2010). As such this SNP or gene could have some yet unidentified role in influencing *EPHX1* or epilepsy.

In single SNP analyses, true associations may be missed because of the incomplete information provided by individual variants (Hirschhorn et al., 2002). Multiple markers across chromosomal regions are thus increasingly being studied in combination, for the identification of relationships between genetic regions and traits of interest, with analysis based on haplotypes potentially more efficient than separate analyses of the individual SNPs (Judson et al., 2000). Of the eight haplotype blocks identified for the six genes included in this study, only one block spanning the gene *UGT2B7* and another spanning the gene *CYP3A4/CYP3A5* were modestly associated with CBZ dose. This finding partly supports the results of the single SNP analysis, where several *UGT2B7* SNPs appeared to associate with CBZ dose. Individually, however, these two haplotype blocks only explain a small amount of the variability present in CBZ dosing (r^2 values of 6.2% and 5.5% for the *UGT2B7* and *CYP3A* blocks respectively). This is similar to the variability accounted for by individual SNPs (r^2 values ranging from 3.0% to 7.1% for the 11 SNPs). The benefit of carrying out additional haplotype analysis was therefore questionable.

These findings were not entirely surprising, given that none of the genes in the panel are known to possess alleles of significant functional effect such as those observed in *CYP2D6* or *CYP2C9* (Ingelman-Sundberg, 2004b, Wilkinson, 2005). The *CYP3A5* gene does possess a null allele (*CYP3A5*3*)(Huang et al., 2004) that has been extensively studied with regard to altered drug metabolism (Hustert et al., 2001, Ingelman-Sundberg, 2004a). Although studies indicate association of this allele with CBZ serum concentrations, the role of CYP3A5 in CBZ metabolism remains controversial (Park et al., 2009, Saruwatari et al., 2010, Meng et al., 2011). It has also been speculated that the loss of CYP3A5 function is potentially compensated by enhanced metabolism mediated by CYP3A4 (Lee Sj Fau - Goldstein and Goldstein, Lamba et al., 2002, Huang et al., 2004).

Rather than seeking functional variants of large effect size, it was anticipated that this study might allow detection of multiple SNPs of small effect size that could be incorporated with non-genetic influences on CBZ dose into a predictive multivariate model. Those non-genetic factors that proved significant in this analysis included age, which would be expected to inversely correlate with dose requirement in an adult population (Bourdet et al., 2001), and source cohort, which was an interesting observation and one that perhaps reflects the differing methods of case ascertainment and drug use in randomised trials and routine clinical care. The

present investigation was however unable to detect a genetic influence on CBZ dosing when several individual SNPs were investigated, even after accounting for the contribution of non-genetic factors. This study therefore failed to support the hypothesis that common variants in the CBZ PK pathway may influence CBZ dosage in.

Two *EPHX1* SNPs, rs1051740 and rs2234922, were previously reported as influential in CBZ maintenance dosing (Makmor-Bakry et al., 2009). These variants were also typed as part of the present study and investigated in the overall genetic analysis and also in isolation in an effort to validate the original finding. The failure to validate the results of the study by Makmor-Bakry *et* al 2009 may be explained by the small number of patients (n=167) in the current analysis, combined with a large number of variables necessitating extensive correction for multiple testing. Power to detect modest associations was therefore limited. The original study was also disadvantaged by a small patient cohort (n=70) and only a weak association was found by the authors (P= 0.002 uncorrected). In addition to *EPHX1* variants, Makmor-Bakry *et al* 2009 also investigated single SNPs in each of *CYP1A2*, *CYP2C8*, *CYP3A4*, *CYP3A5*, and *UGT2B7*, selected on the basis of reported associations and potential functionality. They failed to consider wider variations across each gene region. This arguably limited their ability to detect associations, given that single SNPs are unlikely to explain complex traits. The more sensitive measure of using candidate gene tSNPs in the current study was however similarly unsuccessful, perhaps confirming that genetic variability in DMEs involved in CBZ metabolism does not play an important role in determining dose requirement.

It is possible that this study would have possessed greater sensitivity to detect genetic associations with CBZ PK had serum drug concentration data been available rather than dose data alone. TDM has proven useful for improving the effectiveness and safety of established AEDs, particularly for those with non-linear PK, such as PHT, or with considerable PK variability, as is the case with CBZ (Eadie, 1998, Anderson, 2008). Serum concentrations are more reflective of PK in general and less susceptible to non-genetic influences such as age, sex and body weight, all of which are compensated by dose differences. However, serum levels were not available for the cohorts in question. The use of a mixed-effect population PK approach has been shown to facilitate the delineation of relevant genetic factors, to estimate the magnitude of their effects on the PK variation, and to aid individualised dosing (Saruwatari, Ishitsu et al. 2010).

Using a candidate gene tSNP approach has advantages over the traditional single gene, single variant association method commonly found in PGx studies (Grant and Hakonarson, 2007, McCarthy et al., 2008) as it increases the likelihood of capturing putatively causative SNPs. The trade-off, however, is statistical power to detect associations in studies where a large number of genetic variants are typed in a relatively small cohort of patients. As studies move toward genome-wide analysis of complex traits, such as drug response, problems arise

in how to handle large genetic datasets (in terms of correction for multiple testing) whilst retaining sufficient statistical power. Larger and larger cohorts of patients are required but this may be unrealistic for some phenotypes. It was evident in this study that the lack of power limited the significance of the associations identified, for both single SNP and haplotype analyses. Unfortunately, the study was constrained by the availability of patients who met the inclusion criteria. Even in two of the largest epilepsy pharmacogenetic cohorts worldwide, insufficient numbers of patients were available to allow an association with CBZ dose to withstand correction for the multiple testing.

The haplotype approach to identifying causative genetic factors for both disease association and drug response is relatively new, but the benefits of using gene-based haplotypes as genetic markers is becoming clear (Judson et al., 2000). Determination of haplotypes or combinations of SNPs that are in LD might offer more power to detect associations than simply measuring individual SNPs (Tabor et al., 2002). When the initial test of association with genotypes does not reach statistical significance, further exploration of haplotype-specific effects is thought to increase the chance that at least one significant association will be detected (Colhoun et al., 2003). The ability to preselect SNPs that tag common haplotypes might also increase the prior probability of association with a candidate gene (Johnson et al., 2001).

Unfortunately, in this study, while there were associations between CBZ dose and both single SNPs and individual haplotypes prior to correction for multiple testing, these were lost thereafter. Thus, using a haplotype-based approach did not improve the sensitivity to detect true associations. This may have been because there were no true associations or that statistical power to detect such associations was not sufficiently high. There are also major issues around the use of simple regression for haplotype analysis. These include haplotype uncertainty, when these are derived with computational methods of phase inference, and haplotype complexity, in which the power of haplotype analysis is reduced by the large number of haplotypes that need to be studied (Zhao et al., 2003). With FastPHASE software, an individual is assigned to different haplotype pairs with different probabilities (Scheet and Stephens, 2006) and although this study employed a high threshold probability of 90% for haplotype uncertainty, the problem still exists.

Methods have been developed to reduce the number of haplotypes considered in association studies. One such method divides the whole chromosomal region into smaller regions for analysis and this generally involves a sliding window which is placed on the candidate region, with evidence for association within each window assessed (Zhao et al., 2003). Using the sliding window, the number of haplotype patterns in each window may be significantly less than that in the whole region, so the regression analysis involves fewer parameters and thus should have better power if there is an association between haplotype and

a disease trait. In addition it is assumed that association near the true disease variants is stronger than that in other regions (Zhao et al., 2003). Another common approach is based on the assumption that an unknown mutation occurred at some point in the evolutionary history and became embedded within the historical structure represented by a tree (cladogram) relating different haplotypes, assuming that certain portions of the tree would display the phenotypic effect of the mutation while other portions would not (Zhao et al., 2003). This second approach groups haplotypes into a smaller number before association analysis. Thus, the cladogram defines a nested analysis of variance that simultaneously detects phenotypic effects and localises the effects within the cladogram (Zhao et al., 2003). These other ways of analysis were not considered here instead, a simplistic approach of using haplotype block structures was employed. This simplistic method is helpful in association analysis using block-specific haplotypes (Daly and Day, 2001, Zhao et al., 2003) but there is an argument that the results depend on the definition of haplotype blocks. This method may also not be efficient if there is substantial LD among alleles in different blocks (Gabriel et al., 2002).

Table 3.7 Predicted function for 11 SNPs from FastSNP and SIFT

SNP ID (rs)	Gene	Predicted functional effect FastSNP	Risk	Predicted functional effect SIFT	TF site change	SE/SS change
rs1934956	*CYP2C8*	no known function	0-0	-	-	-
rs2246709	*CYP3A4*	no known function	0-0	-	-	-
rs12721617	*CYP3A4*	enhancer	1-2	-	yes	-
rs4646450	*CYP3A5*	no known function		-	-	-
rs2292558	*TMEM63A /EPHX1*	no known function	0-0	-	-	-
rs4149229	*EPHX1*	Benign	2-3	Tolerated	-	yes
rs2234922	*EPHX1*	splicing regulation	2-3	Tolerated	-	yes
rs4356975	*UGT2B7*	enhancer	1-2	-	yes	-
rs3924194	*UGT2B7*	no known function	0-0	-	-	-
rs7439366	*UGT2B7*	Benign and missense; splicing regulation	2-3	Tolerated	yes	-
rs7375178	*UGT2B7*	no known function	0-0	-	-	-

TF= transcription factor, SE= splicing enhancer site, SS=splicing silencer site.
Risk = Upper and lower risk of functional effect; 0= no effect 1=very low risk, 2=low risk, 3=medium risk, 4=high risk, 5=very high risk (http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp)(http://sift.bii.a-star.edu.sg/)

Drug response is a recognised complex, multifactorial phenotype, likely to involve several classes of genes of potential influence. This study investigated the importance of DMEs in determining the maintenance dose requirement of CBZ by examining numerous SNPs across several candidate genes. While there was evidence of a relationship between common genetic variation and dose, the associations identified were modest and did not survive correction for multiple testing. There are an increasing number of reports showing the importance of UGT enzymes, including UGT2B7 and UGT1A4, in AED PK and PD (Blanca Sanchez et al., 2010, Saruwatari et al., 2010), which would suggest that the results of this analysis have some merit. The lack of statistically significant associations thus does not rule out the possibility that associations may exist with other SNPs in the same genes, not least because the study design was informed by known genetic variation at the time of conception. Variation in DME genes is known to influence the PK and PD of drugs metabolised by CYP2D6, CYP2C9, and CYP2C19. In the case of CBZ, however, the principal DMEs are less polymorphic and likely to have a more subtle influence on inter-individual variability in drug dose, necessitating far larger studies to detect genetic associations. Given their modest contribution in this regard, it is debatable whether such studies are worthwhile or clinically informative.

CHAPTER FOUR

CONTRIBUTION OF A SINGLE FUNCTIONAL VARIANT IN THE SCN1A GENE TO OPTIMAL DOSING OF ANTIEPILEPTIC DRUGS

CONTENTS

4.1. Introduction

Ion channels are pore-forming proteins that regulate the movement of ions across cellular membranes and are therefore integral to a wide range of physiological pathways (Catterall, 1992). Na_V channels are responsible for the generation of action potentials in excitable cells and those expressed in the brain play a central role in the initiation and propagation of action potentials in neurones (Catterall, 1992, Yu and Catterall, 2003). Mutations in this fundamental channel unsurprisingly cause a number of disorders of membrane excitability, including several genetic epilepsies (Rogawski and Loscher, 2004, Meisler and Kearney, 2005).

4.1.1. Structure and function of the voltage-gated sodium channel

The Na_V channel protein consists of two distinct subunits, denoted α and β (Marban et al., 1998) (Figure 4.1). The α-subunit is a large, transmembrane protein composed of 4 homologous domains that are fundamental to channel function (Marban et al., 1998). These domains contain the voltage sensor and pore regions essential to channel gating (i.e. opening and closing of the channel) and ion selectivity, respectively (Clare et al., 2000, Meisler and Kearney, 2005). The four domains associate within the membrane to form a Na^+ permeable pore, through which Na^+ ions flow during propagation of an action potential (Meisler and Kearney, 2005) (Figure 4.2). Each α subunit is also associated with one or more accessory β subunits that are important for the modulation of the Na_V channel as a whole, regulating cell surface expression, voltage dependence and kinetics of the α subunit (Marban et al., 1998, Yu and Catterall, 2003).

Duplication of α-subunit genes during mammalian evolution has generated a number of genes encoding active Na_V channels that differ in tissue specificity and biophysical properties (Yu and Catterall, 2003, Meisler and Kearney, 2005). Ten Na_V α subunit genes (*SCN1A-SCN5A*, *SCN7A-SCN11A*) have been identified in mammals so far, nine of which are expressed in the nervous system (Table 4.1.) (Catterall et al., 2005, Leterrier et al., 2010). The four genes predominantly expressed in mammalian brain are *SCN1A*, *SCN2A*, *SCN3A* and *SCN8A*, which encode the channels $Na_V1.1$, $Na_V1.2$, $Na_V1.3$ and $Na_V1.6$, respectively. $Na_V1.3$ expression is mainly restricted to the early stages of development, while $Na_V1.1$ is the major Na_V channel in inhibitory interneurons and $Na_V1.2$ and $Na_V1.6$ are expressed in the axon initial segment of principal excitatory neurons.

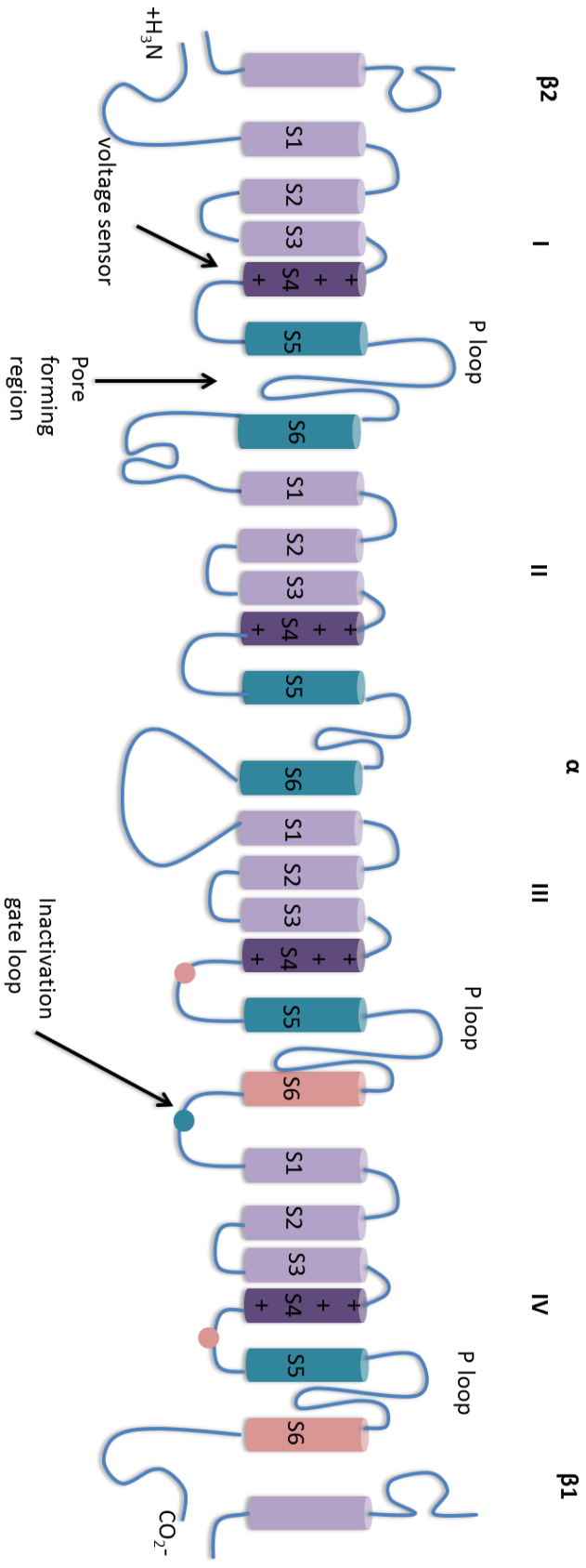


Figure 4.1 Structure of voltage-gated sodium channels Representation of the α-subunit and β1 and β2 subunits of the $Na_v1.2$ channel. The four domains of the α-subunit (I-IV) are indicated including its 6 helices or segments (S1-6). The S5 and S6 helices in each domain (shown in blue) are the pore-lining segments and the S4 helices (dark purple segments) make up the voltage sensors. Pink circles in the intracellular loops of domains III and IV indicate the sites implicated in forming the receptor for the inactivation gate and the blue circle indicates the inactivation gate loop. The pre-entrant loops in each domain (I-IV) form both the ion selectivity filter and outer pore mouth. S6 helices of domains III and IV (pink segments) are regions of modulatory drug binding, including sodium channel-blocking AEDs. Figure has been adapted from Rogawski and Loscher *et* al 2004.

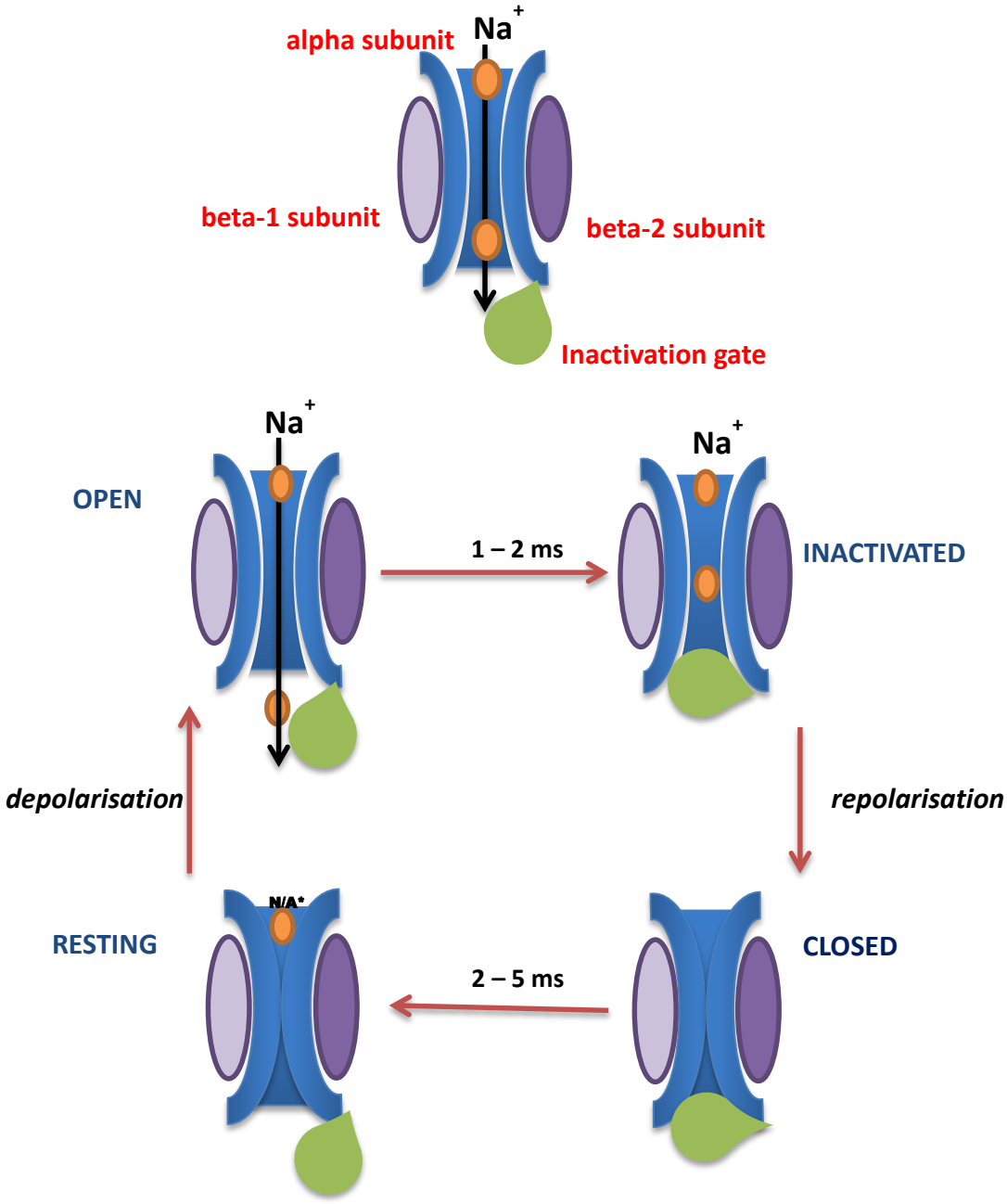


Figure 4.2 Voltage-gated sodium channel gating

Schematic representation of the different conformational states of a voltage-gated sodium channel**.** The voltage-gated sodium channel exists in four conformations, resting, activated (or open), inactivated and closed. The figure shows channel activation and sodium ion gating during the propagation of action potentials. The conformational change of the channel pore required for channel gating is also represented. Figure redrawn from Joseph *et* al 2011.

Table 4.1 Mammalian voltage-gated sodium channel subunits

Tissue distribution and genetic information for mammalian voltage-gated sodium channel alpha and beta subunits. Table adapted from Catterall *et* al 2005. Additional information extracted from the Online Mendelian Inheritance in Man (OMIM) website (www.omim.org).

Channel protein	**Subunit name**	**Gene**	**Tissue distribution**
Alpha			
$Na_V1.1$	Brain type I	*SCN1A*	CNS + PNS + heart
$Na_V1.2$	Brain type II	*SCN2A*	CNS
$Na_V1.3$	Brain type III	*SCN3A*	CNS + heart
$Na_V1.4$	Skeletal muscle	*SCN4A*	skeletal muscle
$Na_V1.5$	Cardiac	*SCN5A*	Heart + minor CNS expression
$Na_V1.6$	Brain type IV	*SCN8A*	CNS + PNS + heart
$Na_V1.7$	PN1	*SCN9A*	PNS
$Na_V1.8$	SNS	*SCN10A*	PNS
$Na_V1.9$	SNS2	*SCN11A*	PNS
Na_X	Atypical heart/glial	*SCN6A/7A*	Heart + uterus + smooth muscle + minor CNS expression
Beta			
$Na_V\beta1$	Beta-1	*SCN1B*	CNS + PNS + skeletal muscle + heart
$Na_V\beta2$	Beta-2	*SCN2B*	CNS+ PNS+ adrenal gland+ kidney
$Na_V\beta3$	Beta-3	*SCN3B*	CNS + PNS+ heart,
$Na_V\beta4$	Beta-4	*SCN4B*	CNS + PNS+ heart, skeletal muscle

Na_V = voltage-gated sodium channel, CNS = central nervous system, PNS = peripheral nervous system

The essential nature of the Na_V channel is further emphasised by the existence of inherited disorders (sodium "channelopathies") caused by mutations in genes that encode these vital proteins (Table 4.2) (George, 2005, Kass, 2005). Many mutations of the neuronal Na_V genes have been described in patients with epilepsy (George, 2005, Kass, 2005). The first of these was identified in the *SCN1B* gene (Escayg et al., 2000, George, 2005, Kass, 2005). Genetic defects in *SCN1A*, *SCN2A*, *SCN3A*, *SCN9A* genes have since been discovered that are responsible for several clinically overlapping epilepsy syndromes, namely generalised epilepsy with febrile seizure plus (GEFS+), severe myoclonic epilepsy of infancy (SMEI), and benign familial neonatal-infantile seizures (BFNIS) (Meisler et al., 2001, Steinlein, 2004, Meisler and Kearney, 2005). The majority of the Na_V channel mutations related to epilepsy can be found in *SCN1A* (Lossin et al., 2003, Mulley et al., 2003, Lossin, 2009, Meisler et al., 2010). Over 700 mutations of the *SCN1A* gene have been identified that cause a range of infantile epileptic encephalopathies with varying phenotypic severities, making this the most commonly mutated gene in human epilepsy (Lossin, 2009, Meisler et al., 2010). A small number of mutations have been identified in the other three principal, brain-expressed α subunit genes and only a handful are known for *SCN1B* (Lossin, 2009, Meisler et al., 2010).

4.1.2. Function of the α–subunit as a antiepileptic drug target

The *SCN1A* encoded $Na_V1.1$ protein functions as a major molecular target for numerous clinically important AEDs (Rogawski and Porter, 1990, Ragsdale and Avoli, 1998). Most AEDs have multiple cellular targets, however the majority of widely used AEDs have shown at least some Na_V blocking activity (Kwan et al., 2001). AEDs with Na_V channel blocking properties include PHT, LTG, CBZ, OXC, ZNS, FBM, TPM and VPA (Rogawski and Porter, 1990, Kwan et al., 2001). These bind to the Na_V channel and facilitate the selective inhibition of Na^+ currents (Macdonald and Kelly, 1995, Kwan et al., 2001). These currents are involved in the repetitive high-frequency spike firing of neurons, which is believed to occur during the spread of seizure activity in epilepsy (Rogawski and Loscher, 2004). AEDs have highest affinity for the Na_V channel protein in the inactivated state and their binding slows the otherwise rapid recycling process (Ragsdale and Avoli, 1998, Brodie and Sills, 2011). As a result, these drugs produce a voltage-and frequency-dependent reduction in channel conductance that limits repetitive neuronal firing with little effect on the generation of single action potentials (Ragsdale and Avoli, 1998, Brodie and Sills, 2011).

Table 4.2 Inherited disorders of voltage gated sodium channels

Muscle sodium channelopathies (SCN4A)
Hyperkalemic periodic paralysis
Paramyotoniacongenita
Potassium-aggravated myotonia
Painful congenital myotonia
Myasthenic syndrome
Hypokalemic periodic paralysis type 2
Malignant hyperthermia susceptibility
Cardiac sodium channelopathies (SCN5A)
Congenital long QT syndrome (Romano-Ward)
Idiopathic ventricular fibrillation (Brugada syndrome)
Isolated cardiac conduction system disease
Atrial standstill
Congenital sick sinus syndrome
Sudden infant death syndrome
Dilated cardiomyopathy, conduction disorder, arrhythmia
Brain sodium channelopathies (SCN1A, SCN2A, SCN1B)
Generalized epilepsy with febrile seizures plus
Severe myoclonic epilepsy of infancy (Dravet syndrome)
Intractable childhood epilepsy with frequent generalized tonic-clonic seizures
Benign familial neonatal-infantile seizures
Peripheral nerve sodium channelopathies (SCN9A)
Familial primary erythermalgia

(Table adapted from George 2005)

4.1.3. Binding sites of antiepileptic drugs

Site-directed mutagenesis experiments show that AEDs and local anesthetics bind to a common receptor site in the pore of the Na_V channel that is formed in part by three critical amino acid residues in transmembrane segment S6 in domains I, III and IV, with the IVS6 segment playing the dominant role (Catterall, 1999, Rogawski and Loscher, 2004, Catterall et al., 2005). Studies using PHT, CBZ and LTG have shown that these drugs contain a common motif (two phenyl groups separated by one or two C–C or C–N single bonds), which is thought to be crucial to this common binding (Figure 4.3) (Kuo, 1998, Rogawski and Loscher, 2004).

Evidence from mutational analysis has identified phenylalanine (F1764) and tyrosine (Y1771) residues in the S6 domain IV region as crucial for use-dependent block by both PHT and LTG (Ragsdale et al., 1996). These residues are brought into the pore during channel gating, thereby facilitating drug binding. Mutational analysis has also revealed that the pore-lining residues leucine 1465 and isoleucine 1469 in S6 of domain III of S6 may also form a portion of the high-affinity binding site for Na_V blocking AEDs (Figure 4.1)(Catterall, 2000, Rogawski and Loscher, 2004, Yarov-Yarovoy et al., 2012).

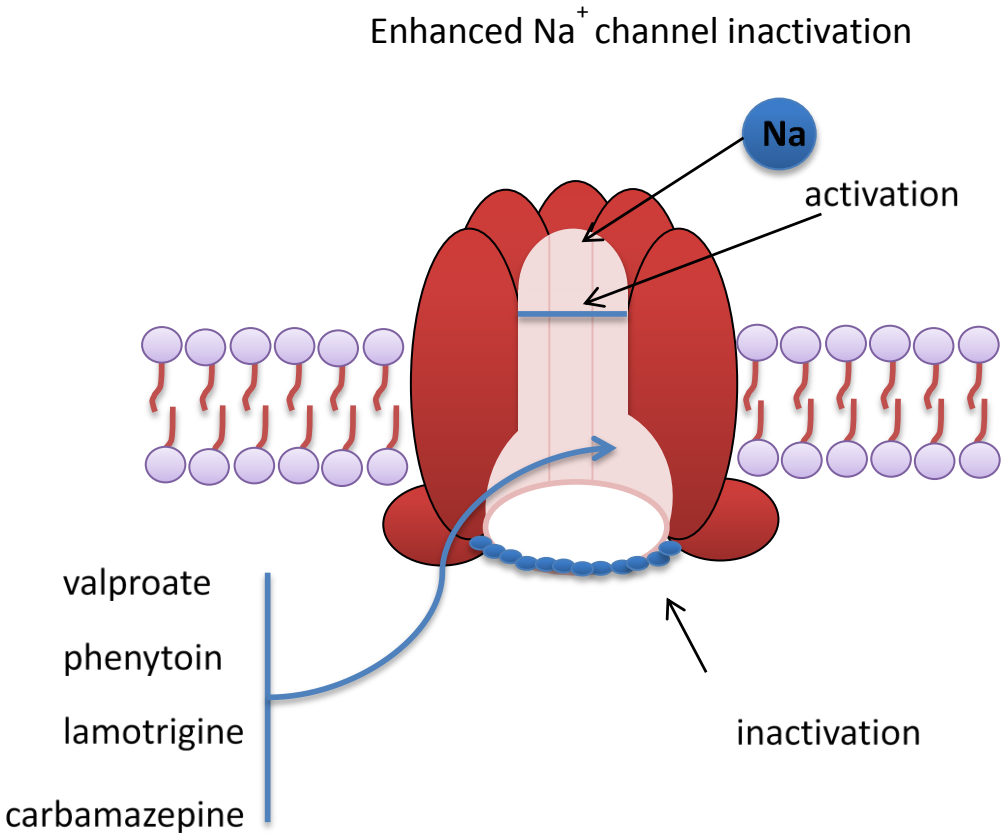


Figure 4.3 Molecular view of sodium channel and proposed site for drug binding
Experimental evidence has shown that antiepileptic drugs and local anesthetics with sodium channel blocking properties bind to receptor sites in the pore that is formed in part by amino acid residues in transmembrane segment S6 of domain III and IV of the channel (Catterall 1999; Ragsdale *et* al 1998).

4.1.4. Variable response to antiepileptic drug treatment

AED therapy typically comprises of a low starting dose, which is titrated upwards until an individual therapeutic dose whereby seizures discontinue or AEs become intolerable (Lo Monte et al., 2011). AEDs are known to have a relatively narrow therapeutic index and to be responsible for a wide variety of clinically important AEs (Perucca et al., 2001, Kwan and Brodie, 2004, Depondt and Shorvon, 2006, Ferraro et al., 2006, Schachter, 2007). Side effects are thus a major cause of medication intolerance and noncompliance, particularly within the first six months of therapy, and major AEs are reported to contribute to initial treatment failure in around 40% of patients taking established AEDs (Sander, 2004, Cavalleri et al., 2011).

4.1.5. Pharmacodynamic variation and antiepileptic drug response

Variability in AED dose requirement at an individual patient level can be broadly attributed to a combination of genetic and non-genetic factors (Cavalleri et al., 2011). Non-genetic influences include body mass index, gender and diet and are reasonably well characterised, although their clinical utility in terms of dose estimation is limited (Cavalleri et al., 2011). Some genetic polymorphisms that affect the PK of AEDs have also been identified and shown to influence the AED dose requirement (Tate et al., 2005, Klotz, 2007, Loscher et al., 2009, Park et al., 2009), particularly those in the CYP enzyme family (section 3.1).

Neuronal drug binding involving the Na_V channel is the first PD pathway to be directly associated with AED dosing (Tate et al., 2005, Kasperaviciute and Sisodiya, 2009). The *SCN1A* gene was originally implicated in AED response through early studies in Mendelian epilepsies. These demonstrated that *SCN1A* mutations can cause Dravet's syndrome or SMEI. SMEI patients are not only resistant to several AEDs but their seizures are typically aggravated following Na_V channel blocking AED treatment (Guerrini et al., 1998, Mulley et al., 2003, Yu et al., 2006, Abe et al., 2008).

4.1.6. *SCN1A* gene Isoforms

The *SCN1A* gene is 81-kb in size and is located on the long arm of chromosome two, situated at position 2q24.3. SCN1A is found as part of a cluster of voltage-gated sodium channel genes; namely *SCN2A*, *SCN3A*, *SCN7A* and *SCN9A* (encoding Nav1.2, Nav1.3, Nax, and Nav1.7, respectively) (Malo et al., 1994). The Nav1.1 protein (encoded by *SCN1A*) open-reading frame is organised into 26 exons and blueprints the instructions for a protein version incorporating between 1976 and 2009 amino acids. The variance in possible length is due to alternative splice junctions at the end of exon 11 that produce a full-length isoform or two shortened versions (Lossin et al., 2002). These differ by 33 bases and result in an 11 amino acid difference between the translated proteins. This splicing variability is the cause for the

inconsistencies in mutation reports across different research groups, as some are referring to full-length Nav1.1, while others reference Nav1.1[-33]; proposed to be more abundant in the brain (http://www.scn1a.info/Isoforms). A second alternative SCN1A splicing site or site of RNA processing variability can be found in exon 5. Here two mutually exclusive exons, 5N and 5A can be found and give rise to a postnatal and an adulthood isoform of the Nav1.1 channel protein (Copley, 2004). The amino-residue coding DNA sequence of these two alternative exons is nearly identical, differing only in three positions (Copley, 2004) (http: //www.scn1a. info/Isoforms).

4.1.7. Implication of the *SCN1A* α-subunit gene in AED response

A direct correlation between *SCN1A* and AED treatment was first reported in 2005 (Tate et al., 2005). The study, using 425 individuals with epilepsy demonstrated that the exon 5 *SCN1A* rs3812718 G>A SNP resulted in a significant shift in the maximal dosage of PHT and CBZ (Tate et al., 2005). Exon 5 of *SCN1A* encodes one of several voltage sensor regions of the Na_V channel (Ragsdale and Avoli, 1998, Tate et al., 2005). The voltage sensor region determines channel gating and so alteration in exon 5 expression can influence sensitivity of channels to blockade by AEDs (Tate et al., 2005, Heinzen et al., 2007). Two alternatively spliced versions of exon 5 are present in human genomic DNA, exon 5A (adult version) and alternative exon 5N (neonatal version), differing by three amino acids in their protein products (Tate et al., 2005).

This variant, located in the consensus sequence of the 5' splice donor site downstream of exon 5N (exon 5N+5G>A) (Figure 4.4) was suggested to alter the regular splicing of *SCN1A* in humans (Tate et al., 2005). The A allele was proposed to disrupt the consensus sequence of the 5N exon, reducing its expression and altering the normal 5N/5A ratio (Tate et al., 2005) (Figure 4.4). This was also demonstrated empirically by the study, which presented altered 5N/5A transcript levels in adult brain tissue from patients with epilepsy (Tate et al., 2005). The ancestral G allele is conserved across vertebrates (Zhang 1998) and is present in homologous CNS Na_V genes that are alternatively spliced within S3-S4 segments in domains I-IV. Maximum AED dosage was reported to consecutively decrease in epilepsy patients with AA, AG and GG genotypes (Tate et al., 2005) and this was suggested to be due to the level of 5N expression; with individuals expressing a greater percentage of 5N (those with two copies of the ancestral G allele) requiring lower drug doses.

4.1.8. Confirmation of the importance of the *SCN1A* polymorphism

Of the few PGx studies that have since investigated the rs3812718 polymorphism, three have managed to confirm the original association with AED dosing (Tate et al., 2006). As

maximum dose is not an accurate determinant of drug efficacy the original study was repeated with maintenance dose data, in a cohort of 71 Chinese patients by the original authors (Tate et al., 2006). The subsequent report confirmed that rs3812718 was also associated with PHT serum concentration at maintenance dose (Tate et al., 2006). However, a more recent study of the *SCN1A* polymorphism and CBZ dosing by Austrian researchers, who similarly used maintenance dose rather than maximum dose in their investigation, found no significant difference in mean CBZ dosing between the rs3812718 genotype groups (Zimprich et al., 2008).

4.1.9. Growing evidence for an influential role of the *SCN1A* gene

Since the first *SCN1A* findings there have been a number of interesting functional studies concerning this polymorphism, primarily exploring its consequences on Na_V channel activity and on response to AEDs (Heinzen et al., 2007, Thompson et al., 2011). Most recently an investigation comparing the biophysical properties of Na_V channels expressing 5A and 5N was performed. The study reported enhanced tonic block and use-dependent block of the 5N version of the Na_V channel by PHT and LTG, indicating increased sensitivity of channels expressing the 5N exon (Thompson et al., 2011). No differences were shown with CBZ (Thompson et al., 2011).

This recent study suggests that the dosing requirement of some AEDs may be altered by the *SCN1A* polymorphism, due to altered channel function, and suggests different pharmacological effects of the *SCN1A* rs3812718 variant dependent on AED. This finding may not only help to determine the true drug dose effect of this polymorphism but may explain the inconsistent findings surrounding this polymorphism to date. The results additionally appear to challenge the original conclusions made by Tate *et al* (Tate et al., 2005), suggesting that there is no effect of rs3812718 genotype on CBZ action (Thompson et al., 2011). This is the first direct evidence that variation in PHT dose requirement originally observed by Tate and colleagues (Tate et al., 2005) arises in part due to differences in how AEDs interact with the alternatively spliced $Na_V1.1$ channel, providing a mechanistic explanation for the association between this polymorphism and AED dosage. However, this data questions previous experimental evidence that suggests a common binding site of AEDs on the S6 helix of domain IV of the Na_V channel (Thompson et al., 2011), as implicates exon 5 (encoding the S3/S4 helices of domain 1) as the site of drug interaction.

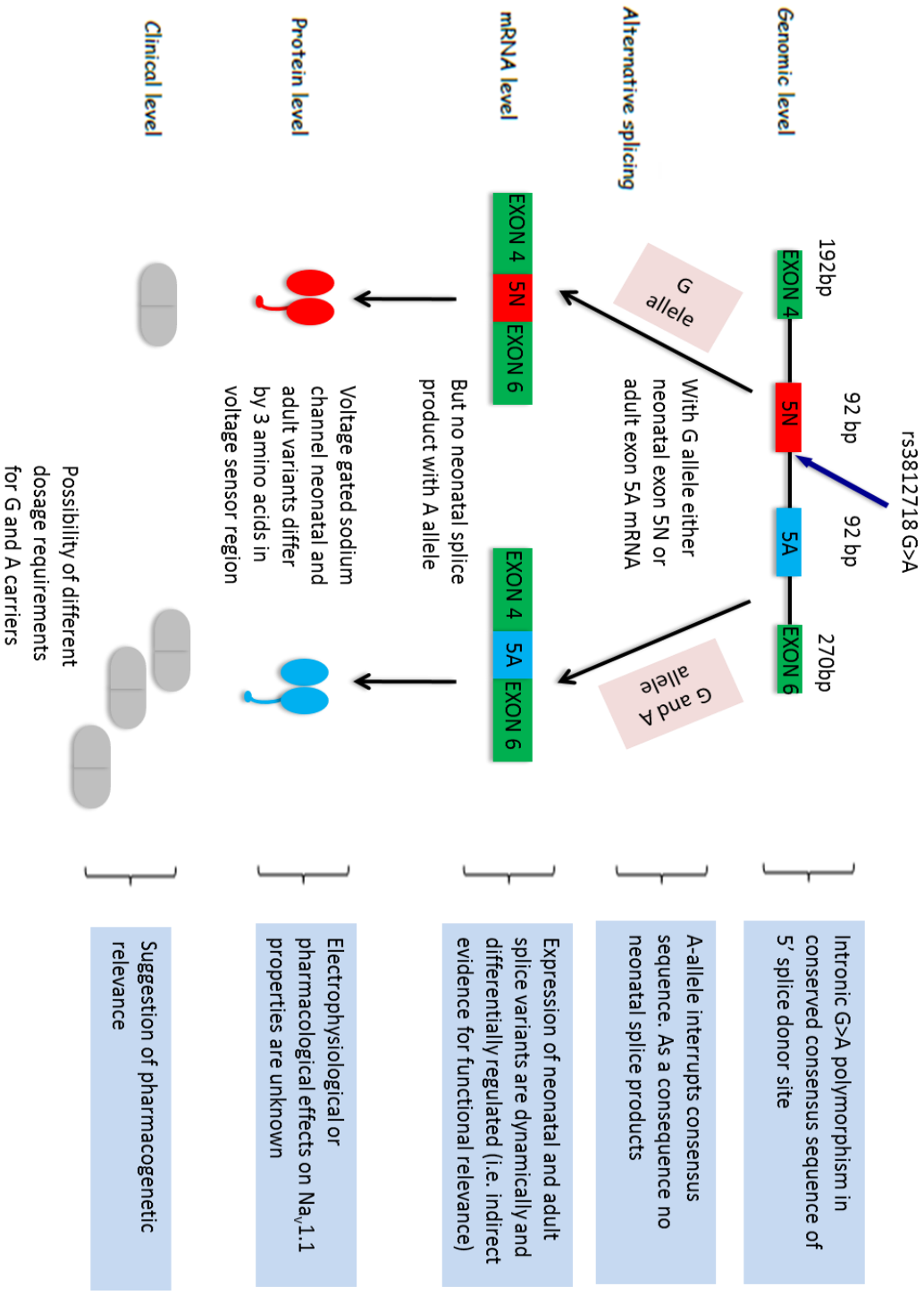


Figure 4.4 Alternative splicing of exon 5 and disruption of 5' splice donor site

Altered splicing of the *SCN1A* gene caused by the rs3812718 variant and how this can produce $Na_V1.1$ channels with different biophysical properties. This proposed mechanism may explain why patients might require different AED doses depending on their genotype. This figure has been redrawn from Loscher *et* al 2009.

The differences in drug block observed in this study were accordingly proposed to result from $Na_V1.1$ 5N/5A isoform-specific activation and/or inactivation gating (Thompson et al., 2011). 5N isoforms exhibited greater tonic and use-dependent inhibition by PHT and LTG, suggesting that binding sites for these drugs may be altered, and that the pharmacologic differences may arise from slower inactivation processes (Thompson et al., 2011).

4.1.10. Summary and research aims

The possibility of adjusting titration schedule based on genotype could lead to more rapid achievement of AED efficacy with adequate tolerability. Although the *SCN1A* genotype does not appear to have a striking influence on maximum PHT, LTG or CBZ doses, initial research results for this SNP suggest the potential for identifying patients who can tolerate higher therapeutic doses of these drugs. PGx data remains both limited and inconsistent for this drug target polymorphism. The growing interest in this gene, in conjunction with the functional evidence that has recently emerged, is however encouraging.

The primary aim of the following study was to examine the pharmacological consequence of the rs3812718 SNP by further investigating the association between rs3812718 SNP genotype and AED dose. A candidate SNP association study was performed to determine the effect of the rs3812718 polymorphism on both maximum and maintenance dose data of numerous AEDs, with a sub-group interaction analysis for individual drugs done in an attempt to validate the original genetic association with maximal dose of CBZ as reported by Tate et *al* 2005. In addition to the original investigation this present study involved maintenance dose data and several AEDs regardless of mechanism.

4.2. Methods

4.2.1. Patient cohort

The study included patients from the SANAD cohort (section 2.2.2), assuming they; i) were treated with AED monotherapy, ii) had information available for AED treatment history (i.e. AED exposure and corresponding dosage data), and iii) had an adequate amount of DNA sample available for SNP genotyping. Individuals who had received treatment with more than one drug during the study period (i.e. those for whom drugs were substituted because of inadequate seizure control and/or AEs) contributed more than once to the analysis.

4.2.2. Outcome and phenotype definitions for patient selection

The *SCN1A* rs3812718 G>A SNP was investigated for association with two outcomes, maximum dose and, where available, maintenance dose. This data was collected for each of

the drugs that individuals had been exposed to. Definitions used in this analysis were; i) seizure freedom, defined as at least 12 months of seizure freedom on an unchanged AED, ii) maximum dose, defined as the highest dose to which a patient had been exposed during treatment, and iii) maintenance dose, defined as the final AED dose that led to at least 12 months seizure freedom during monotherapy treatment. Additional data included in the analysis were patient age at recruitment, gender, and epilepsy type. Epilepsy type was broadly classified into three groups, IGE, LRE and UNC according to the clinical databases from which data was derived.

4.2.3. Data inclusion and extraction

Maximum dose data reflects the upper limit of drug tolerability and could be used to inform individual titration rate. In contrast, maintenance dose data is directly associated with treatment response and so is a better indication of treatment success at a particular dose (Tate et al., 2006). Data on maximum dose (mg/day) and, where available maintenance dose (mg/day), was extracted for each AED to which an individual had been exposed during follow-up. Maintenance doses were unavailable or disregarded for patients in whom there was no remission from seizures or who underwent dosage adjustment; without a single 12 month period without seizure freedom, or had other AEDs added, during the remission period. These patients were included in the maximum dose analysis only.

Many of the commonly used AEDs have multiple, overlapping mechanisms of actions and most possess at least some Na_V channel blocking activity. All AEDs were therefore included in the analysis, regardless of drug class or proposed mechanism of action. The majority of patients had been exposed to two or more AEDs during the course of follow up, usually due to AED switching because of AEs and/or lack of efficacy. Most patients therefore contributed more than one maximum dose to the analysis. In total, the patient population was exposed to nine different AEDs, those with known Na_V channel blocking activity being CBZ, LTG, OXC, PHT, VPA and those with another proposed primary mechanism of action being GBP, TPM, CLB and LEV (See Tables 4.4 and 4.6 for summary of the drugs included in the study). Data for CLB and PHT were excluded from the analysis of maximum dose, due to low numbers of patients taking these drugs ($n<20$).

4.2.4. Genotyping

The rs3812718 genotype for all 817 patients who met the initial inclusion criteria was determined using the Sequenom MassARRAY IPLEX platform in accordance with the manufacturer's instructions and as detailed in chapter 2. The Sequenom platform is designed for the analysis of multiple SNPs using a multiplex approach and would not ordinarily be employed in a single candidate SNP study. Single SNP studies would typically use TaqMan

allelic discrimination assays (Chapter 7) or other such systems. However, in this analysis the rs3812718 SNP was genotyped alongside a panel of five other SNPs (reported in Chapter 5) that were also being typed in the SANAD cohort and, as such, a multiplex approach was deemed most efficient. Thus, a 6 SNP plex was generated using the online Sequenom MassARRAY assay designer software (Chapter 2). Table 4.3 shows the Sequenom assay design output (https://mysequenom.com/Tools/genotyping/default.aspx) for rs38132718 alone *(Gabriel et al., 2009).*

Table 4.3 Primer sequence for Sequenom genotyping

Primer sequence for the rs3812718 SNP as designed by the Sequenom Assay design software

SNP ID (rs)	PCR primer sequence-forward	PCR primer sequence-reverse	Extension primer sequence
rs3812718	ACGTTGGATGACAAAGAGCCTATCCTTTAC	ACGTTGGATGACAAAGAGCCTATCCTTTAC	CCTATCCTTTACTCTAATCACTT

SNP = single nucleotide polymorphism, PCR = polymerase chain reaction

4.2.5. Statistical analysis

All analyses of association were carried out using SPSS statistical software version 16.0 (SPSS Inc., Chicago, IL, USA). In total, three statistical analyses were performed; a univariate analysis to identify any non-genetic confounders that may influence the genetic association analysis, a regression analysis for the identification of genetic sub-groups that differ in maximum or maintenance dosage requirement, and a validation analysis to attempt to replicate the findings of the original study (Tate, Depondt et al. 2005). An additive mode of inheritance was assumed for all genetic analyses, in line with previous reports of this polymorphism (Tate et al., 2005, Tate et al., 2006).

4.2.6. Data preparation

Since individual doses for different AEDs are not equivalent, it was necessary to normalise the dose data prior to analysis. For this purpose, the defined daily dose (DDD) was referred to, which is the average maintenance daily dose in adults as defined by the World Health Organisation (WHO) (http://www.whocc.no/atcddd/, accessed September 22, 2010)

(Table 4.4). Normalised doses were expressed as prescribed daily dose (PDD)/defined daily dose (DDD) ratios. All dose ratios were then transformed using the natural log function to achieve a normal distribution and this final log-transformed ratio was used for all statistical analysis. Prior to the genetic association analyses, the rs3812718 SNP was tested for deviation from Hardy Weinberg equilibrium (HWE) using Haploview software version 4.1 (Barrett et al., 2005), with a p-value of <0.001 indicating deviation.

Table 4.4 Defined and prescribed daily doses

The World Health Organisation (WHO) defined daily dose (DDD) and the range of maximum and maintenance doses (prescribed daily doses; PDD) for each drug.

AED	WHO DDD (mg)	Cohort Maximum PDD range (mg/day)	Cohort Maintenance PDD range (mg/day)
Carbamazepine	1000	200-2800	100-1400
Gabapentin	1800	600-4800	300-3600
Lamotrigine	300	25-675	50-675
Oxcarbazepine	1000	150-3000	450-1500
Topiramate	300	25-800	37.5-400
Valproate	1500	100-3000	200-1500
Levetiracetam	1500	100-3000	-

AED = antiepileptic drug, WHO = World Health Organisation, DDD = defined daily dose, PDD= prescribed daily dose

4.2.7. Univariate analysis of association with non-genetic factors

To evaluate the individual effect of the SNP, univariate tests of association were conducted with the non-genetic factors alone. Age, gender and epilepsy type were tested for association with the two dose variables (maximum and maintenance) in turn. Univariate linear regression, the independent samples t-test and ANOVA were used for each of the factors respectively. Non-genetic factors found to be significant (P<0.05) were adjusted for in the association analysis with *SCN1A* SNP genotype.

4.2.8. Regression analyses for association with rs3812718

The aim of the regression analysis was to investigate whether rs3812718 genotype associated with AED dose requirement. For each dose variable (maximum and maintenance), two regression models were built and compared using the LRT. The first model, the 'baseline (or non-genetic) model' included demographic and clinical factors found to be significant in the univariate analysis. Dose data for all drugs were included in the regression model as covariates. The second model, the 'genetic model', was the same as the first but also included a covariate representing the SNP.

The nature of the SANAD cohort data used in this study meant that some patients contributed more than one observation to the analysis of maximum dose. Mixed-effect models include additional random-effect terms and are often appropriate for representing clustered, dependent data that are either; collected hierarchically, when observations are taken on related individuals (such as siblings), or when data are gathered over time on the same individuals. For the maximum dose variable, linear mixed models (SPSS-Generalised Linear Model function) were thus fitted to capture the hierarchical structure of the data (Everitt and Howell, 2005) (http://www.wiley.com/legacy/wileychi/eosbs/pdfs/bsa251.pdf).

An additional regression analysis was performed to test for possible drug specific genetic influence or drug-gene-interaction. If the rs3812718 genotype was found to be significantly associated with maintenance and/or maximum dose from the linear regression analysis described above, another regression model was fitted that, in addition to SNP genotype and non-genetic factors, included interaction terms for each drug as additional covariates. This third model was termed the interaction model and was compared to the genetic model, with the LRT again employed for testing association and a chi-square distribution p-value of <0.05 again used to indicate a statistically significant difference.

4.3. Results

The rs3812718 polymorphism did not deviate from HWE ($P>0.001$). Of the 817 individuals for whom DNA was available, 637 had sufficient clinical data to be considered for genotyping (28 individuals had no follow up data whatsoever, 138 took two or more drugs in combination throughout treatment, 5 had missing dose data, and 9 had some ambiguity in either drug administration or dosing). A further 51 patients failed genotyping for the rs3812718 polymorphism. The remaining 586 individuals were treated with AED monotherapy and had dose and genotype data for inclusion in the data analysis. Patient characteristics are summarised in Table 4.5. Frequencies for the SCN1A rs3812718 G>A polymorphism in these 586 individuals were 28%, 54% and 18% for the GG, GA and AA

genotypes, respectively. Figures 4.5A and 4.5B show the distribution of standardised maintenance and maximum dose ratios.

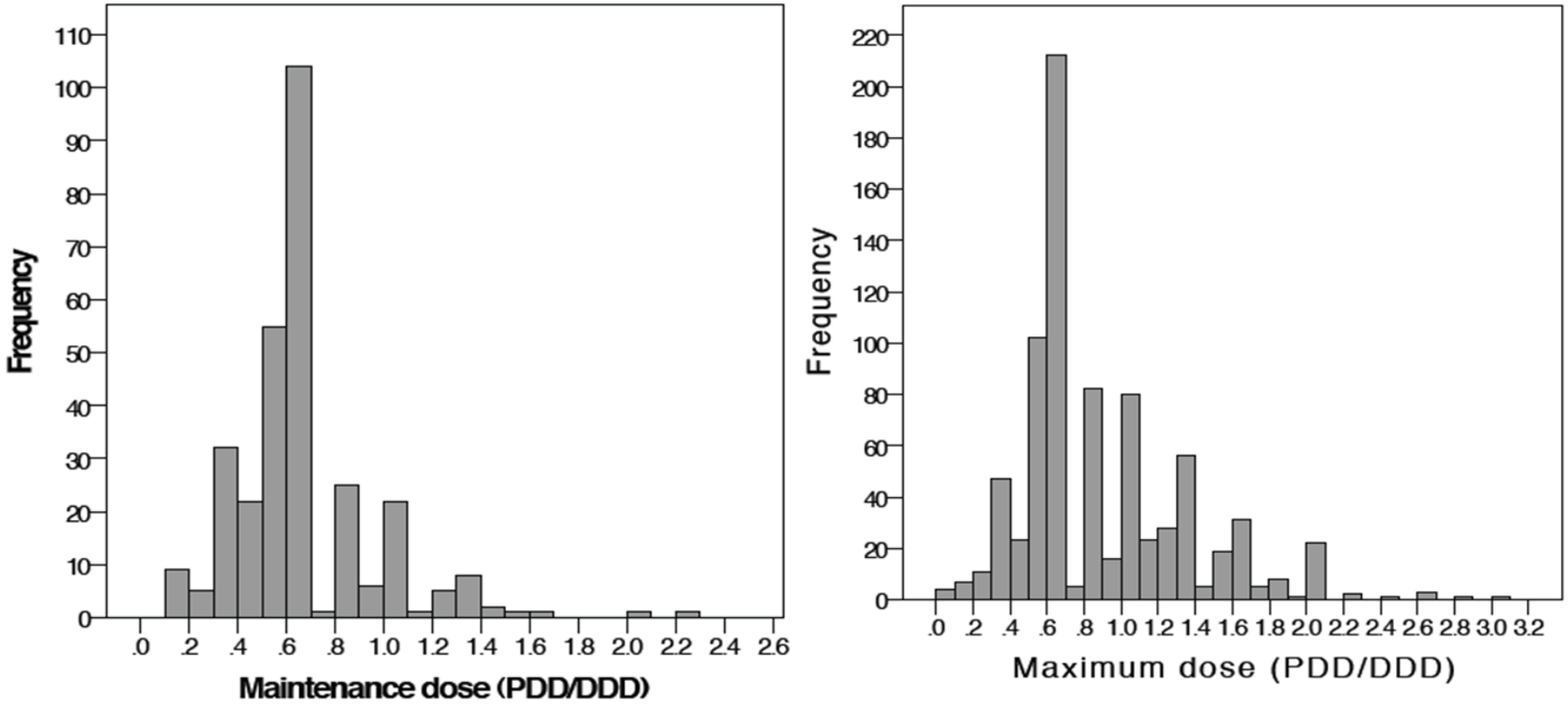
Table 4.5 Characteristics of the patient population

		Cohort n=586
Age (years)	Minimum	5
	Mean (± SEM)	39.26 (± 0.76)
	Maximum	84
Gender (n)	Male	322
	Female	264
Epilepsy type (n)	IGE	95
	LRE	399
	UNC	92
Maximum dose (PDD/DDD ratio)	Minimum	0.1
	Mean	0.87
	Maximum	3
Maintenance dose (PDD/DDD ratio)	Minimum	0.1
	Mean	0.64
	Maximum	2.25

IGE = idiopathic generalised epilepsy, LRE = localisation-related epilepsy, UNC = unclassified epilepsy, PDD = prescribed daily dose, DDD= defined daily dose, SEM = standard error of the mean

Figure 4.5 Distribution graphs for maintenance and maximum dose ratios

Distribution plots for maximum and maintenance dose ratios among the 586 patients included in the analysis. All data are expressed as prescribed daily dose/defined daily dose (PDD/DDD) ratios. A total of 301 patient maintenance dose (A) and 795 patient maximum dose data was available for analysis (B).



4.3.1. Results of the maintenance dose analysis

Maintenance dose data was available for 301 of the 586 patients (Table 4.6 and Figure 4.5A). Tables 4.7 and 4.8 show results of the univariate analysis for non-genetic investigation and regression analyses. Maintenance dose ratio showed no statistically significant association with any of the non-genetic variables (P>0.05). Similarly no significant association was identified with rs3812718 genotype in the regression analysis (P=0.324, Figure 4.6). Distribution of dose ratios for all six AEDs included in the analysis are displayed in Figure 4.7.

Table 4.6 Dose data for each drug included in the study

The median and range of maximum and maintenance doses for each drug included in the analysis, plus the total number of maximum and maintenance doses for each drug.

	Maximum Dose data (mg)	**Maintenance dose data (mg)**	**Total maintenance doses (n)**	**Total maximum doses (n)**
Carbamazepine				
Median	800	600	53	168
Range	2600	1300		
Gabapentin				
Median	1800	1200	30	96
range	4200	3300		
Lamotrigine				
Median	200	150		
Range	650	625	87	211
Oxcarbazepine				
Median	1200	900		
Range	150	1050	21	57
Topiramate				
Median	150	150	66	143
Range	775	363		
Valproate				
Median	1000	1000		
Range	2900	1300	44	97
Levetiracetam				
Median	1500	-		
Range	2900	-	-	23

Table 4.7 Univariate analysis of non-genetic factors and maintenance dose ratio

	Analysis	**F-statistic**	***P*-value**
Age	Continuous	0.77	0.73
Gender	Categorical (male / female)	1.02	0.31
Epilepsy type	Categorical (IGE, LRE, UNC)	1.02	0.36

IGE = idiopathic generalised epilepsy, LRE = localisation-related epilepsy, UNC = unclassified epilepsy

Table 4.8. Contribution of clinical and genetic variables to maintenance dose

Covariate	**Parameter**	**t-statistic**	**Parameter *P*-value**
Non-genetic model	CBZ	0.03	0.73
Individual drugs only	OXC	0.40	0.00
	TPM	-0.11	0.12
	VPA	0.02	0.76
	GBP	0.25	0.01
	LTG	0a	.
Genetic model	CBZ	-0.25	0.03
Individual drugs plus SNP	OXC	-0.22	0.41
	TPM	0.17	-0.10
	VPA	-0.35	0.03
	GBP	-0.21	0.25
	LTG	0.00	0a
	rs3823728	-0.04	-0.04
LRT *P*-value			0.32

LRT = Likelihood Ratio Test, 0a = parameter set to 0 by SPSS as redundant in model

Figure 4.6 ***SCN1A*** **rs3812718 SNP genotype association with maintenance dose ratio**

Box and whisker plots of *SCN1A* rs3812718 SNP genotype association with maintenance dose ratio (prescribed daily dose/defined daily dose (PDD/DDD)) for n=301 individuals. Boxes represent the 25th and 75th percentiles of dose ratio, solid lines represent the median dose ratio, and whiskers represent minimum and maximum dose ratio in each genotype group.

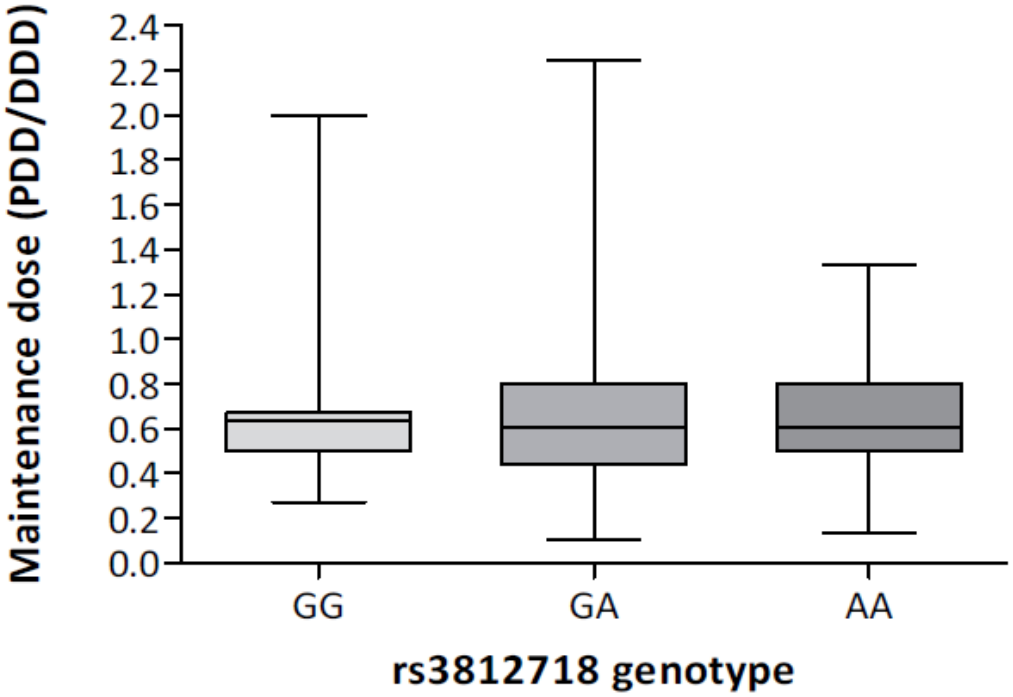
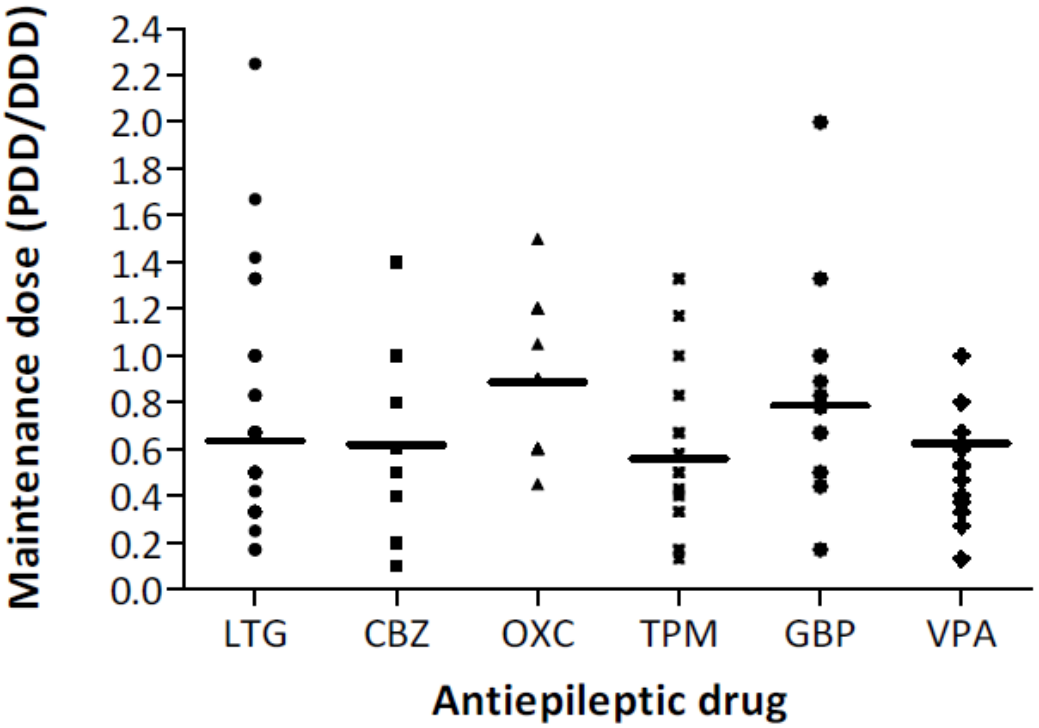


Figure 4.7 Distribution of maximum dose ratio for each drug

Aligned dot plot of maintenance dose ratio (prescribed daily dose/defined daily dose (PDD/DDD)) distribution for each antiepileptic drug included in the analysis. Lines represent the mean dose ratio and dots represent dose ratios for individual patients (n=301). LTG = lamotrigine; CBZ = carbamazepine; OXC = oxcarbazepine; TPM = topiramate; GBP = gabapentin; VPA = valproate



4.3.2. Results of the maximum dose analyses

Multiple maximum dose data was available from some individuals and so in total 795 different maximum doses for n=586 individuals were included for analysis (168 CBZ, 96 GBP, 211 LTG, 23 LEV, 57 OXC, 143 TPM, 97 VPA) (Table 4.6). Univariate analysis for testing association of maximum dose ratio with non-genetic factors is shown in Table 4.9. Epilepsy type was found to be associated with maximum dose ratio (P=0.044) (Table 4.9). A higher average maximum dose PDD/DDD ratio was evident in individuals with LRE when compared to those with UNC. No associations were found with either age or gender.

The results of the regression analysis for standardised maximum dose for each drug are presented in Table 4.10. When the non-genetic model and model including epilepsy type and SNP were compared using the LRT, a significant association was found with maximum dose ratio (P= 0.022; Table 4.10, Figure 4.8). Figure 4.9 shows the distribution of maximum dose ratios for all six AEDs included in the analysis.

Table 4.9 Non-genetic association with maximum dose ratio

	Analysis	**F-Statistic**	***P*-value**
Age	Continuous	0.990	0.54
Gender	Categorical (male / female)	1.802	0.11
Epilepsy type	Categorical (IGE, LRE, UNC)	3.141	0.04

IGE = idiopathic generalised epilepsy, LRE = localisation-related epilepsy, UNC = unclassified epilepsy

Table 4.10 Contribution of clinical and genetic variables to maximum dose ratio

Covariate	Parameter	Regression coefficient	Parameter *P*-value
Individual drugs and clinical variables	IGE	0.06	0.06
	LRE	0.03	0.25
	UNC	0a	.
	CBZ	-0.01	0.61
	OXC	0.14	0.00
	TPM	-0.09	0.00
	VPA	-0.03	0.23
	GBP	0.12	0.00
	LEV	0.12	0.02
	LTG	0a	.
Genetic model Individual drugs plus clinical covariates plus SNP	IGE	0.06	0.06
	LRE	0.03	0.25
	UNC	0a	.
	CBZ	-0.01	0.63
	OXC	0.14	0.00
	TPM	-0.09	0.00
	VPA	-0.04	0.21
	GBP	0.12	0.00
	LEV	0.12	0.02
	LTG	0a	.
	rs3823728	0.02	0.17
LRT *P*-value			0.02

IGE = idiopathic generalised epilepsy, LRE = localisation-related epilepsy, UNC = unclassified epilepsy, LRT = Likelihood Ratio Test, 0a = parameter set to 0 by SPSS as redundant in model

4.3.3. Results of drug interaction analysis for maximum dose ratio

Since a statistically significant association was identified with maximum dose ratio in the initial genetic analysis, this suggested a potential contribution of the *SCN1A* SNP to AED maximum dose requirement and an interaction analysis was subsequently performed to investigate whether AED influenced the effect of SNP genotype. Table 4.11 presents the results of LRT for the genetic model comparisons. This regression analysis compared the genetic regression model described above with a model additionally containing interaction terms for each drug. This also showed a statistically significant association with maximum AED dose ratio ($P = 6.46 \times 10^{-4}$; Table 4.11). Box plots for each AED dose ratio association with rs3812718 genotype are presented in Figure 4.11a and b. Table 4.12 presents the mean maximum dose ratio for each genotype group also stratified by individual AED.

Figure 4.8 ***SCN1A*** **rs3812718 SNP genotype association with maximum dose ratio.**

A total of 795 maximum doses were included in the analysis. Boxes represent the 25th and 75th percentiles of maximum dose ratio (prescribed daily dose/defined daily dose (PDD/DDD)), solid lines represent the median dose ratio and whiskers represent minimum and maximum dose ratio in each genotype group.

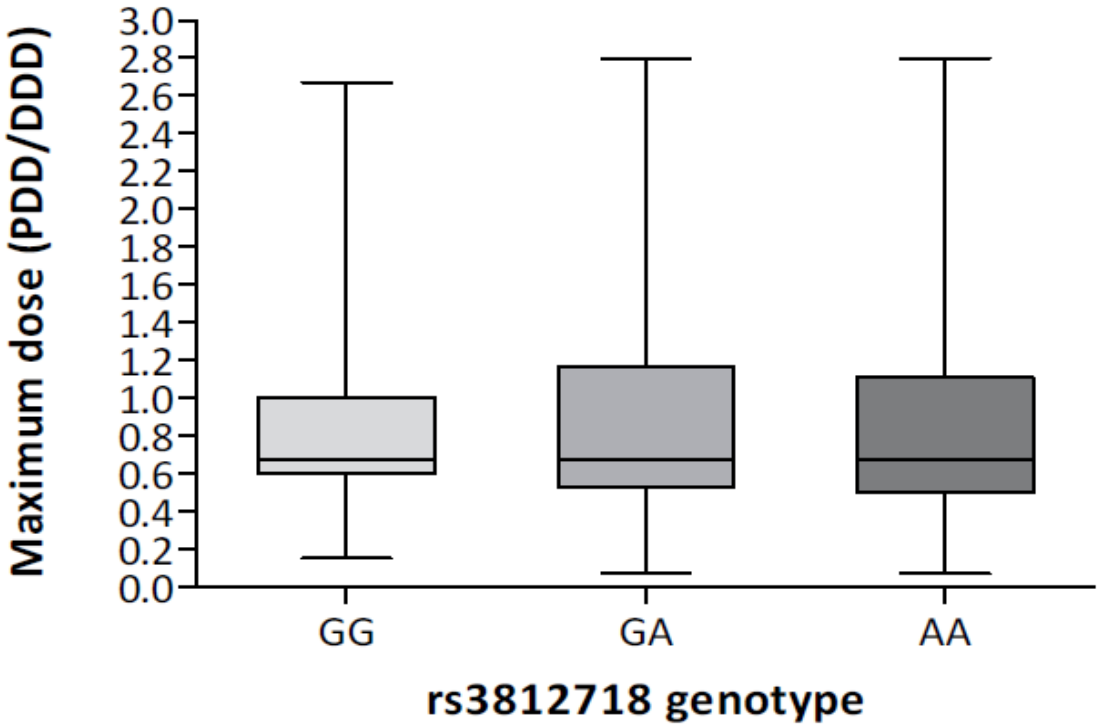


Figure 4.9 Distribution of maximum dose ratio for each drug

Aligned dot plot of maximum dose ratio (prescribed daily dose/defined daily dose (PDD/DDD)) distribution for each antiepileptic drug included in the analysis. Lines represent the mean dose ratio and dots represent dose ratios for individual patients (n=795). LTG = lamotrigine; CBZ = carbamazepine; OXC = oxcarbazepine; TPM = topiramate; GBP = gabapentin; VPA = valproate.

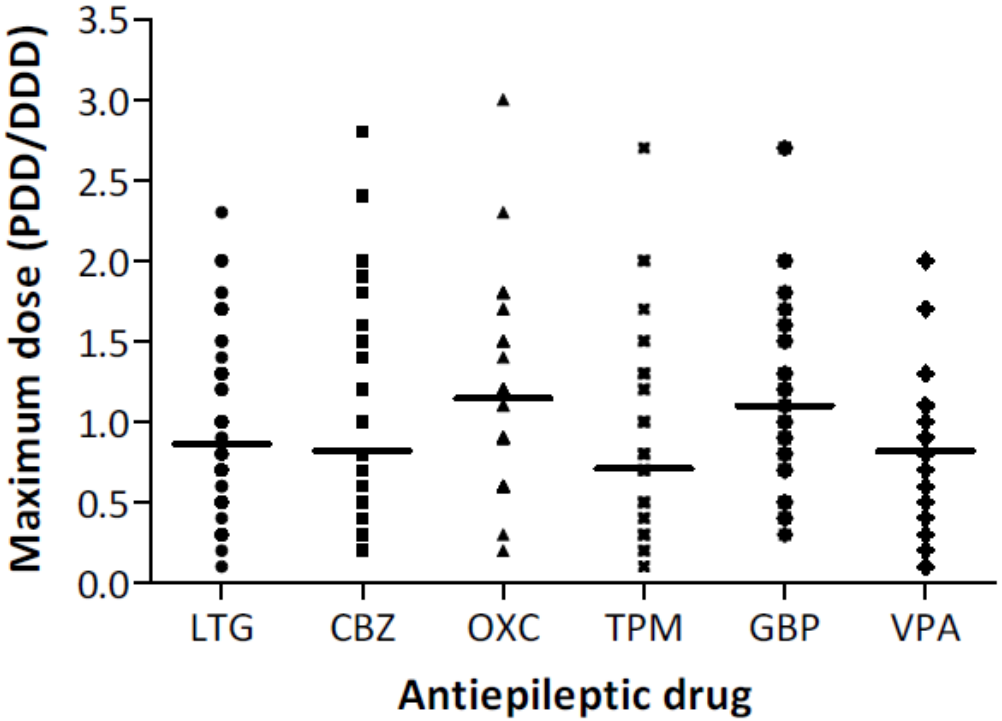


Table 4.11 Regression analysis of interaction effects with maximum dose ratio

Covariate	Parameter	Regression coefficient	Parameter *P*-value
Individual drugs and clinical variables	IGE	0.06	0.06
	LRE	0.03	0.25
	UNC	0a	.
	CBZ	-0.01	0.63
	OXC	0.14	0.00
	TPM	-0.09	0.00
	VPA	-0.04	0.21
	GBP	0.12	0.00
	LEV	0.12	0.02
	LTG	0a	.
	rs3823728	0.02	0.17
Genetic model Individual drugs plus clinical covariates plus SNP	IGE	0.06	0.06
	LRE	0.03	0.23
	UNC	0a	.
	CBZ	-0.02	0.53
	OXC	0.14	0.02
	TPM	-0.08	0.05
	VPA	-0.06	0.24
	GBP	0.15	0.00
	LEV	0.13	0.12
	LTG	0a	.
	rs3823728	0.02	0.47
	CBZ * rs3823728	0.01	0.69
	OXC * rs3823728	0.00	0.94
	TPM * rs3823728	-0.01	0.69
	VPA * rs3823728	0.02	0.61
	GBP * rs3823728	-0.04	0.37
	LEV * rs3823728	-0.01	0.86
	LTG * rs3823728	0a	.
LRT *P*-value			0.0006

IGE = idiopathic generalised epilepsy, LRE = localisation-related epilepsy, UNC = unclassified epilepsy, LRT = Likelihood Ratio Test, 0a = parameter set to 0 by SPSS as redundant in model

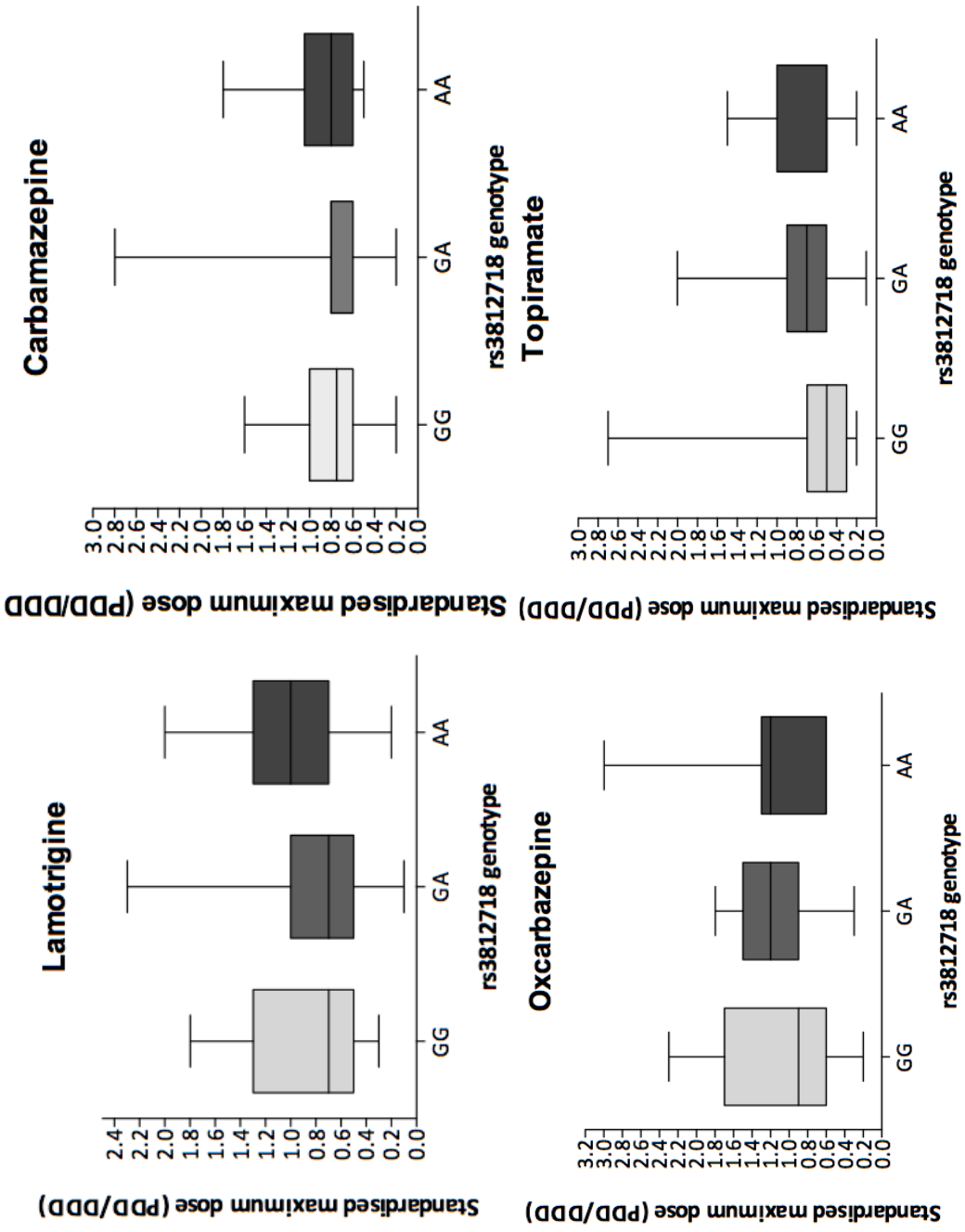


Figure 4.11a Box plots for distribution of maximum dose ratios with genotype Distribution of maximum dose ratios of AEDs lamotrigine, carbamazepine, oxcarbazaepine and topiramate, based on *SCN1A* rs3812718 genotype groups. Boxes represent 25th and 75th percentile, solid lines represent the median dose ratio and whiskers represent minimum and maximum dose ratio in each genotype group.

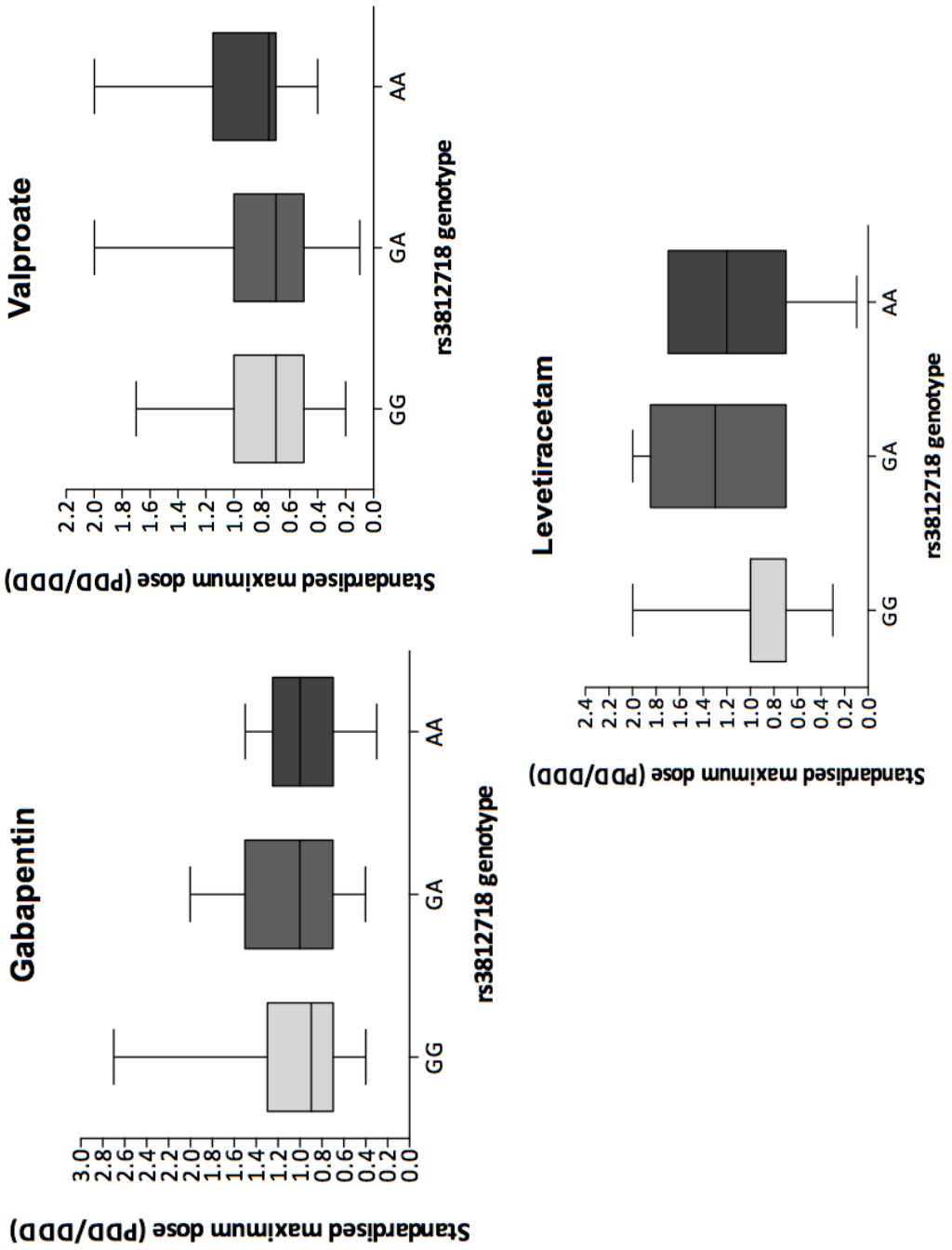


Figure 4.11b Box plots for distribution of maximum dose ratios with genotype Distribution of maximum dose ratios of AEDs gabapentin, valproate, and levetiracetam, based on *SCN1A* rs3812718 genotype groups. Boxes represent 25th and 75th percentile, solid lines represent the median dose ratio and whiskers represent minimum and maximum dose ratio in each genotype group.

Table 4.12 Mean maximum dose ratio for each genotype group stratified by individual antiepileptic drug

	rs3812718 Genotype		
Antiepileptic drug	**GG**	**GA**	**AA**
Lamotrigine	0.88	0.81	1.02
Carbamazepine	0.78	0.81	0.93
Oxcarbazepine	1.11	1.13	1.19
Topiramate	0.70	0.73	0.67
Gabapentin	1.11	1.12	0.94
Valproate	0.80	0.75	0.93
Levetiracetam	0.95	1.30	1.03

4.4. Discussion

A previous report has implicated the *SCN1A* rs3812718 G>A variant in influencing the maximal dose of two established Na^+ channel blocking AEDs (PHT, CBZ) administered to individuals with epilepsy. Although not all attempts to validate the potential significance of this SNP in several independent PGx investigations were successful in identifying a similar association with AED dosing, the evidence produced by a number of recent functional studies is promising.

Despite failure to replicate the original association with CBZ, the present investigation provides support for the role of rs3812718 in maximum AED dosing. The effect of this polymorphism on maximum and maintenance doses of AEDs was examined, irrespective of their primary drug target, using a normalised drug dose (PDD/DDD ratios). A significant association between AED maximum dose ratio and the rs3812718 polymorphism was identified in this investigation, suggesting the variant genotype of this SNP may influence AED maximum dose in newly-diagnosed epilepsy. Individuals with the rs3812718 variant AA genotype showed a significantly higher PDD/DDD ratio than those with the rs3812718 GG genotype (Table 4.10, Figure 4.8). In contrast no significant genotype effect on dosing was observed for maintenance dose (Table 4.7, figure 4.6).

4.4.1. *SCN1A* genotype affects maximal antiepileptic drug dosage

The association between SNP genotype and maximum AED dose identified in this study appears to confirm the initial hypothesis that polymorphisms in Na_V channel genes can influence the dosing of AEDs. The original study by Tate *et* al 2005 reported that for both PHT and CBZ, average maximum dose differed by genotype in the order AA>AG>GG (Tate et al., 2005). The rs3812718 polymorphism was also shown to affect the proportion of alternative transcripts in brain tissue from individuals with a history of refractory epilepsy, which could in turn affect channel sensitivity to sodium blocking activity of AEDs (Tate et al., 2005, Tate et al., 2006). The authors thus proposed that common polymorphisms in *SCN1A* alter the sensitivity of the Na_V channel to Na^+ channel blocking drugs (Tate et al., 2005, Tate et al., 2006). The present investigation included dose data from several different classes of AEDs including the previously associated AED CBZ, and so was not confined to those known to exhibit Na^+ channel blocking activity, though association was still evident.

4.4.2. Non-specific effect of the *SCN1A* variant on maximum dose

The mechanism of action of most AEDs is not completely understood (Kwan et al., 2001). The majority of commonly utilised AEDs are generally classed into three main types based on their individual molecular site of action, mainly; i) those modulating voltage-dependent ion channels (Na^+, Ca^{2+}, K^+), ii) those enhancing GABA mediated inhibitory neurotransmission, and iii) those involved in the attenuation of excitatory (particularly glutamate-mediated) transmission (Meldrum, 1996, Kwan et al., 2001). Out of the seven AEDs analysed CBZ, LTG and OXC are known to principally modulate Na_V channels (Kwan et al., 2001, Schachter, 2007). On the other hand VPA and TPM have been proposed to display a number of mechanistic pathways. In addition to Na_V blocking properties they are also associated with Ca^{2+} blockade and facilitation of the effects of the inhibitory neurotransmitter GABA (Kwan et al., 2001, Schachter, 2007). GBP appears to bind to the α2δ subunit of neuronal voltage-gated calcium channels, but has been suggested to have some Na_V blocking activity. LEV is associated multiple mechanisms. In addition to the novel mechanism of SV2A protein binding, LEV has actions on neuronal GABA- and glycine-gated currents and K^+ currents, though it's exact mechanism of action is unknown. The hypothesis of the present investigation was that the functional *SCN1A* variant which has previously been demonstrated to affect the pharmacological and/or structural properties of the Na_V channel (Thompson et al., 2011), and alter maximum dose for PHT and CBZ could also alter therapeutic dosage requirements of some and/or all these additional AEDs (Tate et al., 2005, Tate et al., 2006).

4.4.3. Evidence for drug-gene interaction and differential drug effect

A recent study concerning the Na_V channel and *SCN1A* rs3812718 variant demonstrated an alteration in Na_V channel sensitivity to LTG and PHT but not CBZ, suggesting a differential effect of the rs3812718 splicing variant on the binding properties of the Na_V channel (Thompson et al., 2011). The variant $Na_V1.1$-5N channel was shown to exhibit greater tonic and use-dependent inhibition by PHT and LTG than the $Na_V1.1$-5A channel, suggesting that binding sites for these AEDs could slow down inactivation processes, which result in pharmacologic differences between AEDs (Thompson et al., 2011). At therapeutically relevant concentrations, the $Na_V1.1$-5N channel was more sensitive to PHT and LTG. The authors proposed an alteration in LTG and PHT dose requirement due to this increase in channel sensitivity (Thompson et al., 2011).

The interaction analysis performed for maximum dose in the present investigation identified a stronger genetic effect when drug type was taken into account. This appears to confirm the original SNP association reported by Tate *et* al 2005 and also implies that this association may additionally be influenced by the AED administered (Thompson et al., 2011). Due to the low numbers of individuals that were prescribed each of the individual AEDs in the present study, however, this finding could not be fully stratified by drug type. And so, although the present findings present a strong statistical association (P<0.01) between genotype and dose when drug interaction was considered, one cannot accurately distinguish which AEDs were most influenced by rs3812718 genotype. The present results can however be taken to signify that a drug specific genotype effect may exist. Further investigation is necessary for more conclusive evidence for drug-SNP interaction in AED maximum dosing.

Finally, the original association reported with CBZ maximum dose and rs3812718 was not validated. No association between CBZ maximum dose and genotype was evident from the interaction regression analysis or when CBZ was tested alone (Interaction regression P = 0.598; ANOVA P = 0.207; Figure 4.11a and b; Table 4.12). This implies lack of influence of rs3812718 on CBZ maximum dose.

4.4.4. rs3812718 variant genotype does not influence maintenance dose

Maintenance dose can be used as a measure of the dose at which optimum response is observed (Patsalos and Bourgeois, 2010, Talati et al., 2011), and presumed to reflect dose at which seizures were controlled in this study. Maximum dose is, in contrast, most likely a measure of an individual's tolerability to an AED (Dlugos et al., 2006). Maintenance dose may therefore be a more informative measure of seizure control than maximum dose and a more accurate depicter of clinical effect (Tate et al., 2006). This was acknowledged by the authors of the original report (Tate et al., 2006), who in a subsequent study attempted to correlate

maintenance dose of PHT with rs3812718 genotype and presented a modest SNP-dose association. The present analysis similarly examined maintenance dose for several AEDs, though failed to find an association between AED maintenance dose and the rs3812718 SNP.

Reasons for the lack of association with maintenance dose in the present analysis includes; the limited availability of maintenance dose data. In addition, there were considerably fewer individuals with maintenance dose data than maximum dose data in the present study (n=301 vs. n=795, respectively). This was not surprising given the stringent definition of maintenance dose, the fact that only 60-70% of all newly-diagnosed patients can expect to achieve a 12 month remission, and that multiple maximum doses were available for some patients.

This methodology for dose analysis was for the most part beneficial, as it allowed the combination of data from different AEDs. Dose standardisation however can dilute drug specific genotype associations, i.e. AEDs that primarily block the Na_V channel as presumed in the original hypotheses (Tate et al., 2005, Tate et al., 2006). Although stratification by drug adjusts for a drug specific dose-SNP effect, an interaction analysis was only performed if an association was identified in the initial linear regression analysis of SNP vs. dose and this was not the case with maintenance dose.

The original PGx study concerning rs3812718 in AED dosing was the first publication presenting the potential effect of a primary AED target polymorphism on the clinical use of anticonvulsant drugs (Tate et al., 2005) and was strengthened by functional evidence from brain tissue expression data (Tate et al., 2005, Heinzen et al., 2007, Thompson et al., 2011). Other studies have investigated the rs3812718 SNP, with one showing association with LTG dosing in Caucasians, and also a significantly higher frequency in epilepsy patients compared to controls, implying that this polymorphism may additionally contribute to the pathogenesis of epilepsy (Krikova et al., 2009). The AA genotype has also been shown to be significantly more frequent in Japanese patients resistant to CBZ treatment (Waldegger et al., 1999).

In another study, no relationship was found in 377 Chinese patients between Na_V blocking AEDs and rs3812718 genotype (Kwan et al., 2008). More recently a study by Mann *et* al 2011 investigating CBZ and OXC in drug-resistant and drug-responsive subjects from Italy similarly concluded no major role of the SCN1A rs3812718 polymorphism as a determinant of AED response (Manna et al., 2011). The failure to identify an association with CBZ in the present analysis could be attributed to the possibility that the effect size of the original association may have been overestimated. If the original relationship identified between CBZ and PHT and the variant allele of rs3812718 was only modest in size, it may not have been detected in the present patient population, which was smaller than that used in the original study (CBZ treated patients n=168 and n=425 respectively).

Despite the lack of consistency in PGx data surrounding the *SCN1A* rs3812718 variant, and failure to replicate the original association with CBZ, the present investigation provides support for the role of rs3812718 in maximum AED dosing. However, it is likely that only a small proportion of the variation in AED dose can be explained by the variant. Because of the complexity of drug response phenotypes it is expected that additional genetic factors are also involved in the variability of AED dosing, and this could additionally explain the lack of consistency in previous studies. Most previous investigations solely involved Na_V channel blocking drugs and patients with unknown (or at least unstated) causes of epilepsy. Different cohorts are also likely contain different ratios of genetic or non-genetic epilepsy syndromes, and so the influence of rs3812718 on AED dosage may be obscured or outweighed in some patient populations. Discrepancies between these studies may thus mainly result from varying; i) cohort size, ii) heterogeneity of epilepsy syndromes in population samples, or iii) differences in ethnic backgrounds (so far Caucasian, Chinese and Japanese patient cohorts have been investigated). Finally some studies have shown that in experimental epilepsy models there is a significant change in expression of Na_V channels in response to seizures, and so seizure frequency, and/or epilepsy severity may also be a contributing factor to changes in response to AEDs (Gastaldi et al., 1997, Aronica et al., 2001).

The report by Tate *et* al 2006 identified an association between rs3812718 with serum concentrations of PHT at maintenance dose, with no associations observed for maximum dose (Tate et al., 2005, Tate et al., 2006). This additional drug concentration data eliminated PK factors as a source of variation, revealing a relationship with maintenance dose. This could explain our findings, in that the lack of concentration data may be masking any genotype effect on maintenance dose. Serum AED levels, if available, may be more successful for identifying associations between the *SCN1A* variant and drug doses or treatment outcomes in future analyses.

Limitations have also been recognised in the original investigation that potentially confound the dependability of the observations reported. The main issue being the sole use of maximum dose data by the authors. Maximum tolerated dose could be a useful indicator of individual dose ceiling, however, in the treatment of epilepsy moderate doses of AEDs are usually sufficient for seizure control and patients may never reach their individual limit of tolerability. Maintenance dose data can serve as a more accurate and informative measure of clinical response. Another potential confounding factor in the original report was the inclusion of both monotherapy and polytherapy patients which could affect the reliability of any associations observed. AEDs are highly susceptible to drug-drug interactions (Patsalos and Perucca, 2003, a, Perucca, 2006), therefore are associated with altered serum drug concentrations, often necessitating dosage adjustment, and can also influence AED tolerability (Johannessen et al., 2003, Anderson, 2008). Age and concomitant medication are additional

factors that can greatly impact drug PK thereby altering the amount of AEDs required by individual patients. The original authors likewise did not provide basic data such as age at onset of treatment, disease aetiology and syndrome type (discussed above). Considering these limitations, one can argue that, the rs3812718-AED dose association is unproven. The results presented in this study in combination with the recent positive associations and latest functional evidence may however still provide a link between patient genotype for drug target variants and AED response.

In summary, our data suggests that the *SCN1A* rs3812718 G>A polymorphism may influence maximum dose of AEDs. However, the modest effect size would question its clinical utility. Further analysis of the effect of this polymorphism on individual Na_V blocking AEDs may be useful. The validation of the original hypothesis of Tate *et* al 2005, 2006, by identifying a genotype-dose association is promising, and indicates the relevance of drug target polymorphisms in individual AED treatment. Nevertheless, the necessity for consistent results to confirm the true effect of rs3812718 on AED dosing remains. Replication of the current results of *SCN1A* genotype-drug and dose associations using a broad selection of AEDs is required. Likewise future investigations utilising dosage, and serum concentration data, ideally in a larger cohort of patients, with homogeneous epilepsy phenotypes could prove beneficial to the *SCN1A* rs3812718 polymorphism story.

CHAPTER FIVE

VALIDATION OF A MULTIGENIC MODEL FOR TREATMENT RESPONSE IN NEW-ONSET EPILEPSY

CONTENTS

5.1. Introduction

The majority of PGx research concerning AED response has, until recently, focused on a relatively small number of SNPs in a few selected candidate PK and PD genes (Depondt, 2006b, Loscher et al., 2009), an overview of which has been given in an earlier chapter (Chapter 1). However, many PGx investigations, including those in the epilepsy field, now routinely include multiple genes implicated in both disease pathogenesis and the PK and PD pathways of drugs (Grant and Hakonarson, 2007, Petrovski et al., 2009, Motsinger-Reif et al., 2010, Cavalleri et al., 2011). This has led to an interest in relevant analytical methods for modeling large volumes of data that take into account the complex, multifaceted network of genes involved in such drug response phenotypes. Accurate classification and prediction algorithms from systems biology methodologies are thought to help meet this data analysis challenge (Baksh and Kelly, 2007). These are not only designed to allow for the multigenic-multifaceted nature of drug response data and gene-network interactions, but are also considered better for the statistical challenge of detecting multiple, small associations in high-dimensional data and thus may ensure more efficient data analysis (Hirschhorn et al., 2002, Ritchie and Motsinger, 2005, Baksh and Kelly, 2007, Pander et al., 2010, Rodin et al., 2011, Vanneschi et al., 2011).

5.1.1. Systems biology approach to genomic analysis

A systems biology approach to analysis of pharmacogenomics data typically involves three-steps; i) selection of variables (SNPs), ranked in order of effect on drug response phenotype, ii) a modelling step involving the generation of a predictive model using SNPs and other relevant factors, and iii) evaluation of generated models using conventional statistical analysis methods (Koster et al., 2009).

5.1.2. Machine learning methods in genomic prediction

ML is a computer-based data mining method derived from the field of artificial intelligence and concerned with the design and development of algorithms to allow machines to learn, make predictions, or extract knowledge from data. It represents a powerful approach to identifying non-linear/complex patterns in high-dimensional datasets (Hastie et al., 2001, McKinney et al., 2006, Zhang and Rajapakse, 2009) and makes intelligent decisions based on knowledge from data or to make predictions on new data (Hastie, Tibshirani et al. 2001) (Figure 5.1).

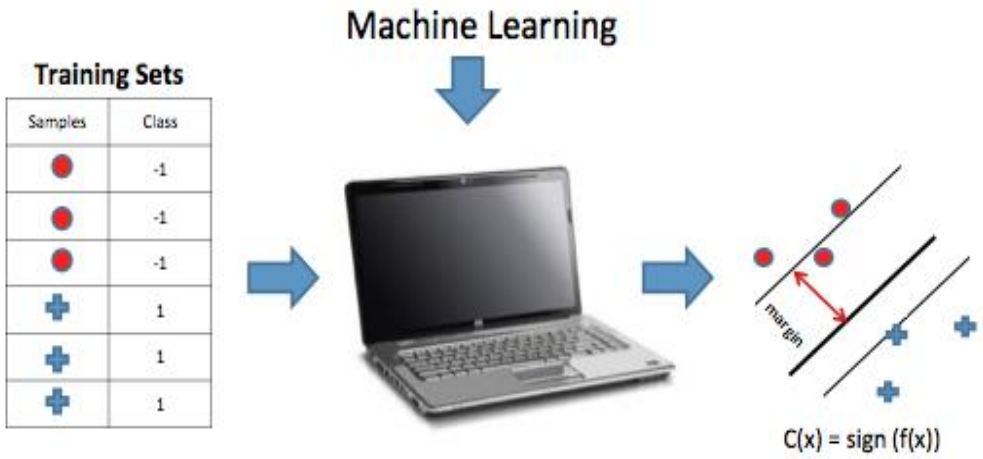


Figure 5.1 The machine learning approach to data inference

A schematic representation of the computer based machine-learning approach to data analysis. This entails learning or data inference from existing data of known values (training data). Algorithms are then generated and used to make predictions for new data of unknown values.

Typical ML methods applied to genomic studies model data using Bayesian networks, which allow the inferential exploration of previously undetermined relationships among genetic and clinical variables, and describe these relationships once identified (an hypothesis- or model-free approach) (Hoppe, 2005, Zhang and Rajapakse, 2009, Rodin et al., 2011). In recent decades ML approaches have been successfully applied to computational biology and bioinformatics (Hastie et al., 2001, Larranaga et al., 2006, Zhang and Rajapakse, 2009) and are now becoming routine in the biological domains of genomics, proteomics, microarrays and systems biology (Bhaskar et al., 2006a, Larranaga et al., 2006).

ML approaches can be used for the development of prediction models that allow integration of the interactions between multiple genetic variables i.e. SNPs in addition to clinical variables and disease phenotype and so overcome the main limitation of traditional statistical approaches through their ability to model high-dimensional data (Hoppe, 2005, Wilke et al., 2005). Additional advantages of ML methods include robustness of parametric assumptions, high power and accuracy (useful for extracting information from underpowered association studies), ability to model non-linear effects, and the availability of numerous well-developed algorithms (Moore and Ritchie, 2004, McKinney et al., 2006). ML models for data

classification may better identify patterns of genetic variants that associate with phenotypes of interest in high-dimensional data (Lee et al., 2008).

5.1.3. Supervised and unsupervised learning methods

Learning scenarios for ML application can be categorised as either supervised or unsupervised (Hastie et al., 2001). In unsupervised learning, there is no outcome measure, and the goal is to describe how the data are organised or clustered (objects are often classified by a similarity measure that defines how closely related those objects are). The goal of supervised learning is to predict the value of an outcome measure based on a number of input measures or features (i.e. using prior knowledge from existing data for training). The classifier is then used to generalise from new instances (Hastie et al., 2001, Kotsiantis, 2007, Emmert-Streib and Dehmer, 2010). Supervised learning algorithms usually produce classifiers in the form of a function (Emmert-Streib and Dehmer, 2010) and are more relevant and applicable to the mining of genetic data for disease association analysis (McKinney et al., 2006).

5.1.4. Machine learning prediction models or classifiers

In a typical supervised ML scenario, an outcome measurement (i.e. cardiac arrest/no cardiac arrest) is predicted based on a set of features such as diet and clinical measurements. There is also a training set of data in which the outcome and feature measurements can be observed for a set of seen or known objects (i.e. patients) (Hastie et al., 2001). Using this training data, a prediction model (or classifier) is built using a function that enables prediction of outcome to be made for new unseen objects (Hastie et al., 2001). Several methods exist for assessing ML classification models. Model performance is usually assessed by how well the classifier can predict outcomes for independent test datasets based on the rules it has learned from the training data (Hastie et al., 2001, Larranaga et al., 2006). Other common methods of assessment include cross-validation and bootstrapping. Figure 5.2 shows a schematic representation of ML classification models and the main stages involved in model building. A more detailed description of the development and assessment of ML models for disease classification, including details of several ML approaches commonly applied to genomic data, can be found in Chapter 6.

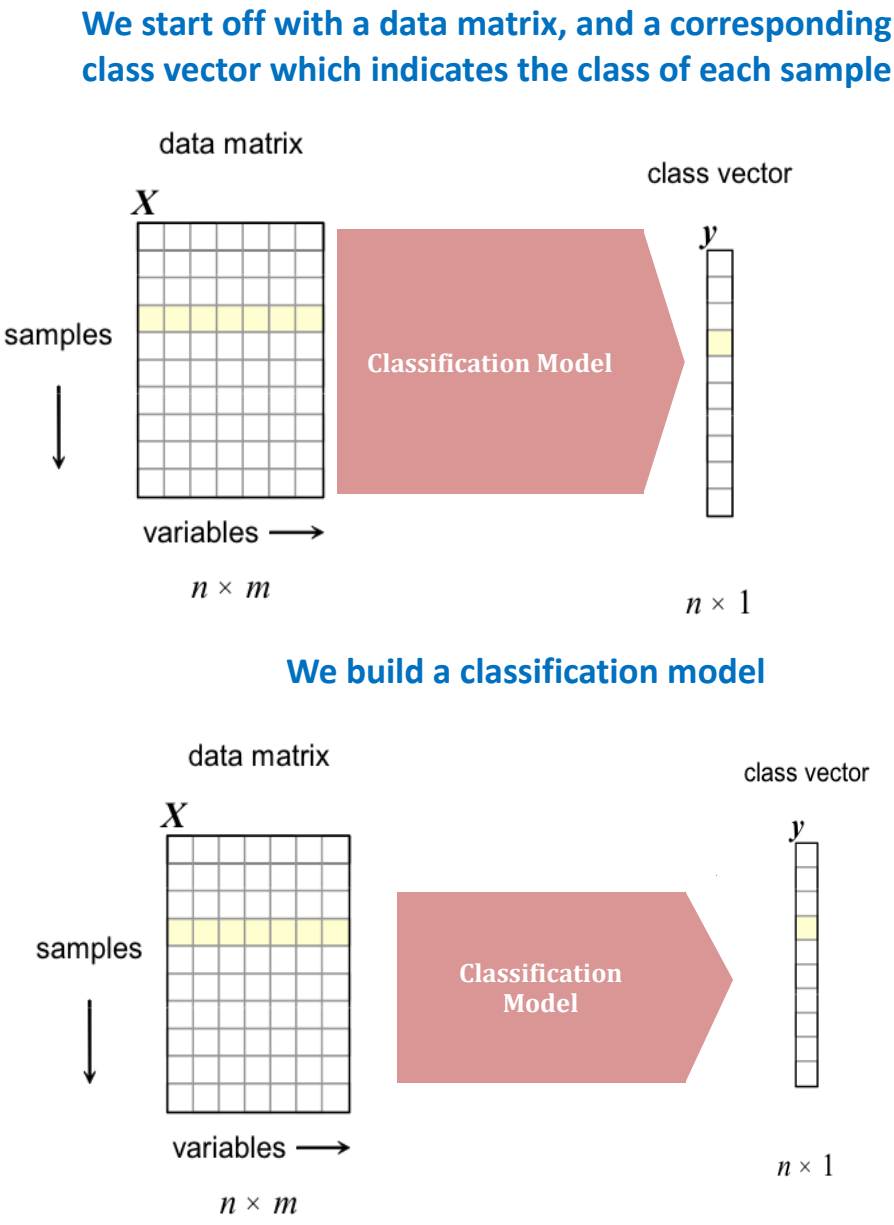


Figure 5.2 Building machine learning classification models

5.1.5. Machine learning algorithms for genomic classification

The use of systems biology data analysis methods derived from ML have steadily grown in genomics over the last two decades (Hastie et al., 2001, Emmert-Streib and Dehmer, 2010). ML approaches, known as classification algorithms, are well suited to genomic data and are widely employed (McKinney et al., 2006, Kotsiantis, 2007, Rodin et al., 2011). Classification algorithms are able to predict (or assign) a class (e.g. case or control) to each data point, based on the values of potentially predictive variables (e.g. SNPs) and additionally build a model capturing the relationships between the variables (Kotsiantis, 2007, Rodin et al., 2011). Classification algorithms were first applied to SNP data within bioinformatics for the prioritisation of candidate polymorphisms by predicting their likely impact on disease susceptibility (Shah and Kusiak, 2004, Rodin et al., 2011, Zhao et al., 2011).

There are numerous ML approaches available for both selecting informative features and/or combining them into a classifier, ranging from simple linear classifiers to complex nonlinear functions (Hastie et al., 2001, Kotsiantis, 2007). Examples of those commonly found in multi-locus traits (i.e. diabetes, heart disease, alcoholism and breast cancer) include multi-factor dimensionality reduction (MDR)(Ritchie and Motsinger, 2005, Vanneschi et al., 2011), neural networks (Lucek and Ott, 1997, Motsinger et al., 2006), random forest (RF) (Yoon et al., 2003, Bureau et al., 2005), support vector machine (SVM) (Yoon et al., 2003, Yu and Shete, 2005) and *k*-nearest neighbor (*k*NN), for which complex disease data is only just emerging (Szymczak et al., 2009). Examples of the application of these ML approaches can also be found for gene discovery in GWAS data from multiple sclerosis and type II diabetes (Szymczak et al., 2009, Ban et al., 2010, Goldstein et al., 2010).

5.1.6. Machine learning approaches in pharmacogenetics

ML approaches have also recently been proposed for modeling SNP data to produce PGx SNP classifiers, which may be more effective for predicting treatment outcome in drug response data than the standard linear regression data modeling approach (Pander et al, 2010). Whilst there is a large amount of literature on the development of ML approaches for the analysis of high dimensional data, much of this research is in the context of disease status, susceptibility and activity, whereas the application of ML approaches to PGx data is a relatively new phenomenon (Simon, 2005, Simon and Wang, 2006, Pander et al., 2010). The development of predictive models incorporating the interplay of numerous genetic factors (amongst other features) potentially contributing to this multi-genic phenotype presents an ideal strategy for confirming which genes and/or gene-gene or gene-environment interactions are of most significance, and is thus potentially a critical step on the road to individualised prescribing and new drug discovery (Simon and Wang, 2006).

PGx studies utilising ML methods are still emerging (Lee et al., 2008, Pander et al., 2010). A number of recent examples of the application of ML algorithms to pharmacogenomics data from several important disease domains can be found. These include predicting treatment response to chemotherapeutics in oncology (Ad et al., 2002, Ritchie and Motsinger, 2005, Simon, 2005, McKinney et al., 2006, Simon and Wang, 2006, Wang, 2007, Lee et al., 2008), anti-retroviral therapy (Altmann et al., 2007), toxicity to statin therapy (Ritchie and Motsinger, 2005) and, more recently, predicting warfarin dose (Cosgun et al., 2011).

5.1.7. Machine learning methods for detecting epistatic interactions

As discussed previously, gene-gene interactions or epistasis is a well-known challenge in data analysis for complex traits and has been recognised as a problem that needs to be addressed in PGx (Ritchie and Motsinger, 2005, Pander et al., 2010). A growing number of researchers are now considering the use of data-reduction ML techniques previously used for advanced genetic interaction or environmental factor analysis, including the MDR method mentioned above (Moore et al., 2004, Moore and Ritchie, 2004, Ritchie and Motsinger, 2005, McKinney et al., 2006, Moore et al., 2006). MDR is a ML method specifically designed to identify interacting combinations of genetic variants associated with increased risk of common, complex, multifactorial human disease (Ritchie et al., 2003, Moore, 2004).

5.1.8. Application of machine learning to epilepsy pharmacogenetics

The application of ML approaches is one of the latest developments in epilepsy PGx. So far there are only three published examples of the development of predictive models for treatment response in epilepsy research. These include two investigations utilising the MDR data reduction method (Kwan et al., 2008, Jang et al., 2009, Kim et al., 2011b) and a single study applying a ML data-mining approach (Petrovski et al., 2009).

5.1.9. *k*NN machine learning method in epilepsy pharmacogenetics

A ML data-mining approach was applied in a recent, proof of principle study examining PGx data from patients with epilepsy (Petrovski et al., 2009). The study utilised a *k*NN algorithm to develop a multi-SNP classification model that was proposed to predict response to initial AED treatment in Australian patients with newly treated epilepsy, with a predictive accuracy of 83.5%. A total of 4041 SNPs from 279 candidate genes were genotyped in 115 patients, five of which were ranked as having the most influence on treatment outcome (Petrovski et al., 2009). The ML supervised learning *k*NN algorithm was then used to develop

a classification model based on the genotype of these five SNPs (Petrovski et al., 2009). The *k*NN classifier algorithm was designed and implemented using in-house software by the original authors (Petrovski et al., 2009). The predictive value of the model was subsequently confirmed in two small, independent Australian cohorts. It was reported to have a sensitivity of >80% and in each of these replication cohorts, the multigenic model proved to be more accurate in predicting drug responsiveness than any of the single SNPs alone (Petrovski et al., 2009).

5.1.10. Summary and research aims

As discussed previously, a substantial proportion of people with epilepsy continue to have seizures despite treatment with appropriate AEDs (Kwan and Brodie, 2000a, Duncan et al., 2006, Szoeke et al., 2006). It is not currently possible to accurately predict the likelihood of seizure control with any given AED treatment. Success or failure in terms of efficacy consequently cannot be adjudged until a therapeutic dose is reached, often many weeks or months after treatment initiation (Kwan et al., 2010). As such, the identification of biological markers that may provide improved prediction of treatment response in an individual patient is likely to be of significant clinical value. The Australian multigenic *k*NN classifier not only represents a successful application of a ML approach to epilepsy PGx data but also identified biological markers that might prove clinically significant to AED response. However, this classifier requires validation in larger cohorts and application across different populations and health care systems to adequately assess its reliability prior to consideration for use in clinical care.

There were two main aims of this study; (i) to assess the broader clinical utility of the Australian multigenic *k*NN model, and (ii) to assess relevance of the five SNPs comprising the classifier to treatment response in non-Australian populations. On that basis, the Australian classifier was applied to genetic data from two independently collected UK cohorts to assess whether it could successfully classify treatment outcome. The relative influence of the five SNPs on treatment outcome was also tested both individually and collectively in these patients.

5.2. Methods

5.2.1. Source populations

Patients from three independent cohorts of newly treated epilepsy were included in this analysis, the Australian cohort in which the multigenic *k*NN model was originally derived (Petrovski et al., 2009) and two UK cohorts (Glasgow and SANAD cohorts). The UK cohorts are described in detail in the general methods section (section 2.2). Patients were initially selected from UK cohorts if they: (i) were newly treated for epilepsy, (ii) were treated with AED monotherapy during their first year of treatment, (iii) had sufficient clinical information available, defined as at least one year of follow-up with detailed drug, dose, and outcome information, (iv) were of self-reported European ancestry, and (v) had provided a DNA sample for genotyping.

5.2.2. Data extraction

Clinical information was extracted from patient databases and or clinical notes for each of the UK cohorts and included age at recruitment, gender, epilepsy type, seizure type(s) and also drug treatment history for first 12 months of treatment (including AEs and reasons for switching AED). Epilepsy type was classified into three categories, IGE, LRE and UNC.

5.2.3. Phenotype definitions for patient selection

Response to AED treatment in the UK cohorts was determined in accordance with the definitions used to phenotype the Australian cohort in the original study (Petrovski et al., 2009). Patients were considered to be "responders" if they remained free from seizures throughout the first 12 months of AED treatment. Seizures arising in the first month of treatment (i.e. during drug titration) and those associated with short-term non-compliance with medication or significant provocation (e.g. sleep deprivation) were discounted. In contrast, patients who continued to experience unprovoked seizures during the first year of therapy despite adequate AED exposure were considered to be "non-responders". Where the first ever AED was discontinued within the initial 3 months of treatment as a result of intolerable AEs, the second AED was considered the 'initial drug' for the purposes of this analysis. Patients in whom clinical information was insufficiently detailed to allow a confident classification of response or who were suspected to be non-adherent with medication were excluded from the analysis, as in the original study (Petrovski et al., 2009).

5.2.4. Study populations

Treatment outcome phenotypes for the Glasgow and SANAD cohorts were identified by interrogation of the trial or clinical databases and/or clinical notes. A total of 285 and 520 individuals were included for analysis from Glasgow and SANAD cohorts, respectively.

5.2.5. Genotyping

The five SNPs that comprised the Australian multigenic classifier were rs658624 and rs678262 from the *SCN4B* gene and rs2808526, rs4869682, rs2283170 from the *GABBR2*, *SLC1A3* and *KCNQ1* genes respectively. All 285 samples from the Glasgow cohort were genotyped for these five SNPs at the Australian Genome Research Facility using an iPLEX Gold assay on the Sequenom MassARRAY compact analyser (Sequenom Inc., San Diego, California, USA) (Petrovski et al., 2009). SANAD samples (n=520) were genotyped in the Department of Molecular and Clinical Pharmacology, University of Liverpool on a Sequenom MassARRAY iPLEX platform in accordance with the manufacturer's instructions (Gabriel et al., 2009) and as described in detail in section 2.5.

5.2.6. The development of the Australian kNN multigenic model

The methods involved in the development and validation of the Australian multigenic *k*NN classifier model (Petrovski et al., 2009) were used for model development for this investigation and are described in detail in the original report (Petrovski et al., 2009). For model generation in the present study, the SANAD cohort was randomly stratified into training (70% of patients, n=343) and test (30% of patients, n=148) datasets in a manner similar to that originally described for the Australian cohort (Petrovski et al., 2009). The 70% training dataset was used for model development (training set patients are used to identify optimum parameters for accurate patient classification and testing association of five SNPs with patient outcome) and a 30% test dataset for assessing the predictive potential of the model. Each patient in the test dataset (30% of patients) was positioned in an N-dimensional space (in this case N=5, representing the SNP genotypes in the five SNP model) defined by the training dataset (70% of patients), with response predicted by simple majority of known treatment responses amongst its *k*-nearest neighbours (i.e. the individuals with the most similar genetic profiles at this combination of five SNPs). The number of nearest neighbours found to give optimal prediction of drug response in the Australian cohort was $k = 9$. A 20% cross-validation methodology was adopted for building the *k*NN model on the training dataset, where the training dataset is divided into five equally sized groups, each of which is excluded in turn and the remaining four groups used to test the model. This cross-validation step in the initial model development also allowed the determination of the optimal number of *k*- nearest neighbours to use (Petrovski et

al., 2009). An overview of the *k*NN procedure, as used to develop the five-SNP classifier is illustrated in Figure 5.3.

5.2.7. Statistical analysis for assessment of cohort differences

The initial statistical analysis examined patient demographics and drug treatment outcomes across the UK and Australian cohorts to identify differences that might confound subsequent analyses. Age at enrolment was assessed using ANOVA, while gender, initial AED, epilepsy type and drug treatment response were all assessed using Chi-square tests. Each of the five SNPs was also assessed for independent association with AED response in each of the two UK cohorts using the Cochran-Armitage test for trend. Any systematic differences identified in the demographic or genetic variables (SNPs) were adjusted for prior to any subsequent model assessment.

5.2.8. Approaches for classifier assessment

Several methods exist for assessing ML classification models. The Australian multigenic *k*NN classifier model was evaluated in the UK cohorts in three ways: (i) treating UK patients as independent test sets and using Australian patients to predict response in UK patients, (ii) re-deriving the *k*NN classifier using the SANAD cohort as both training and test datasets, and (iii) testing the performance of the *k*NN model in UK training datasets using a cross-validation n-1 approach.

Given that the kNN model is not based on a fixed algorithm, but rather a five-dimensional training dataset, its reliability is dependent on similar frequencies of drug response in the training and test datasets. When comparing the original Australian cohort with the two UK cohorts, there was a difference in treatment response frequencies between the groups. In the Australian cohort, 28% of patients were unresponsive to their initial AED, compared with 47% and 52% in the Glasgow and SANAD cohorts, respectively (Table 5.1). As a result, using the Australian cohort as the *k*NN training dataset to predict treatment response in either of the UK cohorts was expected to result in an over-estimation of the number of responders (i.e. false positives). Therefore, in addition to a direct test of the Australian five-SNP model, a secondary analysis using predictions derived from the UK cohorts themselves was required to obtain a realistic understanding of the influence of this combination of five SNPs on treatment outcome in newly treated epilepsy patients from the UK.

5.2.9. Data stratification to account for UK cohort differences

Due to differences between the Australian and UK cohorts, the Glasgow and SANAD cohorts do not represent *direct* validation cohorts for the Australian cohort (Table 5.1). Thus, to allow evaluation of whether the five SNPs from this *k*NN model are relevant for treatment response in Glasgow and SANAD patients, the UK datasets were first stratified by age, gender, epilepsy type, response to AED treatment, and initial AED (Table 5.1). Initial AED was arranged into two groups, patients whose initial treatment was with either CBZ or VPA (n=118 and n=123 for the Glasgow and SANAD cohorts respectively), and those who were initially treated with one of the newer generation drugs such as GBP, LTG, or TPM (n=51 and n=123 for Glasgow and SANAD cohorts, respectively) (Tables 5.2-5.8). The latter group was known as the 'other AED treatment' group.

CBZ and VPA were the two most commonly prescribed initial AEDs in the Australian cohort (96% of the patients), whereas Glasgow patients were largely treated with either LTG or VPA (approximately 40% and 30% of total, respectively) and SANAD patients mostly received LTG or CBZ (approximately 30% and 25% of total, respectively) (Table 5.1). Stratification on the basis of initial AED thus controlled for this difference between the Australian cohort, when used as the training dataset, and the UK cohorts as test datasets. It also allowed determination of whether, the five *k*NN SNPs are universal markers of treatment response or selective for specific drugs (i.e. CBZ/VPA). For the purposes of the leave-one-out (n-1) cross-validation analysis, the Glasgow and the SANAD cohorts were additionally sub-divided into those initially treated with LTG (n=112 and n=97, respectively). The SANAD training and test datasets also included some patients treated with OXC (n=29 and n=12, respectively) and these were included in the CBZ group.

5.2.10. Australian *k*NN model validation in UK cohorts

In this analysis the original Australian cohort (n=115) was treated as the training dataset and each of the UK cohorts (Glasgow n=285, SANAD n=491) was used as independent test datasets for which predictions were made. This is a direct approach for model testing and was used in validation of the original Australian classifier using two independent Australian populations (Petrovski et al., 2009) (Figure 5.3, Figure 5.4).

5.2.11. Re-deriving the five-SNP kNN model in the SANAD cohort

This secondary analysis comprised a UK only prediction, with the aim of developing and testing a *k*NN model using UK training and test cohorts (70% and 30% respectively), in a similar manner to the original Australian model development (Petrovski et al., 2009). The SANAD training dataset was first investigated for association between the five SNPs and AED

response and subsequently used to predict response of the patients in the test dataset. This ensured that patients in training and test datasets are well-matched, particularly with regard to frequency of AED response. The *k*NN parameters used for this replication were identical to those employed in the original Australian study (i.e. five SNPs [N] and nine nearest neighbours [*k*]). Investigators running the model were blinded to the treatment responses of the test dataset.

5.2.12. *k*NN model validation in a UK population

To investigate whether the five SNPs were predictive of treatment response in the UK cohorts a leave-one-out approach (n-1) was adopted. Individual patients within each of the UK cohorts were classified using a *k*NN model built on a "leave-one-out" training dataset comprising the remaining samples (n-1) in that cohort to determine the overall performance of the five-SNP model. With this cross validation method, the *k*NN model used the genetic profiles of the remaining patients to predict the individual test patient that was left out, and this was then repeated for all of the Glasgow and SANAD patients. In addition to accounting for differences in response frequencies between the Australian and UK cohorts, this approach also eliminated any population genetic differences between the Australian and UK patients at these five SNPs. In the leave-one-out cross validation approach, the individual being predicted did not contribute to the training dataset prediction, thus ensuring model optimisation for the modestly sized UK cohorts whilst avoiding over-fitting.

5.2.13. Statistical analysis for assessing model performance

All model development and application processes were performed in Melbourne (Department of Medicine, University of Melbourne) and data analyses were performed in both Liverpool and Melbourne. Model assessment was performed with the *k*NN model using SAS Enterprise Miner software (SAS Institute Inc., Cary, NC) and SPSS version 18. Logistic regression was also performed for confirmation of the leave-one-out analysis results (SAS Enterprise Miner, SPSS version 18). For each model, the likelihood of predicting successful treatment outcome was determined by calculation of the odds ratio (with 95% confidence interval), positive and negative predictive values (with 95% confidence intervals), and Pearson's chi-squared p-value.

5.2.14. Biological significance of the Australian five-SNPs

In addition to the genetic analyses, each of the five SNPs was investigated to identify potential biological significance using several freely accessible online genomic databases. The HapMap website (www.HapMap.org, data source Rel#24/phase II Nov 08), dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/), Haploview version 4.1, and the UCSC Human Genome Browser (www.genome.ucsc.edu/cgi-bin/hgGateway) were used for extracting information on genomic structure (section 2.4.2), including LD structure (section 3.2.4). Regulatory changes were investigated through utilities available on these browsers and using online bioinformatics analysis tools, Fast SNP and TESS, for further functional and transcriptional analysis (Yuan et al., 2006) (section 3.2.7).

5.3. Results

5.3.1. Comparison of Australian and UK cohorts

Comparing the clinical characteristics of the Australian cohort (Petrovski et al., 2009) and the Glasgow and SANAD cohorts showed that both initial AED and drug response frequency were different between the Australian and UK patients (P <0.0001; Table 5.1). In addition, the SANAD cohort had a significantly higher frequency of unclassified epilepsy compared to both Australian and Glasgow cohorts (P <0.001), whereas there were no significant differences in epilepsy type between the Australian and Glasgow cohort.

5.3.2. Single SNP associations with treatment outcome

Of the 285 Glasgow patients, four failed genotyping at one or more of the five SNPs, leaving 281 available for subsequent analysis. Of the 520 SANAD patients, 491 were successfully genotyped for each of the five SNPs. Since the genotypes at the five SNPs employed in the multigenic model were proposed to predict or influence AED response, these were also independently tested for association in both UK cohorts to account for any single SNP influence on AED response. No association was identified with this independent SNP analysis (Table 5.2). A similar lack of association was noted in the original Australian cohort (Petrovski et al., 2009).

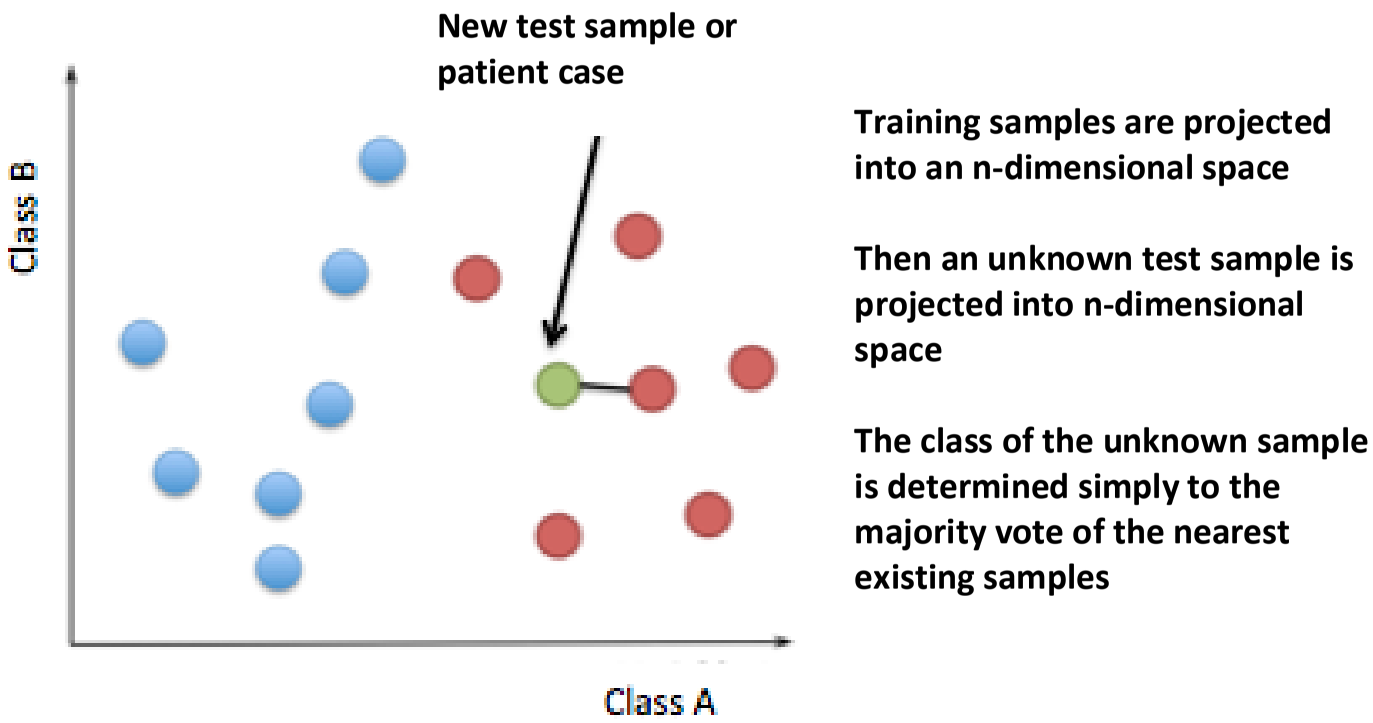


Figure 5.3 ***k*NN supervised machine learning approach**

Illustration of the nearest neighbour method used by the *k*NN supervised learning approach for classifying data

5.3.3. Results of the UK replication of the Australian five-SNP *k*NN model

The results of approach one, where both UK populations were treated as independent test datasets (and Australian patients used as the training dataset) for model validation showed that the Australian kNN classifier did not predict treatment response when applied directly to either UK cohort. A *k*NN model built on five SNP genotypes and drug response phenotypes from the original Australian cohort (n=115) failed to usefully predict treatment response, on the basis of genotype alone, in either the Glasgow (n=281) or SANAD (n=491) cohorts (both $P > 0.05$).

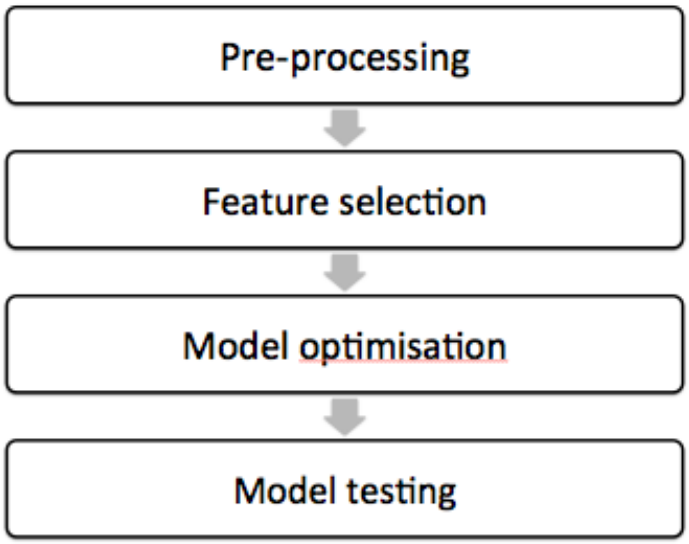


Figure 5.4 Summary of the general approach employed for the development of classification or predictive models using machine learning methods

Several basic steps encompass the development of predictive models. These often involve the pre-processing of large and/or complex data for the selection of the most influential variables and thereon model testing and parameter optimisation.

Table 5.1. Characteristics and comparison of the Australian cohort with the Glasgow and SANAD cohort

			Australian (n=115)	Glasgow (n=285)	SANAD (n=491)	*P*-value
Age	Mean (±SD)		43 (±19.7)	41 (±14.7)	39 (±18.3)	ns
Sex	N (%)	Male	61 (53.0%)	157 (55.1%)	269 (54.8%)	ns
		Female	54 (47.0%)	128 (44.9%)	222 (45.2%)	
Initial AED	N (%)	CBZ	66 (57.4%)	26 (9.1%)	123 (25.0%)*	<0.0001
		VPA	44 (38.3%)	92 (32.3%)	50 (10.2%)	
		LTG	2 (1.7%)	115 (40.4%)	139 (28.3%)	
		Other	3 (2.6%)	52 (18.2%)	179 (36.5%)	
Epilepsy type	N (%)	IGE	27 (23.5%)	92 (32.3%)	80 (16.3%)	<0.001†
		Focal	84 (73.0%)	185 (64.9%)	332 (67.6%)	
		UNC	4 (3.5%)	8 (2.8%)	79 (16.1%)	
Outcome at 12 months	N (%)	Responder	128 (71.9%)	152 (53.3%)	234 (47.7%)	<0.001
		Non-responder	50 (28.1%)	133 (46.7%)	257 (52.3%)	

*AED = antiepileptic drug, CBZ= carbamazepine, IGE= idiopathic generalised epilepsy, LTG= lamotrigine, N= number, ns= non-significant, SD= standard deviation, UNC= unclassified epilepsy, VPA= valproate. *Includes 35 patients initially treated with oxcarbazepine, †Statistical difference between SANAD cohort and both Australian and Glasgow cohorts*

5.3.4. Results of the UK cohort developed kNN classifier

This analysis used SNP genotypes and drug responses in the SANAD training dataset (n=343) to re-derive the *k*NN classifier, albeit using the original parameters, and to use it to predict treatment response, in a blinded manner, in a test dataset (n=148) of patients, also from the SANAD cohort. When using the SANAD training dataset (Table 5.3) to predict response in the SANAD test dataset, the *k*NN five-SNP classifier (using nine nearest neighbors) correctly identified 26 responders and 52 non-responders but incorrectly identified 26 non-responders as responders (false positives) and 44 responders as non-responders (false negatives) (P = 0.4; Table 5.3). Using only those patients that were initially prescribed either CBZ or VPA (n=50), the re-derived *k*NN classifier was internally predictive in the SANAD training dataset (Table 5.3), but again failed when applied to the SANAD test dataset, correctly classifying 10 responders and 14 non-responders but incorrectly identifying 10 non-responders as responders (false positives) and 16 responders as non-responders (false negatives) (P = 0.7; Table 5.3).

5.3.5. Results of the cross validation analyses

A "leave-one-out" cross validation analysis was performed in each of the two UK cohorts by predicting treatment response for each individual patient using five SNP genotypes in the remainder (n-1) of the respective cohort. The Glasgow cohort consisted of all successfully genotyped patients (n=281). Thus, for the Glasgow cohort, prediction was based on treatment response and SNP genotypes in a dataset comprising 280 patients. In the SANAD cohort, prediction was based on the training dataset of 342 patients (n=343 minus one). The 148 patients forming the SANAD test dataset were excluded from this analysis on the basis that investigators were blinded to treatment outcome in this sub-group.

Table 5.2. Uncorrected SNP genotype association with treatment outcome

		Glasgow cohort		**SANAD cohort***	
SNP	AED	*P* -value for trend	Odds ratio (95% CI)	*P*-value for trend	Odds ratio (95% CI)
rs2283170	CBZ/VPA	0.8	1.0 (0.6 – 1.9)	0.6	1.2 (0.7 – 1.8)
	LTG	0.6	1.1 (0.7 – 2.0)	0.2	0.7 (0.4 – 1.2)
	Other AED	0.004	0.3 (0.1 – 0.7)	0.8	1.1 (0.7 – 1.7)
rs2808526	CBZ/VPA	0.3	0.8 (0.4 – 1.3)	0.8	1.0 (0.7 – 1.6)
	LTG	0.4	1.3 (0.7 – 2.2)	0.7	0.8 (0.6 – 1.5)
	Other AED	0.9	1.1 (0.4 – 2.2)	0.3	0.8 (0.5 – 1.3)
rs4869682	CBZ/VPA	1.0	1.0 (0.6 – 1.7)	0.1	1.4 (0.9 – 2.1)
	LTG	0.3	0.8 (0.5 – 1.3)	0.3	0.8 (0.5 – 1.3)
	Other AED	0.8	0.9 (0.4 – 2.0)	0.5	1.2 (0.8 – 1.8)
rs658624	CBZ/VPA	0.2	1.4 (0.8 – 2.4)	0.5	1.1 (0.7 – 1.8)
	LTG	0.3	1.4 (0.8 – 2.3)	0.6	0.9 (0.6 – 1.4)
	Other AED	0.5	1.1 (0.6 – 2.8)	0.8	0.8 (0.5 – 1.3)
rs678262	CBZ/VPA	0.4	0.8 (0.5 – 1.4)	0.1	0.7 (0.5 – 1.1)
	LTG	0.07	0.6 (0.4 – 1.0)	0.5	1.2 (0.7 – 1.9)
	Other AED	0.2	0.6 (0.3 – 1.4)	0.3	1.3 (0.8 – 1.9)

*AED= antiepileptic drug, CBZ= carbamazepine, CI= confidence interval, LTG = lamotrigine, SNP= single nucleotide polymorphism, VPA = valproate, p-value for trend calculated by Cochran-Armitage test. Glasgow cohort numbers: CBZ/VPA=118, LTG=112, Other AED=51. SANAD cohort numbers: CBZ/VPA=173, LTG=139, Other AED=179, *Thirty five patients initially treated with oxcarbazepine were included in the carbamazepine group*

Table 5.3. Predictive performance of the 5-SNP *k*NN model on the SANAD test dataset (n=148) on the basis of initial AED

AED	n	TP	FP	TN	FN	PPV (95% CI)	NPV (95% CI)	Odds ratio (95% CI)	*P* -value
CBZ/VPA*	50	10	10	14	16	50% (23.7-76.3)	47% (25.2-69.4)	0.9 (0.28-2.72)	0.7
Other AED	98	16	16	38	28	50% (28.5-71.5)	58% (41.6-72.1)	1.4 (0.58-3.17)	0.3
Combined	148	26	26	52	44	50% (32.8-67.2)	54% (41.1-66.7)	1.2 (0.60-2.32)	0.4

*AED = antiepileptic drug, CBZ = carbamazepine, CI = confidence interval, kNN = k-nearest neighbour, NPV = negative predictive value, PPV = positive predictive value, SNP = single nucleotide polymorphism, VPA = valproate, TP = true positive (responders correctly classified as responders); FP = false positive (non-responders incorrectly classified as responders), TN = true negative (non-responders correctly classified as non-responders); FN = false negative (responders incorrectly classified as non-responder), *Twelve patients initially treated with oxcarbazepine were included in the carbamazepine group*

In the Glasgow cohort, the five SNP combination was found to be significantly predictive of treatment response in those patients initially prescribed either CBZ or VPA (positive and negative predictive values of 67% and 60% respectively, $P = 0.003$) but not those prescribed any other AED (LTG, $P = 0.3$; all other AEDs, $P = 0.8$; Table 5.4). In the SANAD cohort, the leave-one-out analysis showed a similar drug specific association. The five-SNP combination showed positive and negative predictive values of 69.1% and 55.6% in SANAD patients initially prescribed either CBZ or VPA ($P = 0.008$) and positive and negative predictive values of 57.4 and 60.5%, respectively, in those initially prescribed other AEDs (namely GBP, LTG or TPM; $P = 0.02$). The results indicate that these five SNPs are associated with treatment response in UK patients (Table 5.4), particularly when CBZ or VPA is used as the first AED, even though the independent *k*NN model (described in section 5.3.3) failed to have predictive value.

5.3.6. Logistic regression and permutation analysis for drug specific prediction

A logistic regression analysis was also performed to confirm the association identified in the leave-one-out cross-validation. A fully-fitted logistic regression model incorporating all five SNPs was built using the SANAD training dataset (n=343) which supported observations from the leave-one-out analysis that the five classifier SNPs are predictive of treatment response in patients initially prescribed CBZ or VPA. A regression model that was developed for patients from the SANAD training dataset who initially received CBZ or VPA (n=123) appeared to show a successful prediction of treatment response with positive and negative predictive values of 69% each ($P = 2.5\text{x}10^{-5}$) and model specificity and sensitivity values of 58% and 38.5%, respectively (Table 5.5). The regression model was less powerful when developed on patients prescribed other AEDs and when applied to the training dataset as a whole (n=343), with positive and negative predictive values of 63% and 59% ($P = 0.007$) and 61% and 58% ($P = 0.0006$), respectively (Table 5.5).

A permutation test for the fully fitted logistic model was subsequently performed to identify the likelihood of over-estimation caused by this analysis due to the limited numbers of patients (n=123) in the SANAD training dataset treated with either CBZ or VPA. The permutation results showed that randomised logistic regression models, using the same number of CBZ/VPA responders and non-responders would have been unlikely to achieve a p-value $< 2.5\text{x}10^{-5}$ based on five SNP profiles by chance ($P < 0.05$).

Table 5.4. Predictive performance of the 'leave-one-out' *k*NN approach in SANAD and Glasgow cohorts on the basis of initial AED

Cohort	AED	n	TP	FP	TN	FN	PPV (95% CI)	NPV (95% CI)	Odds ratio (95% CI)	*P* -value
Glasgow	CBZ/VPA	118	49	24	27	18	67% (51.8-79.6)	60% (40.6-76.8)	3.1 (1.4-6.6)	0.003
	LTG	112	35	30	25	22	54% (38.0-69.0)	53% (34.8-70.8)	1.3 (0.6-2.8)	0.3
	Other AED	51	15	17	8	11	47% (25.9-68.9)	42% (17.6-70.8)	0.6 (0.2-2.0)	0.8
SANAD	CBZ/VPA*	123	29	13	45	36	69% (48.6-84.3)	56% (41.3-69.0)	2.8 (1.3-6.1)	0.008
	LTG	97	18	12	40	27	60% (36.4-80.0)	60% (43.8-73.9)	2.2 (0.9-5.3)	0.06
	Other AED	123	21	17	52	33	55% (34.6-74.3)	61% (47.1-73.7)	1.9 (0.9-4.2)	0.06

*AED= antiepileptic drug, CBZ= carbamazepine, CI= confidence interval, kNN= k-nearest neighbour, LTG = lamotrigine, NPV = negative predictive value, PPV = positive predictive value, VPA= valproate, TP= true positive (responders correctly classified as responders), FP= false positive (non-responders incorrectly classified as responders), TN= true negative (non-responders correctly classified as non-responders), FN = false negative (responders incorrectly classified as non-responders), *Twenty nine patients initially treated with oxcarbazepine were included in the carbamazepine group*

Table 5.5. Predictive performance of the fully-fitted logistic regression model in the SANAD training dataset

AED	n	TP	FP	TN	FN	PPV (95% CI)	NPV (95% CI)	Odds ratio (95% CI)	*P* -value
CBZ/VPA*	123	49	22	36	16	69% (53.5-81.3)	69% (50.9-83.2)	5.0 (2.3-10.9)	2.5×10^{-5}
Other AED	220	26	15	106	73	63% (42.9-80.2)	59% (49.6-68.2)	2.5 (1.2-5.1)	0.007
Combined	343	66	42	137	98	61% (48.6-72.3)	58% (49.9-66.2)	2.2 (1.4-3.5)	0.0006

*AED= antiepileptic drug, CBZ= carbamazepine, CI = confidence interval, NPV= negative predictive value, PPV= positive predictive value, VPA= valproate, TP = true positive (responders correctly classified as responders), FP = false positive (non-responders incorrectly classified as responders) TN= true negative (non-responders correctly classified as non-responders), FN = false negative (responders incorrectly classified as non-responders), *Twenty nine patients initially treated with oxcarbazepine were included in the carbamazepine group*

5.4. Discussion

The exploration of high-level relationships between numerous genetic variants with minimal relative risks is characteristic of the current status of common complex trait investigations. The analysis of genomic data from such investigations has presented researchers with several unprecedented challenges (Kwan and Brodie, 2000a, Lee et al., 2008, McCarthy et al., 2008), including that of detecting multiple, small associations in high-dimensional data. Investigating numerous gene variants simultaneously is often not successful using existing mathematical and computational approaches. Making inferences based on the combination of several lower dimensional methods may not provide a correct understanding of real data occurrences. Moreover, important variants and/or other biological information may be obscured. Predictive models with the capacity to incorporate a collection of weak effects, along with their ability to model potentially complex interactions, offer an attractive alternative to multiple single SNP analyses (Lee et al., 2008).

The supervised classification learning method of data analysis is one form of statistical modeling applied to genomic data to obtain genomic prediction models for different groups of biological subjects. A previous "proof-of-concept" study developed a multigenic pharmacogenomics *k*NN model that successfully predicted response to initial AED treatment in an Australian cohort of patients with newly-diagnosed epilepsy. This was subsequently validated in two additional cohorts of patients, also from Australia (Petrovski et al., 2009). The *k*NN supervised classification learning approach was originally described by Fix and Hodges (Fix, 1951, Silverman and Jones, 1989) and has since become an important classification and clustering tool with diagnostic applications in a number of medical research fields. These include diagnostic and sub-class classification in cancer (Furey et al., 2000, Su et al., 2001, West et al., 2001, Crimins et al., 2005), immunoassay based anti-nuclear antibody tests (Binder et al., 2005), microarray experiments (Kim et al., 2004), drug toxicity (Martin et al., 2006), and rheumatoid arthritis (Liu et al., 2009).

The aim of the analysis described in this chapter was to assess the broader utility of the Australian multi-SNP model by applying it to two independently collected cohorts of patients with newly treated epilepsy from the UK. Definitions of response were adopted from those used in the development of the original model and patient cohorts were accordingly stratified by clinical characteristics. For each of the UK cohorts, the multigenic classifier failed to significantly predict response to the first well-tolerated AED when; i) the original Australian cohort was used as the training dataset, and ii) when the classifier was re-driven in UK patients alone. The failure of direct replication was not entirely surprising. Possible explanations include differences in drug response frequencies, differences in phenotypic definitions and

methods of ascertainment, genomic population differences, differing drug policies, failure to re-calibrate the kNN parameters, or a false positive signal in the original study. These are discussed briefly below.

5.4.1. Importance of drug response frequencies

As discussed previously, classification algorithms involve a training dataset of known outcomes on which a predictive model is built and, from this, new predictions or classifications can be inferred. The *k*NN approach positions the training dataset in an n^{th} dimensional space within which new cases/data can be placed and subsequently assigned a value or classification. Thus, the frequency of outcomes, in this case response or non-response to AED treatment, within the training dataset can affect the classification of any new cases that are presented. The *k*NN model using the Australian training dataset (identified as having fewer cases of treatment failure than UK cohorts) was thus expected to result in an over-estimation of the number of responders (false positives) in the Glasgow and SANAD cohorts, thereby affecting model classification reliability. This discordance in treatment response was arguably the most significant confounder in the attempt to directly validate the original multi-genic classifier.

5.4.2. Differences in data collection and treatment response classification

The Australian cohort constituted a series of newly diagnosed patients enrolled into a prospective PGx study at first clinical presentation. In contrast, the SANAD cohort comprised a sub-set of randomised clinical trial patients (believed to be representative of the trial population as a whole) who were belatedly consented for the donation of DNA and whose clinical information, albeit prospectively collected, was extracted from a trial database that was not designed with a PGx study in mind. The Glasgow cohort comprised a variety of individuals attending outpatient clinics and participating in randomised clinical trials, who were retrospectively consented for donation of DNA and whose clinical information was not collected in a systematic manner. These differences in recruitment and data collection procedures may have introduced inconsistencies in the classification of responder and non-responder status, particularly in the UK cohorts where the clinical information was not specifically collected for PGx purposes. The principal concern in this regard is a lack of sensitivity to exclude seizures occurring in the drug titration period or arising from non-compliance or acute provocation (i.e. sleep deprivation or alcohol misuse) during the first 12 months of follow-up. This might explain, at least in part, the significantly greater frequency of non-response in the Glasgow and SANAD cohorts.

5.4.3. Genomic population differences between UK and Australian cohorts

Another important consideration with regard to study design was the assumption that the five *k*NN SNP markers originally identified in an Australian population would extrapolate directly to UK patients. Although most Australian patients were considered to be of European descent, and patients who self-identified as being of non-European ancestry were excluded, it is possible that subtle ethnic differences existed between the cohorts. Under such circumstances, a discrete set of genetic variants might be more weakly associated with the trait of interest (or with unidentified causal variants) in one population than in another.

5.4.4. Differences in initial AED treatment between cohorts

A clear difference between the Australian and UK cohorts was observed in the relative frequencies of individual AEDs used as initial treatment. The Glasgow cohort was largely recruited via a drug trial comparing VPA and LTG (Stephen et al, 2007), with almost 50% of Glasgow patients initially exposed to LTG. The SANAD cohort also showed a high proportion of patients receiving LTG as the first well-tolerated AED, which was unsurprising given that this drug was included in both arms of the SANAD trial (Marson et al., 2007a, b). In contrast, 96% of the Australian cohort was initially prescribed either CBZ or VPA. These simple differences in drug treatment policy may be sufficient to explain the failure to directly validate the *k*NN model in UK patients.

5.4.5. Adapting model parameters of the Australian *k*NN five-SNP classifier for UK replication

An additional confounder may have been the use of *k*NN model parameters that were employed in the original Australian study. These were accordingly derived from a cohort that was significantly smaller (n=115) than either of the two UK cohorts employed in the current analysis. Failure to re-calibrate the *k*NN parameters in order to accommodate larger training datasets with differing response frequencies may have impacted on the accuracy of treatment response prediction. In the *k*NN classifier, the number of nearest neighbours (*k*) by which classification occurs, is unsurprisingly dependent on the size and phenotype frequency of the training dataset and prediction is based on a simple majority phenotype amongst those nearest neighbours. In a larger training dataset, it is possible that fewer nearest neighbours would be required for accurate prediction in test datasets. For example, where the nearest four neighbours are responders and next nearest five neighbours are non-responders, using k=9 would result in the prediction of non-response, whereas using $k \leq 7$ would result in the

prediction of response. Failure to re-calibrate *k*NN parameters despite differences in treatment response between the Australian and UK cohorts was a potentially significant limitation.

5.4.6. Unreliability of the original Australian association

Finally, there is also a possibility that the original findings in the Australian cohort constituted a false positive signal. This seems somewhat unlikely, given the clear association in this analysis between 5-SNP genotype and response to treatment with either CBZ or VPA in UK cohorts when analysed using both a "leave-one-out" method and a fully-fitted logistic regression. This finding, together with the validation in two independent Australian epilepsy cohorts reported in the original study, lends weight to the significance of these SNP genotypes as biomarkers of response to initial drug therapy in newly treated epilepsy. Whether the classifier is truly specific for CBZ and VPA alone or is indicative of treatment responsiveness in general remains to be determined. Making the distinction will require significantly larger cohorts of patients and a more consistent approach to recruitment and data collection.

5.4.7. Cross-validation validated the predictive capacity of the five SNPs comprising the Australian *k*NN classifier

Despite failure of direct validation of the kNN classifier, a further attempt was made to explore the significance of these five SNPs in UK patients. This was performed using a cross-validation "leave-one-out" approach where each of the UK cohorts acted as their own training dataset. This negated many of the confounders described above that could have potentially impacted on findings of the direct validation method. The analysis was again stratified by initial AED in an effort to determine whether the five SNPs originally identified in the Australian cohort were selectively predictive for CBZ or VPA as initial treatment (96% of Australian patients received these drugs as first ever AED). This approach proved successful, with the "leave-one-out" analysis indicating that the five SNPs had a collective predictive capacity for both Glasgow and SANAD patients treated with either CBZ or VPA but not other AEDs. A subsequent permutation test confirmed that the randomized, fully fitted logistic regression model developed on the SANAD training dataset, and using the same number of CBZ/VPA responders and non-responders, would have been unlikely to achieve a p-value $< 2.5 \times 10^{-5}$. This suggested that the association with CBZ/VPA treatment response in a logistic regression model was unlikely to have occurred by chance.

5.4.8. Biological significance of the five-SNPs

The relative success, when using the "leave-one-out" approach, of these five SNPs in predicting response to CBZ or VPA as the initial AED in UK patients suggests that they possess biological significance. The SNPs were originally identified from an initial panel of 4,041 SNPs across 279 candidate genes, selected on the basis of a known or putative involvement in epilepsy susceptibility, a high expression level in the brain, or an involvement in AED pharmacology (Petrovski et al., 2009). The biological investigation that followed our genetic analyses showed that the variants were comprised of two SNPs in the *SCN4B* gene (encoding the β_4 subunit of the voltage-gated sodium channel) and one each in the *GABBR2* gene (encoding $GABA_B$ receptor subunit 2), *KCNQ1* gene (encoding the $K_v7.1$ subunit of the delayed rectifier potassium channel), and *SLC1A3* gene (encoding excitatory amino acid transporter 1) (Table 5.6). Two of the genes (*SCN4B* and *KCNQ1*) are reported to have limited expression in brain tissue (Waldegger et al., 1999, Yu et al., 2003), all five SNPs are located in intronic regions of their respective genes (Kent et al., 2002), and investigation of the genomic structure using www.hapmap.org (release # 24) failed to identify any biologically functional variants with a minor allele frequency $\geq$ 1% in European populations that were in strong linkage disequilibrium ($r2 \geq 0.8$) with any of these specific SNPs. Further bioinformatics analysis using online tools for functional analysis (Yuan et al., 2006) and assessment of TFBSs (Kent et al., 2002) were similarly unremarkable, although all SNPs except rs678262 (in *SCN4B*) were shown to be located in TF binding domains. The rs2283170 SNP in *KCNQ1* was additionally predicted to effect an alteration in TF binding characteristics. As such, the functional significance of these SNPs and the explanation for their association with treatment response in newly treated epilepsy remains unclear. The fact that these SNPs were selectively predictive for response to CBZ and VPA but no other AEDs with similar mechanisms of action also remains unexplained.

Table 5.6. Genomic information for the five-SNPs comprising the *k*NN classifier

Gene	SNP	Location	Alleles	Amino acid change	HapMap MAF
KCNQ1	rs2283170	Intron	A>G	-	0.339
GABBR2	rs2808526	Intron	A>G	-	0.450
SLC1A3	rs4869682	Intron	G>T	-	0.460
SCN4B	rs658624	Intron	T>C	-	0.458
SCN4B	rs678262	Intron	G>C	-	0.346

MAF = minor allele frequency, SNP = single nucleotide polymorphism

5.4.9. Summary

In summary, this analysis suggests that the "proof of concept" *k*NN model, developed in an Australian cohort of newly-diagnosed epilepsy, is not directly applicable to other epilepsy populations, even those that might be considered ethnically comparable. The model has multiple limitations when applied to populations that differ from the one used in its construction, particularly where response frequencies, drug policies, phenotype determination and methods of ascertainment differ. Nevertheless, the combination of the five SNPs reported in the original Australian study does appear to have a collective influence in predicting response to treatment with either CBZ or VPA in UK patients. This observation, although drug specific, should encourage additional replication attempts with larger cohort sizes to better understand the potential of this multi-SNP model as a biomarker of early seizure control in new-onset epilepsy patients treated with these two drugs.

CHAPTER SIX

APPLICATION OF MACHINE LEARNING APPROACHES TO THE DEVELOPMENT OF MULTIGENIC CLASSIFIER MODELS FOR PRIMARY GENERALISED EPILEPSIES

CONTENTS

6.1. Introduction

As discussed in Chapter 1, epilepsy is a complex and heterogeneous disorder, for which rare and common genetic forms exist. Common genetic epilepsies are, clinically and genetically, a heterogeneous group of complex seizure disorders. Thus, defining the genetic contribution to common epilepsy syndromes has proven to be a formidable task (Dibbens et al., 2007, Rees, 2010). In Chapter 5, a ML approach was used to build PGx classifiers to predict responsiveness to AED treatment. The study described in this results chapter applied this previously used statistical methodology to analyse available genomic and clinical phenotype data for individuals with common genetic forms of epilepsy in order to investigate the genetic contribution to these complex syndromes.

6.1.1. Complex genetic forms of epilepsy

It is estimated that there is an underlying genetic predisposition for epilepsy in approximately half of individuals (idiopathic epilepsies), with monogenic epilepsies accounting for less than 1 percent (Kearney, 2012). The remaining majority are idiopathic or primary generalised epilepsies (PGEs) and non-acquired focal epilepsies (NAFE) that have a strong genetic basis with a complex inheritance pattern in which multiple and environmental factors contribute to epilepsy risk, though these complex genetic epilepsies are poorly understood (Tan et al., 2004).

6.1.2. Primary generalised epilepsy syndromes

The PGEs classically fall into several common and rare recognisable sub-syndromes. Rare IGE syndromes include Benign Myoclonic Epilepsy of Infancy (BMEI), Early Onset Absence Epilepsy, Myoclonic Astatic Epilepsy (MAE), Epilepsy with Myoclonic Absences, Eyelid Myoclonia with Absences and Absence Status Epilepsy (Gardiner, 2005). The common PGEs sub-syndromes are characterised by some or all of the three following seizure types; typical absence seizures, myoclonic jerks and generalised tonic-clonic seizures (GTCS), which can occur in different combinations but typically with one seizure type predominating (Engel, 2006a). Childhood Absence Epilepsy (CAE), Juvenile Absence Epilepsy (JAE), Juvenile Myoclonic Epilepsy (JME) and Epilepsy with Generalised Tonic-Clonic Seizures (GTCS) represent the four more common PGE sub-syndromes. (Shneker and Fountain, 2003, Engel, 2006a).

6.1.3. Classifying complex inheritance or common PGE sub-syndromes

Common PGE has a typical electroencephalographic (EEG) pattern with paroxysms

of generalised spike and wave and polyspike discharges, which is the hallmark of the syndrome and the onset of these common PGEs is usually before the age of 16 (Sander, 2003b). Several features confound the genetic analysis of the more common PGE sub-syndromes (Gardiner, 2005). In particular, each sub-syndrome are themselves heterogeneous in their phenotypic presentation and are often found to overlap (Sander et al., 2000). Different PGE sub-syndromes can be found within a single pedigree (Gardiner, 2005) and different generalised seizure types may emerge in the same patient over time, adding further complication (Sander et al., 2000). CAE typically begins between 4 and 10 years of age (Crunelli and Leresche, 2002, Gardiner, 2005). The main seizure type in CAE is typical absence seizures but, in about 50% of patients, GTCS can also occur although very few individuals additionally experience myoclonic jerks (Sander, 2003b). JME represents 5–10% of epilepsy as a whole and individuals most commonly present between the ages of 8 and 26 with myoclonic jerks predominantly of the upper limbs (Greenberg et al., 1992). Over 90% also have GTCS and 30% have typical absences (Crunelli and Leresche, 2002). This overlap in seizure types suggests commonality in genetic predisposition between the PGE sub-syndromes (Janz et al., 1992, Sander et al., 2000). Age at onset and main seizure type are thus used to classify the more common PGEs into the four main sub-syndromes (Panayiotopoulos and International League against Epilepsy., 2005). Despite these distinct features however, accurate diagnosis of the sub-syndromes is not always possible from the first presentation and so a number of patients with PGE are often difficult to classify (Gardiner, 2005).

6.1.4. Genetic studies for primary generalised epilepsy; the picture so far

Since most of the individual genes involved in complex disorders and/or traits are thought to only have a small impact on the clinical phenotype, their identification has presented a major challenge for disease genetics (Hirschhorn et al., 2002). PGEs are similarly thought to arise from additive or interactive effects of more than one susceptibility gene (Dibbens et al., 2007) and so progress in identifying the underlying genetic causes, like most common, complex traits, has been slow (Dibbens et al., 2007, Frankel, 2009). Currently ~20 genes are known to cause Mendelian forms of human epilepsy (Robinson and Gardiner, 2004) and, as might be expected for a disorder of neuronal hyper-excitability, at least two thirds of these encode ion channels (Frankel, 2009). This research suggested ion channel defects as a common pathogenic pathway in a multitude of epilepsies and led to the hypothesis of epilepsies being channelopathies (Berkovic et al., 2006). Although no directly causative ion channel genes have been identified for complex PGEs, this channelopathy concept has provided important positional clues for the pathogenesis of several common PGE sub-syndromes (Gardiner, 2005, Dibbens et al., 2007).

Table 6.1 provides a summary of several of these key genetic studies.

Ion channel targets of current genomic studies have included voltage-gated channels Na^+, K^+, Ca^{2+}, and Cl^- and the ligand-gated channels; nicotinic acetylcholine and $GABA_A$ receptors (Gardiner, 2005). The most important epilepsy gene to be discovered to date is thought to be the previously discussed *SCN1A* sodium channel gene (Chapter 4). *SCN1A* is the most prevalent gene causing Mendelian forms of epilepsy and is the most studied in epilepsy. *SCN1A* mutations have been identified in the familial syndrome of Generalised Epilepsy with Febrile Seizures Plus ($GEFS^+$) as well as severe myoclonic epilepsy (SMEI) or Dravet Syndrome (Gardiner, 2005, Mulley et al., 2005). $GEFS^+$ is a heterogeneous autosomal dominant disorder (also recognised as a complex epilepsy phenotype) in which family members exhibit multiple epilepsy phenotypes including absence, myoclonic, generalised tonic-clonic or partial seizures, as well as febrile seizures (FS), with the FS being phenotypically simple or complex (Mulley et al., 2005). SMEI is also a FS disorder but with a more severe phenotype. Mutations for $GEFS^+$ were first identified in *SCN1A*, and these were shown to alter amino acids within the voltage-sensing S4 segments of the channel, the functional effects of these were altered channel inactivation and a persistent inward sodium current (Gardiner, 2005).

Most of the mutations associated with SMEI are however more critical as they introduce a stop codon with truncation of the protein and predicted to have a loss of function. The *SCN1A* gene has not only lead the way in channelopathy research for the genetic epilepsies, currently presenting the only definitive marker for a phenotype of epilepsy (SMEI), but has also provided insight into the overlap and multigenetic complexity that can underlie these common forms of genetic epilepsies. Another important candidate in PGE is the genes of the $GABA_A$ receptor subunit gene(s), namely *GABRG2* and *GABRA1* (Baulac et al., 2001, Gardiner, 2005, Rees, 2010). The subunits encoded by GABRG2 and GABRA1 are associated with monogenic forms of IGE and also the $GEFS^+$ phenotype. Mutations in the GABA receptor genes have also been implicated with JME (Gardiner, 2005, Rees, 2010).

Table 6.1. Genes implicated in complex idiopathic generalized epilepsies

Summary of genes associated with complex idiopathic generalised epilepsies so far. Genes are presented for the sub-syndromes of juvenile myoclonic epilepsy, childhood absence epilepsies and generalised epilepsies with febrile seizures plus. Data extracted from Huber *et* al 2009 and Rees *et* al 2010

Gene	Chromosomal localisation	Complex Epilepsy phenotype
SCN1A	2q24	$GEFS^+$
SCN1B	19q13.1	$GEFS^+$
CACNB4	2q22–23	JME
CACNA1H	16p13.3	CAE
BRD_2	6p21.3	JME
$CLCN_2$	3q26	CAE
EFHC1	6p12-p11	JME
GABRA1	5q34	JME, CAE
GABRG2	5q34	CAE, $GEFS^+$
GABRD	1p36.3	$GEFS^+$
GABRB3	15q11.2	CAE
EFCH1/*myoclonin 1*	6p12	JME
ME_2	18	JME

SCN1A = sodium channel α1 subunit, SCN1B= sodium channel β1 subunit, CACNB4 = voltage-dependent calcium channel β4, CACNA1H = voltage-dependent T-type calcium channel α1H, BRD_2 *= bromodomain containing protein 2,* $CLCN_2$ *= chloride channel gene 2, GABRA1 =* $GABA_A$ *receptor α1 subunit, GABRG2 = GABA(A) receptor Ƴ2 subunit, GABRD =* $GABA_A$ *receptor δ subunit, GABRB3 =* $GABA_A$ *receptor β3 subunit, EFCH1 = protein with an EF-hand,* ME_2 *= malic enzyme 2,* $GEFS^+$ *= generalised epilepsy with febrile seizures plus, JME = juvenile myoclonic epilepsy, CAE = childhood absence epilepsy*

CACNA1H encodes the T-type calcium channel that is critically involved in the thalamo-cortical network (Mulley et al., 2003, Mulley et al., 2005). *CACNA1H* has additionally been studied extensively in regards to complex PGE genomics and rare variants altering ion channel properties of the encoded T-type calcium channel protein have been observed in patients with PGE in several studies (Chen et al., 2003, Heron et al., 2004, Heron et al., 2007). Variants in *CACNA1H*, were initially associated with CAE (Chen et al., 2003) but the relationship has since been extended to other epilepsy phenotypes (Heron et al., 2004). Variation in *CACNA1H* has consistently been shown to contribute to the pathogenesis of epilepsies with complex genetics, but no variants in *CACNA1H* have been described that are sufficiently pathogenic to cause epilepsy on their own (Dibbens et al., 2007).

As in genetic investigations of drug response in epilepsy (section 1.5), several loci and variants have been studied as possible candidates for complex PGE syndromes but the majority have yielded negative results, with few being pursued any further (Steinlein, 2004, Tan et al., 2004, Kearney, 2012). In those reporting an initial positive association, replication studies have invariably failed to confirm the relationship (Steinlein, 2004, Tan et al., 2004). Failures, particularly in replication studies, are deemed to be due to the inherent effects of phenotypic variability, complex inheritance and genetic heterogeneity (Tan et al., 2004, Dibbens et al., 2007). One recognised assumption is that because of the polygenic nature of PGE, an individual most likely develops PGE only if sufficient variation is present (Dibbens et al., 2007). Under this hypothesised model, only a subset of a large population of susceptibility variants needs to be present. However, to explore this hypothesis requires association studies with much larger sample sizes than currently employed (Dibbens et al., 2007), perhaps in the order of thousands or tens of thousands of subjects (Mulley et al., 2005, Mullen et al., 2009, Ferraro et al., 2012). This lack of power is considered a major issue for complex disease genetic association studies in general (Weller et al., 2006, Mullen et al., 2009).

6.1.5. Recent advancements in disease genomics

Greater success in the identification of genes for PGE may be achieved with unbiased genome-wide surveys in large study populations (Kearney, 2012). Human genomics has moved towards a whole genome approach in the investigation of the genetic architecture of complex traits in an effort to resolve the lack of power or resolution in traditional genetic linkage studies (sections 1.4.9 and 1.4.10) (Ferraro et al., 2012, Kearney, 2012). GWA studies for complex generalised epilepsies have revealed significant linkage at the loci 2q34 and 13q31.3 for myoclonic and absence seizures, respectively (Ferraro et al., 2012, Kearney, 2012, Leu et al., 2012).

The role of CNVs as rare forms of variation has been extensively investigated and has shown a collective contribution to the etiology of a variety of common neurological diseases with complex genetics and has also been implicated in several types of epilepsy. More recently, CNV hotspots have been investigated in the non-Mendelian genetic epilepsies (Mefford and Eichler, 2009, Mefford et al., 2010, Sisodiya and Mefford, 2011, Kearney, 2012). Microdeletions in the chromosomal region 15q13.3 encompassing the *CHRNA7* gene (coding for the alpha-7 subunit of the nicotinic acetylcholine receptor) were identified in approximately 1% of 1,223 PGE patients (Helbig et al., 2009). As such, 15q13.3 microdeletions appear to constitute the most prevalent risk factor for PGEs identified to date (Helbig et al., 2009, Mulley and Dibbens, 2009). In line with current genomic research (section 1.5), recent results from epilepsy genetics have stimulated interest in assessing the contribution of both rare and common variants to the aetiology of epilepsy through the utility of whole-exome and whole-genome sequencing in individual patients and this is the next step in examining the basis of epilepsy genomics (Ferraro et al., 2012, Kearney, 2012).

6.1.6. A machine learning approach to disease or phenotype classification

The univariate nature of linkage mapping, candidate gene analysis and SNP-based disease association studies does not adequately account for the genetics of PGE syndromes, which have so far proven to be too heterogeneous for the detection of strong associations. ML, as described in Chapter 5, is an alternative and efficient way for extracting hidden information in a given dataset (Lee et al., 2008). The main advantages being, i) its ability to extract relationships from high-dimensional data, and ii) efficiency in simultaneously analysing, numerous, often highly interactive variables, of small effect. This may, in turn, allow greater analytical power in cohorts of limited sample size (Lee et al., 2008).

Several steps are involved in the appropriate application of ML approaches (Figure 6.1, 6.2). The processes prior to model development are feature selection for identifying and removing as many redundant variables as possible (Yu & Liu, 2004) and instance selection for filtering noisy data (Kotsiantis, 2007, Derrac et al., 2012). These processes enable models to operate faster and more effectively (Larranaga et al., 2006, Kotsiantis, 2007, Joaqu et al., 2010).

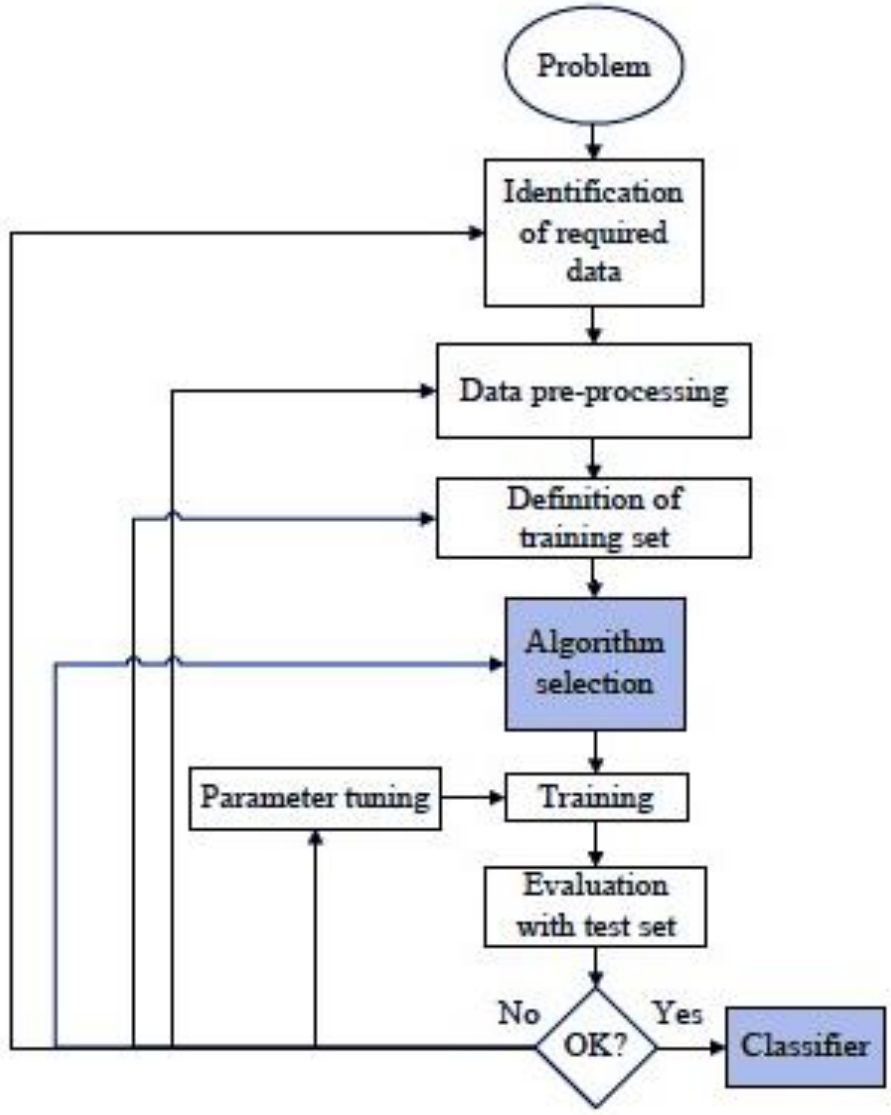


Figure 6.1 An overview of the supervised machine learning process

The supervised approach to machine learning entails the input of data with known values for model development or training. Figure reproduced from Kotsiantis *et* al 2007.

6.1.7. Development and assessment of machine learning models

Once a ML model has been built, it requires evaluation and interpretation and this forms the next stage in model development (Larranaga et al., 2006). The final overall performance of a learning method relates to its prediction capability on independent test data (Hastie et al., 2001). Assessment of this performance is extremely important in practice, since it guides the choice of model and gives a measure of the quality of the chosen model (Hastie et al., 2001). This requires estimating the expected test error and/or prediction accuracy (the percentage of correct predictions divided by the total number of predictions) for a model. From this, the performance of different models can be estimated in order to choose the best one, with prediction error usually estimated on new unseen datasets (Hastie et al., 2001).

The best approach for model development and evaluation is to separate the dataset into three randomly divided individual patient datasets; (i) a training dataset, (ii) a validation dataset, and (iii) a test dataset (Hastie et al., 2001, Mansmann and Winkelmann, 2002, Kotsiantis, 2007, Lee et al., 2008). The training dataset is used to fit the models, the validation dataset is used to estimate prediction error for model selection and optimization, and the test dataset is used for assessment of the generalisation error of the chosen model (Mansmann and Winkelmann, 2002, Kotsiantis, 2007, Petrovski et al., 2009). The test dataset is usually kept hidden and used only at the end of the data analysis for a true, unbiased test of prediction accuracy (Hastie et al., 2001). The number of observations in each of the three parts can depend on the signal-to-noise ratio in the data and the size of the training dataset (Hastie et al., 2001).

Additional techniques that can be used to partition data for initial model development include cross-validation (CV) and leave-one-out CV (e.g. the *k*NN ML method described in Chapter 5). CV is the simplest method for estimating prediction error and involves dividing the training dataset into mutually exclusive and equally-sized subsets, with each individual subset subsequently trained on the composite of all other subsets (the average error rate of each subset is therefore an estimate of the error rate of the model) (Hastie et al., 2001, Mansmann and Winkelmann, 2002, Kotsiantis, 2007). In the latter CV method, all test subsets consist of a single instance (Kotsiantis, 2007).

6.1.8. Application of machine learning to complex disease genetics

A number of classification-based ML methods are available (Kotsiantis, 2007). These include logical or symbolic techniques, such as classification trees, decision learning trees (DLTs), perception-based techniques such as NNs, and statistical techniques such as *k*NN and SVM (Kotsiantis, 2007)(Table 6.2). Several of these MLAs, particularly MDR (Ritchie and Motsinger, 2005, McKinney et al., 2006), NN (Lucek and Ott, 1997, Motsinger et al., 2006),

RF (Yoon et al., 2003, Bureau et al., 2005) and SVM (Yoon et al., 2003, Yu and Shete, 2005), have previously been used for the study of multi-locus association traits (Dinu et al., 2007), including diabetes, coronary heart disease, alcoholism and breast cancer (Yu and Shete, 2005, Bhaskar et al., 2006b, Silva et al., 2011).

6.1.9. The *k*NN machine learning approach can successfully identify high-order patterns in complex disease traits

The Australian multigenic pharmacogenomic classifier study, described in Chapter 5, exemplifies a recent and successful attempt at applying ML approaches to a large amount of genetic data (Cavalleri et al., 2007, Petrovski et al., 2009) namely common variation in the form of SNPs, to predict AED response in epilepsy patients. The *k*NN classifier is one of the most well-known classifiers that is based on the instances contained in the training dataset (Cover and Hart, 1967, Joaqu et al., 2010). Thus, the effectiveness of the classification process relies on the quality of the training data (Joaqu et al., 2010). Its main drawback is its relative inefficiency when the size of the dataset to be used in the modeling process increases (Kotsiantis, 2007). Instance and feature selection, which aid data-reduction, are thus also commonly used alongside the *k*NN ML algorithm. Thus, while supervised learning approaches have previously been reported to obtain reliable results in pharmacogenetics (Petrovski et al., 2009), they have not as yet been used to identify genetic predictors of epilepsy or epilepsy syndromes, such as PGE.

6.2. Purpose of investigation

A number of patients with PGE are often difficult to classify (Reutens and Berkovic, 1995). There are sub-syndromes of JME in which patients present not just with myoclonic jerks but also with or without typical absence seizures and GTCS (Gardiner, 2005). CAE is mainly characterised by typical absence seizures that persist into adolescence (Crunelli and Leresche, 2002, Gardiner, 2005), but GTCS can emerge in a significant percentage (up to 90%) of CAE cases when the absence seizures persist into adulthood (Crunelli and Leresche, 2002). CAE can also evolve into JME, with an estimated 18% of all JME patients having an initial diagnosis of CAE (Delgado-Escueta, 2007). With this overlap in seizure types, distinguishing between PGE sub-syndromes can be problematic and making a precise clinical diagnosis may not be possible at the first presentation. Accuracy in classification is however important for the correct prognosis of individual PGE patients and for initiation of the correct treatment, particularly as several AEDs have been shown to exacerbate specific seizure types (Bergey, 2005, Beydoun and D'Souza, 2012).

A recent attempt has been made to reduce the inherent heterogeneity in non-Mendelian PGEs using neurobiologically defined traits, such as seizure types rather than syndrome categories, for the purposes of genetic study (Greenberg and Subaran, 2011, Ferraro et al., 2012). This approach may help cut across phenotypically complex PGE syndromes and facilitate identification of the underlying susceptibility genes (Greenberg and Subaran, 2011). Patients can be segregated into groups according to strict demographic and/or clinical categories (i.e. age at onset, gender, and seizure type) for studying genetic variants (Ferraro et al., 2012). Each of these factors may have unique genetic signatures. Separating patients into clinical categories that allow such factors to be analysed independently may be a better approach than lumping people into often arbitrarily assigned syndromic sub-groups (Ferraro et al., 2012).

6.2.1. Aims

The aim of the study described in this chapter was to apply ML approaches to a large cohort of patients with newly-diagnosed, complex, non-Mendelian PGEs in an effort to identify the underlying genetic signature of these common epilepsy syndromes, and thereby aid in sub-syndromic diagnosis. The potential biological significance of any identified variants was also explored. The specific research objectives were:

Objective 1: to use ML algorithms to build predictive models for the identification of genes and/or gene variants as potential markers for the differentiation of individuals presenting with CAE, JAE and JME on the basis of seizure types.

Objective 2: to use ML algorithms to build predictive models for the identification of genes and/or gene variants as potential markers for the differentiation of individuals with PGE and focal epilepsies (or LREs).

In each case, the *k*NN ML algorithm (Petrovski et al., 2009) (Chapter 5) was employed. An additional aim was to explore a number of other ML approaches and identify the ML approach with the best overall performance when applied to this particular genomic dataset.

6.3. Methods

6.3.1. Study populations

Clinical and genetic data from two independent cohorts of newly treated PGE patients were employed in the analysis; SANAD study patients (Marson et al., 2007a, b) and a cohort recruited at two epilepsy centres in Australia (The Royal Melbourne Hospital in Melbourne and Austin Health in Heidelberg) (Cavalleri et al., 2007, Petrovski et al., 2009). Genotype and clinical data was available for a total of 436 patients with newly treated PGE, 296 from Australia and 140 from the UK (Table 6.3). For Objective 2, LRE patients acted as non-PGE controls, with a total of 760 LRE patients (628 SANAD and 132 Australian) possessing sufficient clinical and genetic data for initial inclusion (Table 6.4). Clinical information on PGE patients was extracted from hospital notes and existing databases in order to identify syndromes, sub-syndromes and seizure types.

Table 6.3 Characteristics of UK and Australian patient cohorts for PGE (seizure classification cohort)

			Australian (n=136)	**UK (n=68)**	**Total (n=204)**
Age at randomisation	Mean (±SD)		12 (±8.27)	19 (±11.6)	14 (±10.0)
Sex	n (%)	Male	59 (43.4%)	35 (51.5%)	94 (46.1%)
		Female	77 (56.6%)	33 (48.5%)	110 (53.9%)
Epilepsy syndrome	n (%)	CAE/JAE	94(69.1%)	32 (47.1%)	126 (61.8%)
		JME	42(30.9%)	36 (52.9%)	78 (38.2%)

SD = standard deviation, CAE = childhood absence epilepsy, JAE = juvenile absence epilepsy, JME = juvenile myoclonic epilepsy

Table 6.4 Characteristics of UK and Australian newly treated epilepsy patient cohorts for classification of epilepsy type

			Australian (n=428)	UK (n=189)	Total (n=617)
Age at randomisation	Mean (±SD)		13 (±8.4)	28 (±17.4)	23 (±18.1)
Sex	n (%)	**Male**	192 (64.8%)	104 (35.1%)	296 (48.0%)
		Female	236 (73.5%)	85 (26.5%)	321 (52.0%)
Epilepsy type	n (%)	**PGE**	296 (69.3%)	131 (30.7%)	427 (69.2%)
		LRE	132 (71.7%)	58* (31.5%)	184 (29.8%)

** Removal of 570 patients from total LRE available to maintain consistent ratio of Australian to UK patients with LRE*

6.3.2. Phenotyping and patient inclusion for objective one

For objective 1, all patients were classified into the following 2 groups, based on seizure type: Group 1 - patients with typical absence seizures (with or without GTCS) but not myoclonic jerks, and Group 2 - patients with myoclonic jerks (with or without GTCS) but not typical absence seizures. Patients with GTCS alone were excluded (n=88) (Table 6.3). Of the total PGE cohort (n=436), individuals were also excluded from the analysis if they exhibited both myoclonic jerks and typical absence seizures (n=71), if they exhibited both focal and generalised seizure types (n=7), or if there was evidence of CAE later evolving into JME (n=60). Finally, six additional patients were removed at the QC stage due to inadequate phenotype data and/or missing genotypes. In Group 1, no distinction was made between patients diagnosed with either CAE or JAE. The remaining 204 individuals (136 Australian, 68 SANAD) thus had PGE characterised by either myoclonic jerks (n=78) or typical absence seizures (n=126), with or without GTCS (Table 6.3).

6.3.3. Phenotyping and patient inclusion for objective two

For objective 2, all patients considered for inclusion in the study had PGE or LRE, with unclassified epilepsies excluded. Of the 436 available PGE patients, 427 (296 Australian, 131 SANAD) were included in the analysis, with 9 patients again removed after QC due to

inadequate phenotype data and/or missing genotypes. Of the 760 available LRE patients, only 190 (132 Australian, 58 SANAD) were included in the analysis. This was a significant but deliberate exclusion considered necessary to achieve consistency in the ratio of Australian to SANAD patients in both PGE and LRE groups (Table 6.4).

6.3.4. Patient stratification

For each of the research objectives, the respective patient groups were randomly allocated into a developmental dataset and a test dataset. Datasets were matched where possible for age, gender and cohort of origin (Australian or SANAD) to eliminate the influence of demographic variables. The developmental datasets were used to build predictive classifiers and the test datasets used to assess the predictive capacity of those classifiers.

6.3.5. Developmental and test datasets for objective one and two

For the distinction of seizure types, i.e. typical absence seizures and myoclonic jerks, 162 patients (80%) were allocated to the developmental dataset and 42 patients (20%) were allocated to the test dataset. The developmental dataset was further split into training (65%; n=105) and validation (35%; n=57) datasets. For the distinction of PGE patients from non-PGE controls, a larger initial cohort was available, which allowed a more optimal sub-division into the required datasets. Thus, 447 patients (72%) were allocated to the developmental dataset and 170 patients (28%) were allocated to the test dataset. The developmental dataset was further split into a training (62%; n=277) and validation (38%; n=170) datasets.

6.3.6. Genotyping and genetic variants

Australian patients had been genotyped on the Illumina GoldenGate™ platform for 4,041 candidate SNPs from 279 candidate genes in a previous multiple candidate gene study (Cavalleri et al., 2007) that also formed the basis for the Australian five-SNP pharmacogenomic classifier described in Chapter 5 (Petrovski et al., 2009). The 279 candidate genes were selected on the basis of suspected involvement in epilepsy, as part of an international collaboration to detect variants that may influence the development and treatment of common forms of epilepsy (Cavalleri et al., 2007). The gene panel included all known members of the voltage-gated sodium and calcium channel families, selected chloride and potassium channels, and key receptors, metabolic enzymes, and transporters of the major neurotransmitters (GABA, glutamate and acetylcholine) (Cavalleri et al., 2007). In contrast, SANAD patients were genotyped on the HumanHap660 Illumina bead chip (Illumina 660™) at the Wellcome Trust Sanger Institute (Cambridge, UK), yielding 550,000 genome-wide

tSNPs and 120,000 additional SNPs targeting CNVs.

6.3.7. Quality control methods and SNP inclusion

Available genetic data, as described above, was interrogated to identify a total of 2,087 SNPs that were common to both SANAD and Australian patients. Genetic data for all 2,087 SNPs were subjected to QC procedures before inclusion in the analysis. SNP QC included; i) comparison of genotyping consistency, ii) deviation from HWE, and iii) consistency in MAF. A total of 1,840 SNPs survived QC and were used in the subsequent model building and data analyses.

6.3.8. Statistical analysis and machine learning modeling software

HWE and MAF for the initial QC checks were performed using SAS® Enterprise Miner version 5.3 software (SAS®). Statistical analysis was performed using the online Cochran-Armitage test for trend (http://ihg.helmholtz-muenchen.de/cgi-bin/hw/hwa1.pl) and the Chi-square test in SPSS. ML methods were explored and employed in the development of predictive models using SAS®, with the *k*NN ML approach executed using the *k*NN algorithm as described in Chapter 5 (Petrovski et al., 2009).

6.3.9. Model development using SAS® Enterprise Miner

Many different data-mining algorithms and tools are currently available. A variety of supervised learning classification-based methods are available on SAS® Enterprise Miner for the development of predictive models for pattern recognition. These differ not only in the type of data they prefer (i.e. continuous, categorical, heterogeneous) but also in complexity of the data (i.e. interactions and relationships that may exist within data), use of functions, approaches used for algorithm generation, and overall data classification. No single supervised learning method is best suited for a particular dataset, so several algorithms were applied in this analysis. The different ML approaches used are briefly described in Table 6.2. Each SAS® data-mining approach was applied to the training dataset and this was used for the initial model fitting. Next, the validation dataset was used to monitor and tune the model weighting and for initial model assessment. Finally, the test dataset was used to determine the predictive capacity of the model.

6.3.10. *k*-Nearest Neighbour approach for model development

The *k*NN ML approach for classifying objects (described in Chapter 5) is based on closest training examples in the feature space (Hastie et al., 2001). An observation is classified by the average of the observations that are its *k* nearest neighbours, (*k* is a positive integer, typically small) and the nearest neighbour is the one with the smallest Euclidean distance in the n^{th}-dimensional feature space (Hastie et al., 2001, Kotsiantis, 2007, Petrovski et al., 2009). The contributions of the neighbours are weighted, so that the nearer neighbours contribute more to the average observation than the more distant ones (Kotsiantis 2007; Petrovski, Szoeke et al. 2009). The neighbours are identified from observations made in n^{th}-dimensional space for the training dataset, in which the correct classification is known (Kotsiantis 2007; Petrovski, Szoeke et al. 2009). The best choice of *k* is dependent upon the data; generally, larger values of *k* reduce the effect of noise on the classification but make boundaries between classes less distinct (Kotsiantis 2007; Petrovski, Szoeke et al. 2009). The optimum *k* for any given classification model is determined by various techniques, including cross validation. In the analyses described below, *k* was optimised in the validation dataset, which represented 35% of the total developmental dataset (or 38% in the PGE vs. LRE analysis) (Kotsiantis 2007; Petrovski, Szoeke et al. 2009).

6.3.11. Model building process

Initially for objective 1, all genomic data (n=1,840 SNPs) was used to build classification models. This was undertaken as part of a preliminary explanatory analysis. However, only a subset of SNPs (selected by a specific data filtering method) were later employed in formal classification models for both objectives. This was for an effort to reduce data complexity as to allow more efficient data analysis, and for the application of the *k*NN approach which preferentially functions on a smaller set of variables (Petrovski et al., 2009). The stages of data analyses for each classification task were; (i) cohort stratification, (ii) independent univariate analysis for all SNPs, (iii) application of SAS® ML approaches to the (a) training, (b) validation, and (c) test datasets (objective one), (iv) a performance test for each model using the Chi-square test (objective one), (v) feature selection using the Golub test to reduce SNP number, (vi) re-application of ML methods to each dataset using the SNP subset, (vii) application of the kNN approach, and finally, (viii) a performance test for each subset SNP model using the Chi-square test.

In total, 6 different ML algorithms were run to predict PGE seizure type. These were DLT, NN, Logistic Regression (LOGREG), Ensemble, Partial Least Squares and SVM. Where an ML approach failed to generate a result because it was unsuited to the data, an alternative ML method was applied. Individual models were developed using only the developmental dataset (training and validation datasets).

6.3.12. Dimension reduction and SNP selection using the developmental dataset

Data reduction procedures are of vital importance to ML and data mining (Czarnowski and Jędrzejowicz, 2008). Most ML algorithms employ a data reduction step whereby any irrelevant attributes are removed and a subset of variables are selected according to their influence on the outcome variable. These are often found to be embedded in ML programs (Moore et al., 2010). Identification of a suitable subset of SNPs was achieved by randomly assigning each patient in the developmental dataset into one of five independent groups, with equal numbers of cases with absence and myoclonic seizures or LRE and PGE in each group (Table 6.5). For each of these five groups, the SNPs were ranked according to their influence on seizure type/epilepsy type using the Golub score (Golub et al., 1999). The Golub score methodology has been described previously (Petrovski et al., 2009) (Figures 6.2 and 6.3, Table 6.5). Only the SNPs that ranked among the top 30 in two or more of the five independent groups was selected for further analysis.

6.3.13. Application of kNN approach

The *k*NN approach used previously for the investigation of drug response phenotypes (Chapter 5) involved an n-1 leave-one-out cross validation for model optimisation and initial assessment. In the current analysis, however, this step was performed in specific validation datasets, representing 35% and 38% of the development datasets for objectives 1 and 2, respectively. This was performed by randomly dividing data into five equally sized groups, stratified by seizure (objective 1) or epilepsy (objective 2) type. The prediction model was then fitted to four of the subgroups and validated by calculation of prediction error in the fifth subgroup. This process was repeated each of the five subgroups in turn and final estimates of the prediction error combined (Tables 6.11. and 6.14).

6.3.14. Optimisation and assessment

The performance of each of the ML models was first internally assessed using the validation dataset. This allowed optimisation of the models and was additionally critical to the selection of the best model parameters. For the *k*NN classifier approach, a cross-validation method was used at this stage (Chapter 5). Independent validation of the ML models was performed on the test datasets for each objective (20% of total population for objective 1, 28% for objective 2) to confirm the predictive accuracy of the models. This involved re-running each of the ML approaches, including *k*NN, using only test dataset patients. This step was performed on both the full SNP set (n=1,840) and the filtered SNPs.

Sensitivity indicates a test's ability to correctly classify those with the phenotype of interest and is analogous to the true positive rate. Specificity measures the ability of a test to correctly classify those without the phenotype of interest and is akin to the true negative rate (Kotsiantis, 2007). Classification accuracy of each model on the validation and test datasets was measured to determine the chance likelihood of the predictions; a 2x2 contingency table was generated and the difference between actual (TP, TN) and predicted values (FP, FN) assessed using Fisher's exact test (www.langsrud.com/fisher.htm). Sensitivity and specificity values of ≥80% and PPV of ≥80% were used to indicate good model performance (Petrovski et al., 2009). Sensitivity, specificity and positive (PPV) and negative (NPV) predictive values were calculated for each of the models using the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) rates automatically generated by each of the SAS® ML models. For the *k*NN model, the TN, TP, FN, FP values were calculated manually using the actual and predicted outcome generated by the algorithm (Larranaga et al., 2006).

6.3.15. Investigation of the biological significance of SNPs

In addition to the genetic analyses, each subset of SNPs identified after the data reduction stage (i.e. those found to be most predictive of seizure type for objective 1 and epilepsy type for objective 2) were subject to bioinformatics analysis to identify their potential biological significance (see section 2.4.2 and 3.2.7) (fastsnp.ibms.sinica.edu.tw) (www.ensembl.org)(www.cbrc.jp/research/db/TFSEARCH.html)(Yuan et al., 2006) (section 3.2.7). Information was also extracted on genomic structure; including gene/SNP LD structure (section 3.2.4).

6.3.16. Application of the Golub score method in each classification task

For objective 1 of the top 30 SNPs in each of the Golub subgroups, 10 SNPs were found in three of the five subgroups and one in four of the five subgroups (Figure 6.2). This subset of SNPs (N=11) was subjected to ML modeling for seizure type. For objective 2 one

SNP was found across all five subgroups, seven were found in four of the five subgroups, five in three subgroups, and three SNPs in two of the five subgroups (Figure 6.3) and this subset of 16 SNPs (N) was subjected to ML modeling for epilepsy type.

6.3.17. Statistical analysis

Chi-square statistics (SPSS) and Cochran-Armitage test for trend (http://ihg.helmholtz-muenchen.de/cgi-bin/hw/hwa1.pl) were used for all univariate analyses, with $P \leq 0.05$ indicative of statistical significance for both single SNP association and model assessment

Table 6.5 Random stratification of PGE patients

Patients were stratified into five independent groups with equal number of individuals with each seizure or epilepsy phenotype for the application of the Golub filtering approach

Seizure classification	**Group1 n=32**	**Group2 n=32**	**Group3 n=32**	**Group4 n=33**	**Group5 n=33**
JME	12	12	12	13	13
ABS	20	20	20	20	20
Syndrome classification	**Group1 n=90**	**Group2 n=90**	**Group3 n=90**	**Group4 n=89**	**Group5 n=88**
PGE	62	62	62	62	61
LRE	28	28	28	27	27

PGE = primary generalised epilepsy JME = myoclonic seizures, ABS = absence seizures, PGE= primary generalised epilepsy; LRE= localised related epilepsy

Figure 6.2 Random stratification of the developmental cohort into five independent groups for the application of the Golub filtering approach for seizure type

The top 30 ranked single nucleotide polymorphisms within each of the five independent cross-validation subgroups were partitioned from the training set. Arrows and colours indicate common SNPs across three or more of the five groups that were selected for the additional model development.

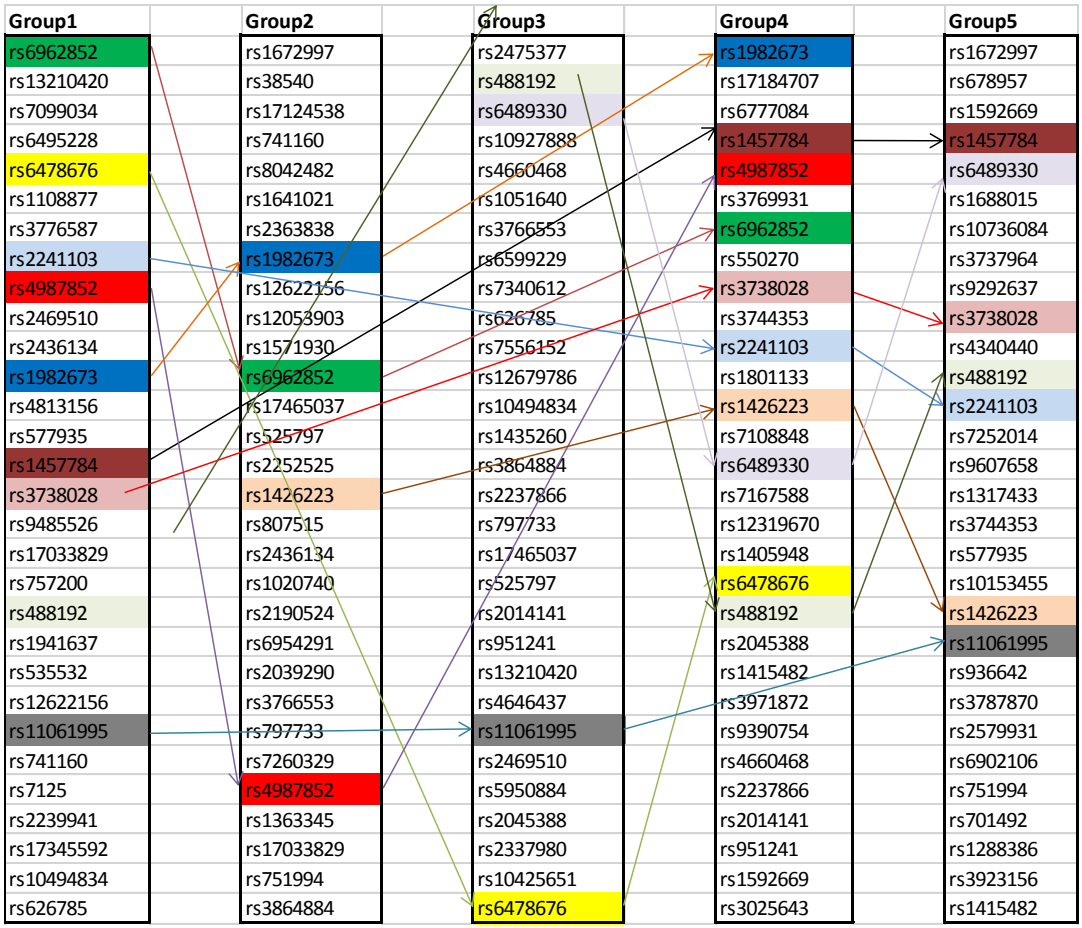
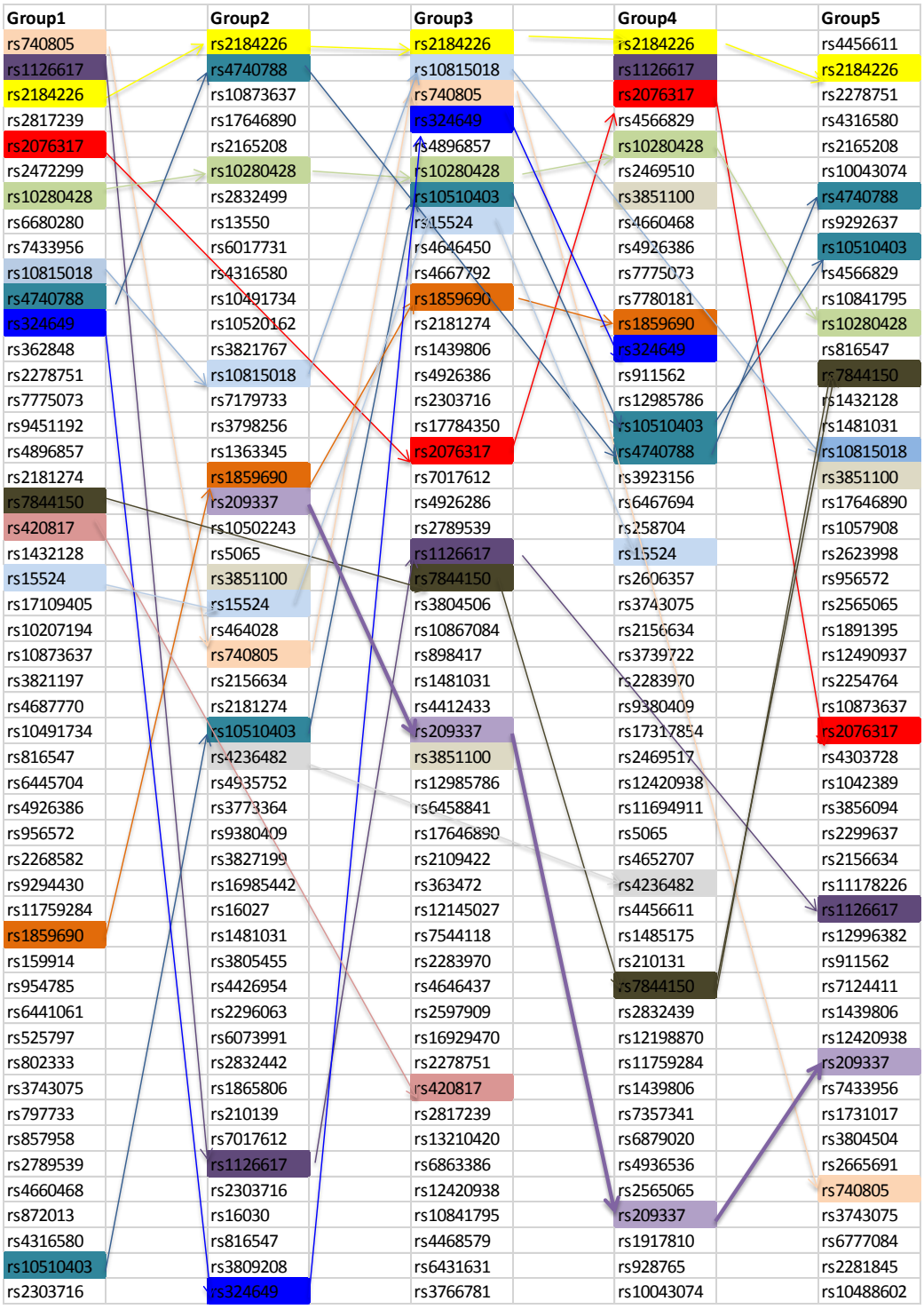


Figure 6.3 Random stratification of the developmental cohort into five independent groups for the application of the Golub filtering approach for epilepsy type

The top 30 ranked single nucleotide polymorphisms within each of the five independent cross-validation subgroups were partitioned from the training set. Arrows and colours indicate common SNPs across two or more of the five groups that were selected for additional model development.



6.4. Results

6.4.1. Development of a classifier for seizure type using all 1,840 SNPs

Since the original 1,840 SNPs were selected from epilepsy related genes (Cavalleri et al., 2007), the first stage of the model building process was a univariate test to identify any individual SNPs associated with seizure type. From these SNPs five SNPs were found to have a p-value of 0.01 or below, before correction for multiple testing. All 1840 SNPs were subsequently used in ML model development. When all 1,840 SNPs were investigated, good performance was found with the training dataset across all ML models, with SVM being the most accurate and showing the lowest error (no incorrect classifications in the training dataset) (Table 6.6). However, none of the models were able to accurately classify seizure type in either the validation or test datasets, as shown in (P>0.05). This poor overall model performance was expected, given the large number of SNPs employed. This task was mainly performed for the purpose of model exploration and to assess the benefit of the subsequent data reduction stage to predictive modeling.

6.4.2. Univariate results seizure type and training data

The results of the independent analysis of association between each of the 11 SNPs with seizure type in the training dataset is illustrated in Table 6.7. All SNPs showed association with seizure type (uncorrected for multiple testing) in the training data, with three SNPs showing a Chi-square of P of <0.01. The performance of each of the ML models developed on combining all 11 SNPs is presented in Table 6.8. With this multi-SNP analysis, a lower prediction error was observed across all four models, (P=<0.05 in the training and validation datasets) in most cases. The NN approach appeared to be the most accurate model, correctly identifying 29 typical absence and 17 myoclonic jerk cases but incorrectly identifying 6 myoclonic jerks as typical absence (false positives) and 7 typical absence cases as myoclonic jerks (false negatives); $P=4.2\text{x}10^{-27}$ in the training dataset). Although predictive in the training and validation sets, none of the ML models were able to accurately classify individuals in the test dataset (all p-values >0.05; Table 6.8).

Table 6.6 **Classification of seizure type using 1840 SNPs and SAS ML models**

Predictive performance of each SAS machine learning model on training n=103, validation n=59 and test (n=42) data subsets

ML model	n	Data subset	TP	FP	FN	TN	PPV(%)	NPV(%)	Sensitivity(%)	Specificity(%)	*P*-value
DLT	103	Training	64	24	0	15	100.0	38.5			6.2×10^{-8}
	59	Validation	31	18	5	5	86.1	21.7			0.49
	42	Test	22	14	4	2	61.1	33.3	84.6	12.5	0.76
NN	103	Training	64	39	0	0	100.0	0.0			1
	59	Validation	36	23	0	0	100.0	0.0			1
	42	Test	42	26	0	16	100.0	0.0	61.9	-	1
LOGREG	103	Training	46	14	18	25	71.9	64.1			4.5×10^{-4}
	59	Validation	20	12	16	11	55.6	47.8			0.50
	42	Test	9	7	17	9	56.3	34.6	34.6	56.3	0.82
SVM	103	Training	64	0	0	39	100.0	100.0			2.6×10^{-29}
	59	Validation	30	16	6	7	83.3	30.4			0.34
	42	Test	21	15	5	1	58.3	16.7	34.6	56.3	0.96

ML= machine learning, DLT= decision learning tree, NN= neural network, LOGREG= logistic regression SVM= support vector machine, NPV= negative predictive value, PPV = positive predictive value, TP = true positive (CAE/JAE correctly classified as CAE/JAE), FP= false positive (JME incorrectly classified as CAE/JAE), TN= true negative (JME correctly classified as JME), FN= false negative (CAE/JAE incorrectly classified as JME), CAE= childhood absence epilepsy, JAE= juvenile absence epilepsy, JME= juvenile myoclonic epilepsy

Table 6.7 Independent analysis of 11 SNPs and seizure type n=162

Chi-square test for association between genotype and seizure type in the training data set

SNP ID (rs)	*P*-value Training set	SNP ID (rs)	*P*-value Test set
rs65652852	0.20	rs6962852	$1x10^{-4}$
rs488192	0.53	rs488192	$2.2x10^{-3}$
rs6489330	0.26	rs6489330	0.01
rs2241103	0.98	rs2241103	0.01
rs4987852	0.11	rs4987852	0.01
rs1457784	0.69	rs1457784	0.01
rs3738028	0.43	rs3738028	0.01
rs11061995	0.43	rs11061995	0.02
rs6478676	0.95	rs6478676	0.02
rs1982673	0.78	rs1982673	0.02
rs1426223	0.68	rs1426223	0.04

Table 6.8 Classification of seizure type using 11 SNPs and SAS machine learning models

Predictive performance of each SAS machine learning model on training n=103, validation n=59 and test n=42 data subsets

ML model	n	Data subset	TP	FP	FN	TN	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	*P*-value
DLT	103	Training	49	11	15	13	81.7	46.4	76.6	54.2	1.7×10^{-6}
	59	Validation	26	10	10	13	72.2	72.2	72.2	56.5	0.03
	42	Test	19	9	17	7	67.9	29.2	52.8	43.8	0.70
NN	103	Training	64	1	0	38	98.5	100.0	100.0	97.4	$1.7 \times^{27}$
	59	Validation	29	6	7	17	82.9	80.6	80.6	73.9	4.2×10^{-5}
	42	Test	20	9	6	7	69.0	53.8	76.9	43.8	0.14
LOGREG	103	Training	59	9	5	30	86.8	85.7	92.2	76.9	4.3×10^{-13}
	59	Validation	31	8	5	15	79.5	86.1	86.1	65.2	7.2×10^{-5}
	42	Test	20	10	6	6	66.7	50.0	76.9	37.5	0.26
SVM	103	Training	61	10	32	9	65.6	47.4	85.9	22.0	6.1×10^{-14}
	59	Validation	2	11	12	34	15.4	14.3	14.3	75.6	2.2×10^{-4}
	42	Test	24	13	2	3	64.9	60.0	92.3	18.8	0.27

ML = machine learning, DLT = decision learning tree, NN = neural network, LOGREG = logistic regression SVM= support vector machine, NPV = negative predictive value, PPV = positive predictive value, TP = true positive (CAE/JAE correctly classified as CAE/JAE), FP = false positive (JME incorrectly classified as CAE/JAE), TN = true negative (JME correctly classified as JME), FN = false negative (CAE/JAE incorrectly classified as JME), CAE = childhood absence epilepsy, JAE = juvenile absence epilepsy, JME = juvenile myoclonic epilepsy

6.4.3. *k*NN analysis of seizure type

Prior to application of the *k*NN algorithm, the 11 SNPs were independently tested for association with seizure type in the patients (n=42) forming the test dataset (Table 6.9) and no association was identified. For the *k*NN approach, the developmental dataset (n=162) was randomly allocated into five validation groups (V1-V5; n=32) with equal numbers of typical absence and myoclonic jerks cases in each group (Table 6.10). A single group (e.g. V1) was then used as a validation group and the remaining groups (V2-V5) as the training dataset to make a prediction. This process was repeated with each of the five validation groups in turn, with V1-V5 each used as the validation dataset on one occasion only. An average predictive performance estimate across the five runs was then generated. The training dataset was therefore built on an average of 130 patients, with 32 individuals used as a validation dataset for independent cross-validations. This allowed the determination of the best *k* (i.e. number of nearest neighbours) and avoided model over-fitting. A $k=13$ was found to be optimal, with the best predictive performance value and lowest *p*-value. The model parameters were therefore N=11 (11 SNPs) and $k=13$ (13 nearest neighbours).

The classifier was predictive of seizure type in the test dataset but not at a statistically significant level (model $P=0.06$). Table 6.10 shows the development and performance data for the *k*NN model using the training (n=162) and the test dataset (n=42). In summary, five individuals were correctly classified as having myoclonic jerks and 24 individuals were correctly classified as typical absence, two typical absence cases were incorrectly classified as myoclonic jerks (false positive) and 11 myoclonic jerk cases were incorrectly classified as typical absence (false negative). Test cohort PPV and NPV were 71% and 69% respectively.

Table 6.9 Independent analysis of 11 SNPs and seizure type in the test data set

SNP ID (rs)	Odds ratio	*P*-value
rs65652852	0.592	0.20
rs488192	1.649	0.53
rs6489330	0.718	0.26
rs2241103	1.39	0.98
rs4987852	1.518	0.11
rs1457784	0.768	0.69
rs3738028	1.18	0.43
rs11061995	1.839	0.43
rs6478676	0.914	0.95
rs1982673	1.528	0.78
rs1426223	0.913	0.68

SNP = single nucleotide polymorphism, P-values uncorrected

Table 6.10 Results of the kNN machine learning approach for seizure type

Performance of the *k*NN *20*% and n-1 cross validation in the training (n=162) and test dataset (n=42) respectively

kNN model	TP	FP	FN	TN	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Two tail *P*-value
V1	5	4	7	16	41.7	80.0	55.6	69.6	0.20
V2	9	0	3	20	75.0	100.0	100.0	87.0	8x10-6
V3	7	2	5	18	58.3	90.0	77.8	78.3	6x10-3
V4	4	1	9	19	30.8	95.0	80.0	67.9	0.07
V5	7	0	6	20	53.8	100.0	100.0	76.9	4x10-4
V1-5; n=162	32	7	30	93	51.6	93.0	82.1	75.6	2x10-10
n=42	5	2	11	24	31	92	71	69	0.06

TP= true positive (CAE/JAE correctly classified as CAE/JAE), FP= false positive (JME incorrectly classified as CAE/JAE), TN = true negative (JME correctly classified as JME), FN = false negative (CAE/JAE incorrectly classified as JME, kNN= k-Nearest Neighbour, V= 20% cross validation subset of validation cohort, TP = true positive, FP = false positive, TN= true negative, FN= false negative, Sens= sensitivity, Spec= specificity PPV= positive predictive value, NPV= negative predictive value, CAE= childhood absence epilepsy, JAE= juvenile absence epilepsy, JME = juvenile myoclonic epilepsy

6.4.4. Univariate association of SNP genotype with epilepsy type

All 16 SNPs were assessed for independent association with epilepsy type. The results of the Chi-square test of association performed on the developmental dataset (n=447) is shown in Table 6.11. Univariate analysis showed that all SNPs had some association with epilepsy type with five SNPs showing a Chi-square p-value of <0.01.

6.4.5. Development of a classifier for epilepsy type

For objective 2, the same approach was taken for model development and SNP analyses as described for objective 1. The performance of each of the ML models in the developmental dataset is presented in Tables 6.12a and 6.12b. The accuracy and sensitivity of each ML model in predicting epilepsy type in the test dataset is also presented in Table 6.12. When all 16 SNPs were investigated, each of the ML models was able to predict PGE with good predictive accuracy in the training dataset (P =<0.05) but as in objective 1, the majority of models failed to accurately classify epilepsy type in either the validation or test datasets. Modest associations were observed with NN and Ensemble models in the test datasets (P =0.017 and P =0.034, respectively) but overall predictive performance was poor (sensitivity =70.3% and 65.3%, PPV =75.5% and 71.3 %, respectively).

TABLE 6.11 Independent analysis of 16 SNPs and epilepsy type n=447

Chi-square test for testing association between genotype and epilepsy type in training set

SNP ID (rs)	Odds ratio	*P*-value
rs10280428	2.32	3.9×10^{-4}
rs740805	0.59	2.9×10^{-3}
rs324649	0.61	1.2×10^{-3}
rs420817	2.50	2.6×10^{-3}
rs15524	1.61	1.8×10^{-3}
rs4236482	2.14	1.7×10^{-2}
rs209337	0.30	4.4×10^{-4}
rs1126617	1.49	3.4×10^{-3}
rs2076317	0.54	7.7×10^{-3}
rs7844150	0.71	0.02
rs2184226	1.59	0.02

SNP = single nucleotide polymorphism, P-values uncorrected

Table 6.12a Classification of epilepsy type using 16 SNPs and SAS machine learning models

Predictive performance of each SAS machine learning model on training n=279, validation n=168 and test n=170 data subsets

ML model	n	Data subset	TP	FP	FN	TN	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	*P*-value
DLT	279	Training	-	-	-	-	-	-			-
	168	Validation	nd	Nd	nd	nd	nd	nd			nd
	170	Test	nd	Nd	nd	nd	nd	nd	nd	nd	nd
NN	279	Training	192	1	1	85	99.5	98.8			$4.6x10^{-70}$
	168	Validation	91	34	25	18	78.4	34.6			0.06
	170	Test	83	27	35	25	70.3	48.1	75.5	41.7	0.02
LOGREG	279	Training	174	0	19	86	90.2	100.0			$1x10^{-53}$
	168	Validation	61	22	55	30	52.6	57.7			0.14
	170	Test	57	23	61	29	48.3	55.8	71.3	32.2	0.37

ML = machine learning, DLT = decision learning tree, NN = neural network, LOGREG = logistic regression SVM= support vector machine, NPV = negative predictive value, PPV = positive predictive value, TP = true positive (PGE correctly classified as PGE, FP = false positive (LRE incorrectly classified as PGE), TN = true negative (LRE correctly classified as LRE), FN = false negative (PGE incorrectly classified as LRE), PGE= primary generalised epilepsy, LRE= localised related epilepsy

Table 6.12b Classification of epilepsy type using 16 SNPs and SAS machine learning models

Predictive performance of each SAS machine learning model on training n=279, validation n=168 and test n=170 data subsets

ML model	n	Data subset	TP	FP	FN	TN	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	*P*-value
SVM	279	Training	193	0	0	86	100.0	100.0			$2.8x10^{-74}$
	168	Validation	93	35	23	17	80.2	32.7			0.06
	170	Test	92	35	26	17	78.0	32.7	72.4	39.5	0.10
ENSEMBLE	279	Training	193	0	0	86	100.0	100.0			$2.8x10^{-74}$
	168	Validation	91	34	25	18	78.4	34.6			0.06
	170	Test	77	31	41	21	65.3	40.4	71.3	33.9	0.03
PARTIAL	279	Training	193	0	0	86	100.0	100.0			$2.8x10^{-74}$
LEAST	168	Validation	91	33	25	19	78.4	36.5			0.03
SQUARES	170	Test	93	38	25	14	78.8	26.98	71.0	35.9	0.26

ML = machine learning, DLT = decision learning tree, NN = neural network, LOGREG = logistic regression SVM= support vector machine, NPV = negative predictive value, PPV = positive predictive value, TP = true positive (PGE correctly classified as PGE, FP = false positive (LRE incorrectly classified as PGE), TN = true negative (LRE correctly classified as LRE), FN = false negative (PGE incorrectly classified as LRE), PGE= primary generalised epilepsy, LRE= localised related epilepsy

6.4.6. *k*NN analysis of epilepsy type

The 16 SNPs were also independently tested for association with epilepsy type in the test dataset (n=170) (Table 6.13), before application of the *k*NN algorithm. A weak association with epilepsy type was seen with one SNP only in this smaller dataset (rs4740788, *P*= 0.04).

The *k*NN approach was applied to the PGE vs. LRE analysis as described for objective 1. The performance of the *k*NN model within the developmental dataset was again assessed using a 20% CV and different *k*-parameters to identify the optimum *k*. The *k*-parameter that performed best within the developmental dataset based on the average prediction accuracy across the 5-fold CV was a *k* of 13. The model parameters were therefore N=16 (16 SNPs) and *k*=13 (13 nearest neighbours). Table 6.14 presents the results of the *k*NN model using the developmental dataset (n=447) and the test dataset (n=170). Using the test dataset patients, 110 individuals were correctly classified as having PGE and 8 individuals were correctly classified as having LRE, 44 LREs were incorrectly classified as PGE (false positive) and 8 PGEs were incorrectly classified as LRE (false negative), with overall PPV and NPV of 71% and 50%, respectively. Although the classification performance was markedly improved with *k*NN in comparison to other ML approaches, the model failed to adequately classify PGE and LRE patients (test dataset *P* =0.07).

TABLE 6.13 Independent analysis of 16 SNPs and epilepsy type n=170

Chi-square test for testing association between genotype and epilepsy type in test dataset

SNP ID (rs)	Odds ratio	P-value
rs10280428	0.41	0.11
rs740805	0.96	0.78
rs324649	1.03	0.41
rs420817	1.21	0.45
rs15524	1.34	0.57
rs4236482	1.51	0.37
rs209337	0.85	0.50
rs1126617	1.20	0.72
rs2076317	0.42	0.08
rs7844150	1.57	0.12
rs2184226	0.63	0.12
rs3851100	0.89	0.61
rs10815018	1.01	0.84
rs4740788	3.65	0.04
rs10510403	1.10	0.93
rs1859690	0.44	0.13

SNP = single nucleotide polymorphism, P-values uncorrected

Table 6.14 Results of the kNN machine learning approach for epilepsy type

Performance of the *k*NN *20*% and n-1 cross validation in the training (n=447) and test dataset (n=170) respectively

kNN model	TP	FP	FN	TN	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Two tail *P*-value
V1	59	20	3	8	95.1	29.0	75.0	73.0	0.10
V2	57	19	5	9	92.0	32.1	75.0	64.2	0.08
V3	55	19	7	9	89.0	32.1	74.3	56.2	0.06
V4	59	17	3	10	95.1	37.0	77.6	76.9	0.09
V5	59	23	2	4	97.0	15.0	72.0	67.0	0.22
V1-5; n=447	289	98	20	40	93.5	29.0	75.0	67.0	2x10-4
n=170	110	44	8	8	92	15	71	50	0.07

TP= true positive (PGE correctly classified as PGE), FP= false positive (LRE incorrectly classified as PGE), TN= true negative (LRE correctly classified as LRE), FN= false negative (PGE incorrectly classified as LRE), kNN= k-Nearest Neighbour, V= 20% cross validation subset of validation cohort, TP= true positive, FP = false positive, TN= true negative, FN = false negative, Sens= sensitivity, Spec= specificity, PPV= positive predictive value, NPV= negative predictive value, PGE= primary generalised epilepsy, LRE= localised related epilepsy

6.4.7. Biological significance of genetic variants used in ML models

The biological investigation of two groups of SNPs associated with seizure type and epilepsy type are found in Tables 6.15 and 6.16 respectively. For seizure type ten SNPs were intronic region variants and one was a synonymous coding region SNP (Table 6.15). Investigation of the genomic structure using www.hapmap.org (release # 24) failed to identify any biologically functional variants with a MAF ≥ 1% in European populations that were in strong LD ($r^2 \geq 0.8$) with any of these SNPs. Further functional analysis (Yuan et al., 2006) and assessment of potential TF binding sites (Kent et al., 2002) highlighted the synonymous rs6962852 variant (in the *CLCN1* gene) to be located in both TF binding domains and in an enhancer splice site. Although this synonymous SNP was not expected to alter TF binding characteristics (TFSEARCH), it was predicted to alter the number of exonic splicing enhancer (ESE) motifs (ESE finder, Fast SNP). Functional mutations in the corresponding *CLCN1* gene

have previously been identified in rare familial forms of PGE (Dibbens et al., 2007) but not in large cohorts of sporadic PGE, indicating that *CLCN2* is probably not a gene that is commonly mutated in PGEs (Gardiner, 2005, Dibbens et al., 2007). Limited biological significance and a lack of predictive performance in the test dataset would question the relevance of all 11 SNPs in susceptibility to a PGE seizure type.

Of the 16 variant subset used for epilepsy type classification, only two SNPs were located in protein coding regions (one non-synonymous and one synonymous variant) (Table 6.16) and these were also not in strong LD with (HapMap release # 24, $r^2 \geq 0.8$) any other biologically functional variants with a MAF ≥ 1% in European populations that Four of these 16 SNPs were however in strong LD with each other (Table 6.16). Further functional and regulatory region analysis for these is presented in Table 6.17. Of the 16 SNPs The rs1126617 (a non-synonymous SNP in the glycosylphosphatidyl-inositol specific phospholipase D1 (*GPLD1*) gene was predicted to be a low risk splicing regulation polymorphism that resulted in an alteration in a splice site (Fast SNP; ESE/ESS finder). Amongst the remaining 15 SNPs, the rs1020848, rs740805, rs420817, rs3851100 and rs10510403 variants were also predicted to possess low risk in terms of a potential functional effect. These were promoter or regulatory region SNPs that may result in altered TF binding (TFSEARCH). Further investigation of the SNPs and their corresponding genes in the literature also indicated no additional biological implication or disease/epilepsy association.

Table 6.15 Genetic information for 11 SNPs for epilepsy seizure type classifiers

SNP ID (rs)	Gene	Alleles	MAF	Description	a.a change
rs6962852	*CLCN1*	C>T	0.28	synonymous	P.T87T
rs488192	*SLC6A13*	A>G	0.17	intronic	
rs6489330	*CACNA2D4*	G>A	0.19	intronic	
rs2241103	*ABAT*	A>G	0.09	intronic	
rs4987852	*BCL2*	A>G	0.08	3'UTR	
rs1457784	*KCNQ3*	C>A	0.13	Intronic	
rs3738028	*KCNN3*	T>G	0.35	Intronic	
rs11061995	*CACNA2D4*	G>A	0.16	Intronic	
rs6478676	*GABBR2*	G>A	0.43	Intronic	
rs1982673	*BCL2*	T>G	0.12	Intronic	
rs1426223	*GABRB3*	C>T	0.28	intronic	

SNP= single nucleotide polymorphism, MAF= minor allele frequency, a.a= amino acid change, T= threonine

Table 6.16 Genetic information for 16 SNPs for epilepsy type kNN classifiers

SNP ID (rs)	Gene	Alleles	MAF	Location	Amino acids	LD
rs10280428	*CACNA2D1*	A>C	0.08	Upstream		
rs740805	*CACNG5*	T>C	0.12	Upstream		
rs324649***	*CHRM2*	C>T	0.39	intronic		****
rs420817****	*CHRM2*	T>C	0.49	intronic		***
rs15524	*CYP3A5*	T>C	0.04	3' UTR		
rs4236482	*FAM131B*	G>A	0.19	intronic		
rs209337	*GABRG2*	C>A	0.05	intergenic		
rs1126617**	*GPLD1*	G>A	0.36	Exonic	P.V30I	*
rs2076317*	*GPLD1*	A>G	0.39	upstream		**
rs7844150	*KCNQ3*	G>T	0.07	intronic		
rs2184226	*MTHFR*	G>A	0.09	downstream		
rs3851100	*SCN3B*	T>C	0.16	intronic		
rs10815018	*SLC1A1*	A>G	0.39	intronic		
rs4740788	*SLC1A1*	T>C	0.11	intergenic		
rs10510403	*SLC6A1*	A>G	0.17	intronic		
rs1859690	*ZNF498*	A>G	0.04	Exonic	P.E388E	

*SNP= single nucleotide polymorphism, MAF= minor allele frequency, a.a= amino acid change, LD= linkage disequilibrium, V= valine, I=iisoleucine, E= glutamine, LD= linkage disequilibrium; (r2≥0.8), * see corresponding SNP*

Table 6.17 Predicted function and functional effect of 16 SNPs used for predicting epilepsy type (http://fastsnp.ibms.sinica.edu.tw/pages/inputCandidateGeneSearch.jsp)

SNP ID (rs)	Gene	Possible functional effects	Risk	TF binding site change	SE/ SS
rs10280428	*CACNA2D1*	promoter/regulatory region	1-3	yes	-
rs740805	*CACNG5*	promoter/regulatory region	1-3	yes	-
rs1126617	*GPLD1*	splicing regulation	2-3	-	yes
rs4740788	*CACNA2D1*	intronic enhancer	1-2	-	-
rs420817	*CHRM2*	intronic enhancer	1-2	yes	
rs3851100	*SCN3B*	intronic enhancer	1-2	yes	
rs10510403	*SLC6A1*	intronic enhancer	1-2	yes	
rs1859690	*ZNF498*	sense/synonymous	1-1	-	-
rs324649	*CHRM2*	intronic with no known function	0	-	-
rs15524	*CYP3A5*	downstream with no known function	0	-	-
rs4236482	*FAM131B*	intronic with no known function	0	-	-
rs2076317	*GPLD1*	Upstream with no known function	0	-	-
rs7844150	*KCNQ3*	Intronic with no known function	0	-	-
rs2184226	*C1orf167*	Intronic with no known function	0	-	-
rs10815018	*SLC1A1*	Intronic with no known function	0	-	-
rs10510403	*SLC6A1*	Downstream with no known function	0	-	-

TF= transcription factor, SE= splicing enhancer site change, SS=splicing silencer site change. Risk = Upper and lower risk of functional effect, 0=no effect, 1=very low risk, 2=low risk, 3=medium risk, 4=high risk, 5=very high risk (http://fastsnp.ibms.sinica.edu.tw/ pages/ input_CandidateGeneSearch.jsp).

6.5. Discussion

Common forms of genetic epilepsies are among numerous common, complex disorders for which networks of gene regulation and interactions are thought to confer disease risk (Ferraro and Buono, 2006, Ferraro et al., 2012). These genetic factors are likely to be inter-related, probably in a highly complex fashion. Recent assumptions regarding the genetic architecture of complex epilepsies include the popular CDCV hypothesis (Lohmueller et al., 2003, Cavalleri et al., 2007). However, the vast majority of the common variants identified so far confer only small risks (Dibbens et al., 2007). This has led to the proposal of developing novel approaches for simultaneously testing multiple genetic loci of small effect as an attempt to increase the capacity of correctly predicting the likelihood of disease occurrence (McKinney et al., 2006). ML is one such proposed method for efficient data analysis (Hastie et al., 2001). The *k*NN supervised ML approach was able to predict seizure control in newly treated Australian patients with epilepsy but the method has yet to be validated in independent and international cohorts (Petrovski et al., 2009, Johnson et al., 2011b). The advantages of using ML techniques for gene association data include robustness, higher power and greater accuracy than that of parametric statistical approaches, as well as the additional ability to model non-linear effects and high-dimensional data (Lee et al., 2008).

The studies reported in this chapter applied several ML approaches to test their proficiency in the analysis of complex disease association data. Large genomic datasets from two independent epilepsy cohorts were explored. ML was used for both identifying susceptibility variants associated with PGE and the seizure types defining the main PGE sub-syndromes. Two SNP subsets, comprising 11 and 16 SNPs were identified as most significantly associated with seizure type and epilepsy type respectively, and used in the development of two phenotype classifiers.

All individual SNP subsets in each identified subset were found to associate with seizure or epilepsy type at the $p=<0.05$ level when univariately tested in the respective developmental cohorts, thus indicating some initial significance of each of these SNPs to PGE. Most of these however failed when applied to the blinded test cohorts. ML models using the SNP subsets in combination were similarly found to associate with phenotype for both objectives but again only in the training cohorts in both cases. Some predictive value was evident in the epilepsy type classification task when NN and Ensemble approaches were applied; Test cohort PPV= 70.3%, 65.3% and Sensitivity= 75.5%, 70.3%, respectively. These models were however only applied to the analysis concerning all 1,840 SNPs, thus no corresponding data is available for the 16 SNP subset analysis. The NN ML method has previously been used in both linkage and association analyses for the identification of disease susceptibility genes as well as complex traits (Motsinger-Reif et al., 2008). The second

approach, Ensemble is a novel ML method available on SAS that combines the results of multiple classifiers and has been shown to achieve a substantially improved prediction when compared to single classifiers in several reviews (Ahn et al., 2007, Moon et al., 2007).

The main hypothesis was that the *k*NN ML approach in particular, due to previously being able to make successful predictions for complex genomic data (Petrovski et al., 2009), may be a more suitable method for developing classifiers of PGE than traditional ML methods (Petrovski et al., 2009). Indeed, the Australian patients and genomic data used in the current analysis were the same as those used in the original Australian pharmacogenomic *k*NN classifier study (Cavalleri et al., 2007, Petrovski et al., 2009). The *k*NN ML approach is considered to be more user friendly, more easily used for incremental learning, more easily tuned, and better for avoiding over-fitting when compared to other ML approaches (Kotsiantis, 2007). Improved performance was seen with the *k*NN approach for each classification task (p=0.06 and p=0.07 for seizure and epilepsy type analyses, respectively) but the *k*NN ML models similarly failed to predict phenotype when applied to the respective test datasets.

6.5.1. General design considerations

There are several explanations for the lack of success with ML approaches to predicting PGE and PGE seizure types. These include; i) issues with study design, ii) inherent problems with predictive modeling, iii) genomic differences between populations, and iv) complexity of the task itself. Some of the general issues with population genomic differences and ML modeling, particularly in the *k*NN method, have been discussed previously (section 5.4).

<u>Population genomic differences from using international populations</u>

- The potential influence of ethnicity on genetic transmission (Delgado-Escueta, 2007) is likely to be far too diverse for analyses that combine populations.

- Variants common to multiple populations are each likely to be of a small effect size and so undetectable in modestly sized cohorts as available for this study

- Despite the adoption of cross validation method the cohorts may also remain insufficient in size to allow the partitioning required for model building and independent assessment

Methodological issues

- A data reduction step may have excluded potentially causative alleles or loci (Larranaga et al., 2006, Kotsiantis, 2007, Joaqu et al., 2010, Derrac et al., 2012);

- The Golub score filter method for SNP ranking is an easy method to implement, however it does not consider correlation between features or SNPs (Golub et al., 1999).

- Principal limitations of classification algorithms exist including limited scalability, potential for over-fitting (this can result in false-positive results) (Hastie et al., 2001), challenging feature/SNP selection, and difficulty in accounting for gene-gene interactions (Moore and Ritchie, 2004).

- Most existing studies using ML approaches have dealt only with candidate SNP data in which hundreds rather than thousands of SNPs are modeled (Szymczak et al., 2009, Goldstein et al., 2010).

Heterogeneous and multigenic nature of the PGE phenotype

- Likelihood that insufficient numbers of SNPs were employed to detect the multiple common variants thought to contribute to the complex PGEs (Kasperaviciute et al., 2010); The 1,840 variants used in this study constituted fewer than 50% of the 4,041 SNPs in the original SNP panel; GWAS scans that typically provide in excess 500,000 SNP genotypes may be more efficient for the identification of susceptibility loci in oligogenic traits, (Ferraro et al., 2012, Kearney, 2012).

- Complex PGEs have been investigated primarily on the study of SNPs (Ferraro et al., 2012) (Schork et al., 2009). Rare variants including CNVs are typically excluded, despite these less common genetic defects might equally explain inter-individual susceptibility to disease (Mefford et al., 2010, Ferraro et al., 2012).

6.5.2. Summary and future perspectives

The investigation described in this chapter set out to assess the potential value of ML models in epilepsy classification by applying this essentially novel approach to a large subset of genomic data in an effort to identify variants of potential significance to complex PGE phenotypes. Although the current study was unable to model the available genomic data successfully and did not produce classifiers with good predictive performance, overall the classifiers that were developed on a subset of SNPs did appear to show improved association with seizure and epilepsy type than when the same SNPs were tested individually. These present findings further emphasise several points and considerations for future work concerning complex epilepsy genomics, namely that;

- There may be tens of thousands of alleles that constitute the broader epilepsy genome;
- Multiple combinations of these could, increase susceptibility or resistance to epilepsy in any given individual.
- Identification of all or at least sufficient numbers of these genomic markers to differentiate between seizure and epilepsy types remains an elusive task; requiring greater study power, in terms of both numbers of cases and genotyping methodology.
- Positive susceptibility loci require independent confirmation and validation in independent cohorts;
- In ethnically different populations, such susceptibility loci may be entirely different, thus requiring independent validation in multiple ethnic populations before a true phenotype association can be proposed.

To conclude, as so little is known about the genetics of epilepsy, the number of different possible subsets of susceptibility alleles is almost limitless and thus unravelling the phenotypic diversity of complex PGE is and will continue to be an arduous task. The strategy used in the present investigation to detect genetic predisposition in PGEs may be an improvement over methods used in traditional association studies but still fails to capture the genomic heterogeneity that potentially exists.

CHAPTER SEVEN

A CANDIDATE SNP STUDY FOR THE VALIDATION OF A MULTICENTRE GENOME WIDE ASSOCIATION ANALYSIS FOR PREDICTING TREATMENT RESPONSE IN NEWLY TREATED EPILEPSY

CONTENTS

7.1. Introduction

As discussed previously, several clinical and genomic factors have been investigated to help identify factors that predict treatment response in epilepsy (Cockerell et al., 1995, MacDonald et al., 2000, Sillanpaa and Schmidt, 2006, Kaneko et al., 2008, Loscher et al., 2009). Despite identifying a number of clinical factors associated with poor treatment outcomes in newly treated and chronic epilepsy (Hitiris et al., 2007, Kwan et al., 2011) and more recently those associated with seizure remission (Callaghan et al., 2011, Bonnett et al., 2012, Brodie et al., 2012), their predictive power and subsequent clinical utility remains limited (Brodie, 2005b, Mohanraj and Brodie, 2005, 2007). No single clinical factor has been found to accurately predict seizure control (Brodie et al., 2012), though a combination of one or several of these factors may help to define those individuals who are most unlikely to respond to drug treatment (Bonnett et al., 2012). PGx efforts have similarly made little clinical impact on the search for definite genomic predictors of AED drug efficacy (Loscher et al., 2009, Johnson et al., 2011b). Ultimately constructing multivariable, multifactorial models that combine both influential clinical and genetic factors, maybe most useful (Johnson et al., 2011b). PGx research has long recognised that the inherent basis of patient response essentially results from multiple variants of small effect (Evans and Johnson, 2001, Grant and Hakonarson, 2007). However, most current studies investigate polymorphisms univariately, typically in small or modest sized cohorts (Colhoun et al., 2003, McCarthy et al., 2008, Loscher et al., 2009).

7.1.1. Genome wide approach to complex disease genetics

Genome wide association studies of epilepsy disease genetics have recently been published (Kasperaviciute et al., 2010) as have GWAS assessing drug response in epilepsy (Kasperaviciute and Sisodiya, 2009, Cavalleri et al., 2011, McCormack et al., 2011, Ozeki et al., 2011a). GWAS studies can consider the relevance of variation across the entire genome and so are not restricted to exclusively investigating biologically driven or previously reported candidate genes (Crowley et al., 2009, Motsinger-Reif et al., 2010). With the growing number of GWAS across disease genetics, several limitations of GWAS have similarly become apparent (Crowley et al., 2009, Motsinger-Reif et al., 2010, Johnson et al., 2011b). One of the main issues for GWAS concerning complex traits is that the "common variant" hypothesis predicts weak genetic effects, (inherently due to the large number of low penetrance variants being tested). Very large sample sizes are thus required to successfully boost the genetic "signal" over the additional "noise" produced by environmental variables and other genetic factors.

GWA studies additionally only consider common genetic variation which is now thought to at best only have a modest role in the predisposition to complex syndromes and/or traits (Ferraro et al., Barrett et al., 2009, Kasperaviciute and Sisodiya, 2009, Daly, 2010a, Kasperaviciute et al., 2010). The alternative to the common disease common variant hypothesis is that multiple rare variants cause disease at high prevalence in the population (Motsinger-Reif et al., 2010). Common variants are thought to result in subtle effects on gene function, often through changes to gene regulation, whilst rare variants such as non-synonymous variants can have larger effects on gene function, which could lead to large changes in disease risk or trait values (Motsinger-Reif et al., 2010). And so, both these hypotheses can have important implications to common phenotypes (McCarthy et al., 2008, Motsinger-Reif et al., 2010).

7.1.2. Genome wide approach to epilepsy genetics

To date, GWA studies evaluating drug response in epilepsy are scarce (Kasperaviciute and Sisodiya, 2009, Cavalleri et al., 2011) and this is likely to be due to the low number of well phenotyped epilepsy patient cohorts with DNA currently available. Epilepsy comprises of a group of phenotypically and genetically heterogeneous disorders, with an underlying genetic predisposition likely in over half of individuals with epilepsy (Kearney, 2012). The genetic architecture of the epilepsies is consequently likely to be very complex, reflecting this genotypic and phenotypic heterogeneity and high degree of heritability (Kearney, 2012).

DNA samples from hundreds if not thousands of phenotyped epilepsy patients is expected to be required for epilepsy GWAS to be successful and informative (Kasperaviciute et al., 2010, Cavalleri et al., 2011). As published GWAS are underpowered to detect all but the biggest effects, the susceptibility variants identified to date, are probably only a subset of the influential loci yet to be detected and/or may indicate false positive associations (Guessous et al., 2009). Moreover, because the effect sizes of these variants are usually small and the number of false positive findings are expected to be large (McCarthy et al., 2008, Guessous et al., 2009), additional patient cohorts from independent populations will be necessary as replication cohorts (Kasperaviciute et al., 2010).

7.1.3. Genome wide studies in epilepsy

Currently, only two cases of GWAS analysis of disease susceptibility can be found in literature and both of these concern focal or localisation related epilepsy (Kasperaviciute et al., 2010, Guo et al., 2012, Kearney, 2012). The first, reported by Kasperaviciute and colleagues in 2010, used broad phenotype criteria and included individuals with focal epilepsy, regardless of etiology (Kasperaviciute et al., 2010). No genome-wide significance associations

were found, thus little was gained in terms of genes for disease susceptibility (Kasperaviciute et al., 2010). This is however not entirely unexpected, considering the high degree of heterogeneity across specific epilepsy types and the expectation that both rare and common variants contribute only small effects to complex traits (Cavalleri et al., 2007, Kwan et al., 2009, Cavalleri et al., 2011). The second focal epilepsy GWAS, using a two-stage approach and a meta-analysis of both stages proved more successful (Guo et al., 2012, Kearney, 2012).

7.1.4. Meta-analysis of GWAS for an increase in study power

Meta-analysis is a statistical technique for combining the findings from independent studies and in medicine is most often used to assess the clinical effectiveness of healthcare interventions (Egger and Smith, 1997, McCarthy et al., 2008)(www.cochrane-handbook.org)(Davey et al., 2011). The joint effort of independent research centers in combining and analysing genomic data from similar studies is one approach to improve the power of whole genome scans (Nebert et al., 2008a, Cavalleri et al., 2011, Kearney, 2012).

Aggregate data from several scans has previously facilitated detection of variants with small effects (Manolio et al., 2007, Weedon et al., 2008, Zeggini et al., 2008, Lettre, 2012) and such data-sharing efforts could also help achieve success in the anticipated wave of cohort-based GWAS for epilepsy (McCarthy et al., 2008). Meta-analyses of data from multiple epilepsy cohort may have sufficient power to detect any main as well as underlying (gene–gene and gene–environment) genetic effects, explore potential sources of heterogeneity and also inform the selection of the most relevant SNPs for replication efforts (McCarthy et al., 2008). Joint analysis of GWA scans moreover may be used to confirm any reports that have previously implicated susceptibility variants with modest effect sizes (McCarthy et al., 2008).

Meta-analysis of several GWA studies has already demonstrated considerable value in complex disease genetics, with reports of being able to implicate novel disease loci with greater confidence (Zeggini et al., 2008, Barrett et al., 2009). Several new disease risk variants of smaller effect sizes were identified for type 2 (Zeggini et al., 2008) and type 1, diabetes (Barrett et al., 2009) and such meta-analysis of GWAS are likely to similarly prove more effective and more powerful for detecting associations in epilepsy disease genetics (Kearney, 2012).

The EPICURE Consortium have recently published a linkage study in which they attempted to improve power by undertaking the first genome-wide linkage meta-analysis for PGE (Leu et al., 2012). In this meta-analysis significant linkage for myoclonic and absence seizures was reported and these were also in several previously implicated gene loci (Leu et al., 2012). Authors have since reinforced the need to collaborate and pool cohorts to increase

sample sizes to improve strength of evidence in the context of epilepsy genetics (McCarthy et al., 2008, Kasperaviciute et al., 2010, Tan and Berkovic, 2010, Leu et al., 2012).

7.1.5. Phenotypic heterogeneity in epilepsy

Thousands of epilepsy patients have participated in pharmacogenomic studies and have also been GWAS scanned (Marson et al., 2006, Cavalleri et al., 2007, Kasperaviciute et al., 2010) yet GWAS evaluating drug response in epilepsy patients remain limited in number (Kasperaviciute and Sisodiya, 2009, Daly, 2010a, Kasperaviciute et al., 2010, Johnson et al., 2011b). Genomic research towards other complex diseases as a standard now use cohorts of over 90,000 individuals, recruited through multi-centre collaborative efforts. The challenge for epilepsy PGx research is to similarly develop such multi-centre collaborations (Cavalleri et al., 2011).

7.1.6. Defining drug response

Treatment response is characterised by the remission of seizures and responders to drug treatment are currently defined by the ILAE as 'individuals being seizure free for at least 12-months after starting AED therapy' (Kwan and Brodie, 2010). Classifying response in patients with anything less than perfect seizure control however remains challenging (Cavalleri et al., 2011).

Several difficulties exist with this definition of treatment success. As previously discussed in a recent review of PGx studies in newly treated epilepsy (Johnson et al., 2011b), clinical outcome is affected by both therapeutic response and the natural history of a specific epilepsy (Johnson et al., 2011b). The natural tendency for some types of adult and childhood epilepsies is to remit spontaneously over time and so these may appear drug resistant at first, only to remit in later life. Consequently when defining treatment outcome one may be classifying as drug responders those who are i) seizure free because of a pharmacological response to AEDs and ii) those who are seizure-free because their epilepsy has spontaneously remitted.

In addition to this within medicine, there is a tendency for clinicians to dichotomise continuous traits. Individuals with epilepsy thus are usually labeled as AED responders (their seizures stop) or AED non-responders (their seizures continue), though in reality, across a population of patients with epilepsy, there is probably a continuum of therapeutic response (Johnson et al., 2011b). So far several response classification schemes have been proposed but none capture the underlying complexity and dynamic nature of response in epilepsy (Berg and Kelly, 2006, Cavalleri et al., 2011). Further research on defining drug response that both incorporates information relating to aetiology, inherent severity of the epilepsy i.e. seizure

frequency prior to starting treatment and also considers therapeutic response to AEDs as a quantitative trait has been suggested to help resolve these issues (Johnson et al., 2011b).

Larger cohort studies that also incorporate clinical covariates and additionally classify patients according to response to a specific AED may aid the identification of potentially larger and more clinically relevant genetic effects. The incorporation of a wide range of clinical variables to association studies is becoming an increasing approach to PGx study design (Sanchez et al., 2010, Cavalleri et al., 2011, Johnson et al., 2011b). The recent EPICURE GWAS meta-analysis attempt mentioned previously (section 7.1.4) presents a good example of this concept (Leu et al., 2012). Using a broad trait definition, authors did not detect any significant linkage however when stratification by epilepsy subtype was applied, significant linkage was found (Leu et al., 2012).

7.1.7. Genome wide association study meta-analysis for predicting treatment outcome in newly treated epilepsy

Clinical covariates are known to have an important influence on outcomes in epilepsy and hence in PGx studies (Petrovski et al., 2010, Sanchez et al., 2010, Cavalleri et al., 2011, Grover et al., 2011). A recent review by Johnson *et* al 2011 moreover proposed a novel concept of intermediate clinical phenotype where such influential clinical variables were proposed to potentially impact the genetic influence on drug treatment response at numerous levels (Johnson et al., 2011b) (see Figure 7.1 adopted from Johnson *et* al 2011). The assumption was, that if a genetic factor acts via a measured clinical covariate then adjustment for that covariate will confound its detection (Johnson et al., 2011b). Conversely, adjustment for clinical covariates will lead to improved detection of genetic factors influencing outcome via an independent route to a measured clinical covariate (Johnson et al., 2011b).

Authors of the above mentioned review also recently performed a multi-centre meta-analysis of GWAS carried out for treatment response in newly treated epilepsy (unreported). The meta-analysis attempt combined data from two GWA scans (Australian and UK) This work is one of the first PGX projects in newly-diagnosed epilepsy. The initial analysis of this GWAS meta-analysis identified a single variant associated with treatment outcome, with a GWAS significance p-value of $<5x10^{-7}$ (rs622902) within the *GSTA4* gene, and an additional nine top ranking SNPs (see Table 7.1 for a list of these top 10 GWAS SNPs). Most of the identified SNPs can be found on Chromosome 6 and within the *GSTA4* gene. Although the MS Genetics Consortium GWAS (Nature, 2011) applied a GWAS meta-analysis cut off p-value of $<5x10^{-8}$ other GWAS studies have used $<1x10^{-7}$ (Davila et al., 2010) and this is assumed "suggestive" of a causal association (Meyer et al., 2010). The analysis was performed both with the inclusion and exclusion of clinical covariates. The Manhattan plot from the

single SNP logistic regressions with the inclusion of clinical covariates for this GWAS meta-analysis effort, are presented in Figure 7.2. A subsequent re-analysis of the data by the authors; with updated clinical cohort information and additional patients identified a narrower set of 3 associated loci, only one of which was identified in the original analysis (original top *GSTA4* SNP).

Table 7.1 Top 10 ranking SNPs from the initial GWAS meta-analysis

SNP ID (rs)	Chromosome	Position (bp)	Gene	Allele change
rs622902	6	52954433	GST-A4	C/T
rs316132	6	52955925	GST-A4	C/G
rs316133	6	52955510	GST-A4	A/G
rs367836	6	52951090	GST-A4	A/C
rs316131	6	52956108	GST-A4	C/T
rs316130	6	52956159	GST-A4	C/T
rs316141	6	52954117	GST-A4	C/T
rs384505	6	52943531	GST-A4	C/T
rs405729	6	52950740	GST-A4	A/G
rs316128	6	52957105	GST-A4	C/A

rs= reference sequence, SNP = single nucleotide polymorphism, bp = base pair

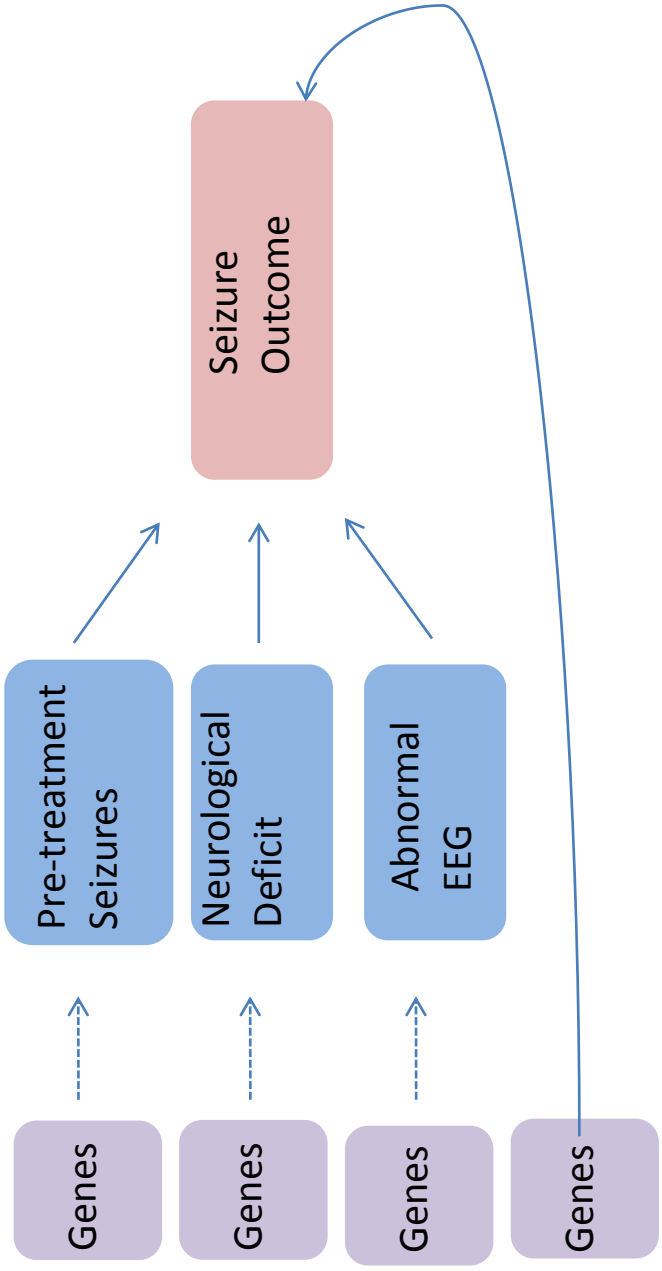


Figure 7.1 Predictive factors in a potential pathway for epilepsy treatment outcome

A schematic representation of how potential genetic and non-genetic, clinical predictive factors may act via different pathways to influence epilepsy treatment outcome. Figure adopted from Johnson *et* al 2011.

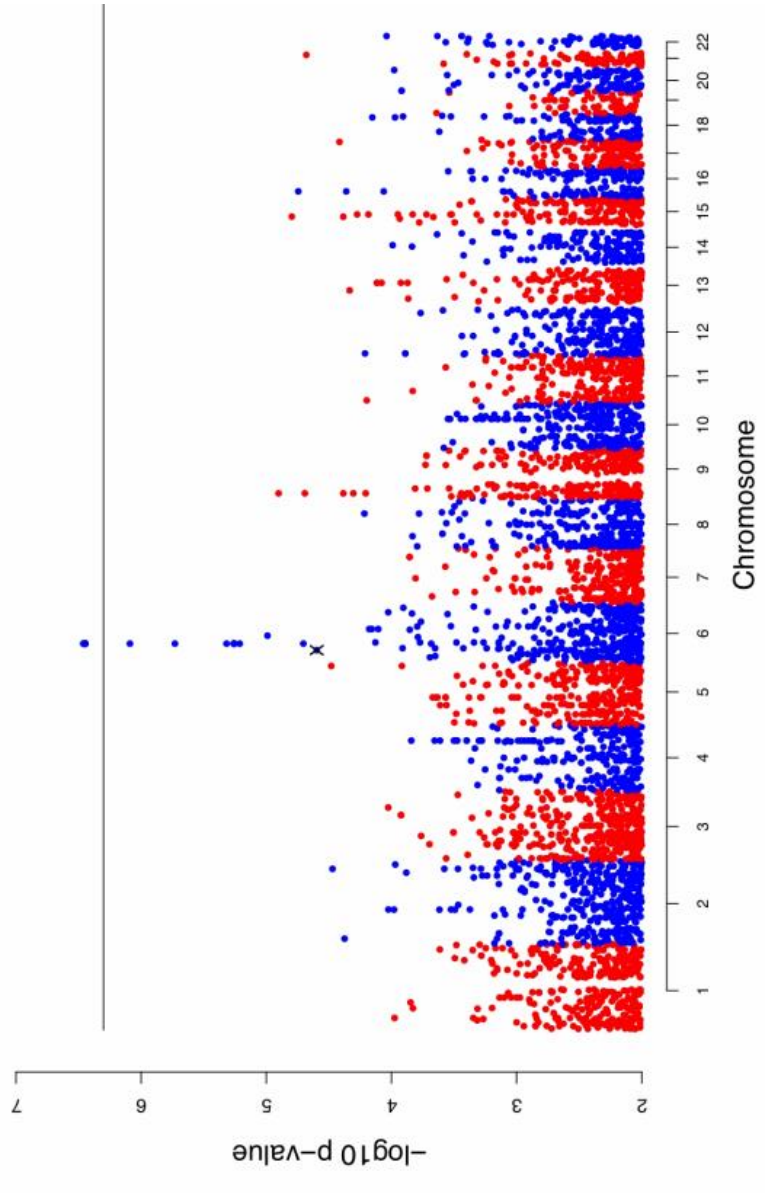


Figure 7.2 GWAS meta-analysis results combining Australian and UK SANAD data for 1-year remission, with the inclusion of clinical covariates

The Manhattan plot above from the initial meta-analysis single-SNP logistic regression also including clinical covariates is presented. The plot indicates a single *GSTA4* loci association at p-value $<5x\ 10^{-7}$ (rs622902 variant).

7.1.8. Biological significance and role of *GSTA4* in epilepsy

GSTs could be considered good candidates for epilepsy PGx due to their general role in detoxification of xenobiotics (Depondt and Shorvon, 2006). There are eight isoforms of soluble GST (**α**, **μ**, **π**, **θ**, **ω**, **ζ**, **σ**, and **κ**) and at least three membrane-bound GST isoforms (MGST1, MGST2 and MGST3) (Salinas and Wong, 1999, Board et al., 2000). GSTs are widely expressed in almost every tissue, though some of the isoforms are tissue specific (Carder et al., 1990, Desmots et al., 2001, Listowsky, 2005).

GSTs have recently been implicated in the hepatic metabolism and clearing of AEDs (Kasperaviciute and Sisodiya, 2009, Tan and Berkovic, 2010, Depondt et al., 2011). Current PGx data concerning the role of GSTs in epilepsy particularly that of GSTA4 is however sparse (Saruwatari et al., 2010). GSTs are known to be involved in the detoxification of reactive CBZ metabolites (Madden et al., 1996, Bu et al., 2005) and the deletion of a GSTM1 allele has furthermore been implicated in CBZ and VPA-related hepatotoxicity in Japanese patients (Ueda et al., 2007, Fukushima et al., 2008b, Depondt et al., 2011). It has also been hypothesised that higher levels of GSTs in the brain-blood barrier may result in low concentration of AEDs potentially leading to medical intractability (Shang et al., 2008). Human GST expression in such patients was recently examined and an association between expression of the GST-π isoform and intractability was reported (Shang et al., 2008). In several animal studies GST isoforms in liver, testis, and brain tissues were additionally reported to be induced by some AEDs (Selim et al., 2000, Thyagaraju et al., 2005), though this proposition of a mechanism of intractability development however remains to be studied (Shang et al., 2008).

7.1.9. Aims and hypothesis

Newly treated epilepsy GWAS and meta-analysis study details:

The initial results from the meta-analysis suggested that G*STA4* may play a role in treatment response (Speed, D *et* al. [Unpublished]). The initial findings identified that the 10 top ranking GWAS SNPs (lowest GWAS p-values) were mostly within chromosome 6 and located in the *GSTA4* gene (all 10 when no clinical covariates considered, and 8 out of 10 when associated clinical covariates were included in the GWAS analysis). Due to the potential biological significance of the *GSTA4* gene, our intention was to attempt to replicate this initial GWAS meta-analysis finding, in an independent cohort of well-defined individuals with newly treated epilepsy, thus adopting a candidate SNP association study approach (McCarthy et al., 2008).

The epilepsy PGx GWAS meta-analysis was based on two, independent genome-wide scans for treatment responsiveness. The two GWAS cohorts of newly treated epilepsy were i) a subset of the UK SANAD cohort and ii) the Australian, Melbourne prospective epilepsy cohort (both genotyped at the WTSI on Illumina 660Q) (both study populations previously described see Chapters 5 and 6 for cohort details)(Marson et al., 2006, Cavalleri et al., 2007). Definitions of seizure outcomes and clinical covariates were harmonised across the cohorts to allow meta- analysis of primary outcome and the clinical covariates included in the analysis. The authors of the GWAS meta-analysis designated the larger SANAD cohort as the Discovery Cohort and the Australian cohort as the replication study group and a total 552144 SNPs from both data sets were meta-analysed. The meta-analysis was performed using a prospectively agreed definition of 1-year remission of seizures (this was presumed to indicate adequate seizure control or treatment response).

Hypotheses of present study:

The primary aim of this results chapter was to perform a validation of the findings from the initial analysis effort of this first newly treated epilepsy, using individuals from the Glasgow data set as an independent cohort of UK patients with epilepsy (see section 2.2.1 for Glasgow source population details). A subset of the 10 GWAS identified SNPs were selected for genotyping and were to be assessed for association with both treatment outcome and time to 12-month remission in our current investigation.

7.2. Study cohort, materials and methods

7.2.1. Phenotype definitions for patient selection

The primary outcome of this present study was treatment success with pharmacotherapy. Individuals were classified as either responders or non-responders to AEDs. Response was defined as achieving a minimum period of 1-year at any stage after starting treatment (Speed, D *et* al. [unpublished]). Thus patients required a minimum follow- up period of 1-year after starting AED therapy. This definition was chosen as it matches the definition proposed by the ILAE (Kwan et al., 2010) and is presently seen as the only relevant seizure outcome consistently associated with meaningful improvement in quality of life (Callaghan et al., 2011, Cavalleri et al., 2011, Johnson et al., 2011b).

7.2.2. Immediate vs. delayed seizure remission and definitions for time to event analysis

In addition to investigating seizure freedom, an additional analysis was performed in order to investigate or account for the probability of delayed or late seizure remission in some patients that would not necessarily be captured by the definition above (see section 7.3 below). This delayed remission was investigated by associating time to 12-month seizure remission where time to outcome has been censored. Remission status was defined as above.

7.2.3. Glasgow validation cohort

Patients from the UK Glasgow cohort (on-going collection of DNA) (see Chapter 2) that were identified to have newly treated epilepsy at time of recruitment were utilised for the analysis on this Chapter. In total 518 patients were identified as having newly treated epilepsy and thus were available for genotyping.

All clinical notes for each of these 518 individuals were reviewed in order to confirm individual phenotype data and their eligibility. From these patients 13 were automatically excluded for either not having epilepsy or newly treated epilepsy on reviewing case notes, a further 2 had only one seizure prior to treatment, thus did not qualify as having epilepsy, 5 were of non- European ancestry and, 66 individuals had less than twelve months follow- up data (required for classification of treatment outcome). The remaining 434 patients were of European ancestry, had sufficient DNA for genotyping and clinical information for phenotyping and thus were eligible for study inclusion and subsequent genotyping.

7.2.4. Clinical data selection and inclusion

Clinical information was extracted from clinical databases. This included the general patient characteristics and disease phenotype; age at recruitment gender and epilepsy type, which are all known to potentially influence treatment outcome (Kwan and Brodie, 2001a, Hitiris et al., 2007). Additionally those clinical factors that were explored and/or included in the GWAS meta-analyses effort were considered for inclusion in this present study. These were i) initial treatment AED (the first AED administered at recruitment or first follow-up), ii) AED at final follow-up iii) EEG status, categorised as; non-done, normal, not-specific, epileptiform (abnormal) and iv) medical imaging status, categorised as; not-done, normal, non-specific, focal (abnormal). For the survival (time to treatment) analysis, AED recorded at remission was recorded and included as a covariate as opposed to final follow-up AED. For the treatment covariates i.e. Initial AED treatment, final AED treatment and AED at remission (survival analysis covariate) drug treatment was categorised as either CBZ, GBP, LTG, OXC,

VPA (the most commonly used AEDs), Multiple AEDs if more than one AED indicated and 'other' for any other drug or for missing treatment information.

7.2.5. Genetic data selection and inclusion

Rather than simply select the most significant SNPs from the GWAS analysis (i.e. those with the lowest meta p-values), validation SNPs were additionally selected on the basis of biological plausibility, functional significance, and expression array data derived from analysis of surgically resected, human epileptic brain (temporal lobe). Selection was undertaken by a collaborator (Dr Michael Johnson, Imperial College London). Five SNPs were ultimately selected for the validation study, the top 2 *GSTA4* SNPs identified by the initial GWAS meta-analysis (rs316132 and rs622902) (Table 7.1 and 7.2) and an additional 3 SNPs (rs17252760, rs12919774 and rs16994558) located in intergenic regions (Table 7.6).

Table 7.2 Top 10 ranking SNPs of the initial GWAS meta-analysis single-SNP logistic regression including clinical covariates

SNP ID (rs)	Chromosome	Position (bp)	Gene	Allele change
rs316132	6	52955925	*GST-A4*	C/G
rs622902	6	52954433	*GST-A4*	C/T
rs316131	6	52956108	*GST-A4*	C/T
rs316130	6	52956159	*GST-A4*	C/T
rs316141	6	52954117	*GST-A4*	C/T
rs316133	6	52955510	*GST-A4*	A/G
rs367836	6	52951090	*GST-A4*	A/C
rs6464296	7	152343151	UNKNOWN	G/A
rs4779485	15	28721754	*ARHGAP11B*	C/T
rs405729	6	52950740	*GST-A4*	A/G

SNP= single nucleotide polymorphism, bp= base pair

7.2.6. Genotyping of candidate SNPs

Genomic DNA samples from the n=434 Glasgow cohort were genotyped for the five candidate SNPs at the Department of Molecular and Clinical Pharmacology, University of Liverpool, using custom TaqMan® SNP genotyping assays (Applied Biosystems, Warrington, Cheshire, UK) in accordance with the manufacturer's instructions. This assay is based on the 5'-3' exonuclease activity of Taq DNA polymerase, using allele-specific TaqMan® fluorescent minor groove binding (MGB) probes VIC® and FAM™ (as specified by the manufacturer's instruction).(http://www3.appliedbiosystems.com/cms/groups/mcbsupport/documents/general documents/cms_042998.pdf).

7.2.7. Experimental details

Briefly approximately 20ng of genomic DNA (pre-dried sample) was amplified in 5uL reaction mixtures containing 1x TaqMan universal genotyping master mix and 1x TaqMan assay mix (containing a premix of the customised SNP primers and the fluorescent probes), in 384-well plates. Reactions were performed on an ABI 7900HT fast Real-Time PCR System (Applied Biosystems). A standard protocol for DNA amplification was followed where after the Taq enzyme was activated at 95°C for 10 min, 40 PCR cycles of denaturation at 92°C for 15 s and 1 min of combined annealing and extension at 60°C were completed on the reaction mixes. In total five runs were performed for each of the 424 DNA samples (1 for each of the five SNP assays). As part of quality control, negative controls (DNA replaced with water) and 10% duplicates were included in every 384-well plate run. After PCR amplification, an endpoint plate read of fluorescence and allelic discrimination was performed using the Applied Biosystems Real-Time PCR System and the Sequence Detection System (SDS) Software (Applied Biosystems). Fluorescence measurements made during the plate read are used to plot fluorescence (Rn) values based on the signals from each well. The plotted fluorescence signals indicate which alleles are in each sample.

7.2.8. Taqman chemistry

Each TaqMan MGB probe anneals specifically to its complementary sequence between the forward and reverse primer sites. When the oligonucleotide probe is intact, the proximity of the reporter dye to the quencher dye results in quenching of the reporter fluorescence primarily by Förster-type energy transfer. AmpliTaq Gold® DNA polymerase extends the primers bound to the template DNA. AmpliTaq Gold DNA polymerase cleaves only probes that are hybridized to the target complimentary sequence. Cleavage separates the reporter dye from the quencher dye, which results in increased fluorescence by the reporter. The increase in fluorescence signal occurs when the hybridized probes are cleaved.

7.3. Statistical analysis

All statistical analyses were performed in SPSS version 18. Deviation from HWE was tested for each of the five SNPs using a Chi-Square test (http://ihg.helmholtz-muenchen.de/cgi-bin/hw/hwa1.pl), with a p-value of <0.001 assumed to indicate deviation from HWE. The MAF of each SNP was also checked and any SNPs with a study cohort MAF of <0.05 were excluded from the analysis (Haploview version 4.1). The SNP MAF was also compared to that of the general population (www.hapMap.org). P-values for all genetic and non-genetic association tests undertaken in the study were adjusted for multiple testing using the FDR and a statistical p-value of ≤ 0.05 after correction was deemed to indicate a statistically significant association (Benjamini et al., 2001).

Due to the growing acknowledgement of the importance of clinical factors in epidemiological association studies (Cavalleri et al., 2011, Johnson et al., 2011b), two assumptions were made for this present study 1) genetic factors could act or influence treatment response via a measured clinical covariate 2) genetic factors could influence treatment response via a pathway un-related to the clinical covariates. We therefore undertook the following analyses i) for genetic factors alone; without the adjustment of any associated clinical covariates and ii) using both genetic factors and any associated clinical covariates.

7.3.1. Univariate tests with treatment outcome for association analysis

To evaluate the individual effect of SNP genotype on outcome, two univariate tests of association were conducted for each of the five SNPs, one making no assumption of underlying mode of inheritance and one assuming an additive mode of inheritance, and the minimum p-value referred to in each case. For univariate analysis of each SNP genotype the Armitage trend test was used (http://ihg.helmholtz -muenchen.de/cgi-bin/hw/hwa1.pl). For the binary clinical covariates (gender, epilepsy type, EEG, imaging initial AED and AED at final follow-up, a Chi-square test (SPSS) or Fisher's exact t-test were used (SPSS and www.langrud.com/fisher.htm) and a t-test (SPSS) was used for the single continuous variable of patient age at recruitment or study admission.

7.3.2. Multiple regression analysis with treatment outcome

The purpose of this analysis was to test for association between each of the five SNPs and 1-year remission in the presence of any clinical covariates found to independently influence treatment outcome, in order to potentially adjust for non-genetic clinical association, with the assumption that these may allow improved detection of any genetic influencers. For

this a multivariate binary logistic regression analysis was performed (SPSS) where two logistic regression models were fitted for each of the five SNPs (a baseline and genetic model) and compared using the LRT test (described in Chapter 3 and 4). The 'baseline' model included clinical factors found significant in the univariate analysis ($P = 0.05$) as covariates. The genetic 'model' was the same but also included a genetic covariate representing an individual SNP. For each SNP the 'genetic model' was again fitted twice, first making no assumption of the underlying mode of inheritance and second assuming an additive mode of inheritance. The minimum p-value was referred to in each analysis.

7.3.3. Survival analysis for time to remission data

Survival or time to event analysis may be a more appropriate analysis as remission could have occurred at any time after starting treatment. Survival analyses can account for censored observations which include i) patients dropping out of the study, ii) death due to a cause that is not the event of interest, iii) termination of the study (the study ends before some individuals have the outcome of interest). Survival analysis may therefore help determine or provide information on which fraction of the population will remit past the final follow-up, and the rate at which seizure remission is achieved. Moreover how particular factors benefit or affect the probability of remission can be investigated.

In order to perform a time to event survival analyses using the genetic data available for this study, time to period of 1-year seizure remission was required to be extracted from the available clinical data and treated as an additional covariate. Time to seizure remission (days) was calculated using dates available for both treatment initiation (study recruitment or admission) and time to achieving remission. A Kaplan-Meier survival analysis and a log-rank test were performed to test for a univariate association of remission status for each of the binary clinical covariates (gender, epilepsy type, EEG, imaging initial AED, AED at remission) and the Cox regression test was used for continuous clinical covariates (age at recruitment/admission).

For the multivariate survival analysis, for each of the five SNPs two Cox regression models were built, a 'baseline' model, again containing statistically significant clinical covariates only ($P = 0.05$) and 'genetic model' containing both clinical covariates and genetic covariates for each SNP. A Chi-square p-value was again generated using log likelihood ratio and LRT. For each SNP models were fitted when making no assumption of the underlying mode of inheritance and when assuming an additive mode of inheritance and the minimum p-value referred to in each case.

Finally a bioinformatics analysis was also performed for each of the five-SNPs using several freely available as described previously (section 5.2.14), in order to investigate potential biological significance.

7.4. Results

From the 434 individuals genotyped for the five candidate SNPS, 10 individuals failed to be successfully genotyped for all five SNPs and so were removed from data analysis. Demographics of the remaining 424 patients are summarised in Table 7.3. The majority of patients were Caucasians with newly treated epilepsy and treated with one or more AEDs. Of the 424 patients included in the data analyses, 304 remained seizure free for a period of 12 months or more at some point through their treatment, while 120 continued to experience seizures, without any period of seizure freedom of at least a year, during their period of follow up. Hence of the 424 patients 28.3 % achieved remission and were treated as cases while 71.7 % failed to achieve remission and were labeled as controls.

Table 7.3 Characteristics of UK Glasgow cohort of newly treated epilepsy

Clinical characteristic			**(n=424)**
Age	Mean (±SD)		37 (±16.96)
Sex	N (%)	Male	234 (54.7%)
		Female	190 (44.4%)
Epilepsy type	N (%)	IGE	79 (18.5%)
		LRE	318 (75%)
		UNC	27 (6.4%)
Remission	N (%)	Yes	304 (71.7%)
		No	120 (28.3%)

SD= Standard deviation, IGE = idiopathic generalised epilepsy, LRE = localisation related epilepsy, UNC = unclassified epilepsy

7.4.1. Univariate analysis of association between SNPs or clinical covariates and treatment outcome

For the final cohort of n=424 patients, population MAF of each SNP was at least 5%, and each SNP achieved HWE. All five SNPs were previously typed by the International HapMap project (NCBI build 36, dbSNP build 126) and HapMap population MAFs (HapMap-CEU European ancestry) did not deviate from those observed in this study. When each of the five SNPs were analysed univariately as to test for an independent effect of each of the five variants, none of the five SNPs were found to individually influence treatment outcome (Chi-square $P > 0.05$). Results of the univariate tests of association for the five genetic covariates are presented in Table 7.4.

Non-genetic influence on outcome was also tested univariately. Of the binary clinical covariates included in the analysis both epilepsy type and EEG result influenced treatment outcome, with a statistically significant p-value of below 0.05 before FDR. Final treatment drug (AED at final follow-up) additionally showed a strong association with treatment outcome ($P < 0.001$ after FDR correction). These clinical covariates were thus included in the regression models for the subsequent genetic analyses (Table 7.5). Figures 7.3, 7.4 and 7.5, present plots for the influence of epilepsy type, EEG and AED at final follow-up on response to drug treatment.

Table 7.4 Univariate analysis of genetic and clinical factors with drug response

Univariate regression analysis for the independent association of genetic variables and clinical covariates with treatment response in newly treated epilepsy.

Factor	Uncorrected *P*-value
Final follow-up AED	0.000
Initial follow-up AED	0.432
Epilepsy type	0.018
EEG	0.002
Age	0.52
Gender	0.867
Imaging	0.21
rs17252760	0.91
rs12919774	0.62
rs16994558	0.22
rs316132	0.78
rs622902	0.5

AED = antiepileptic drug, EEG = electroencephalography

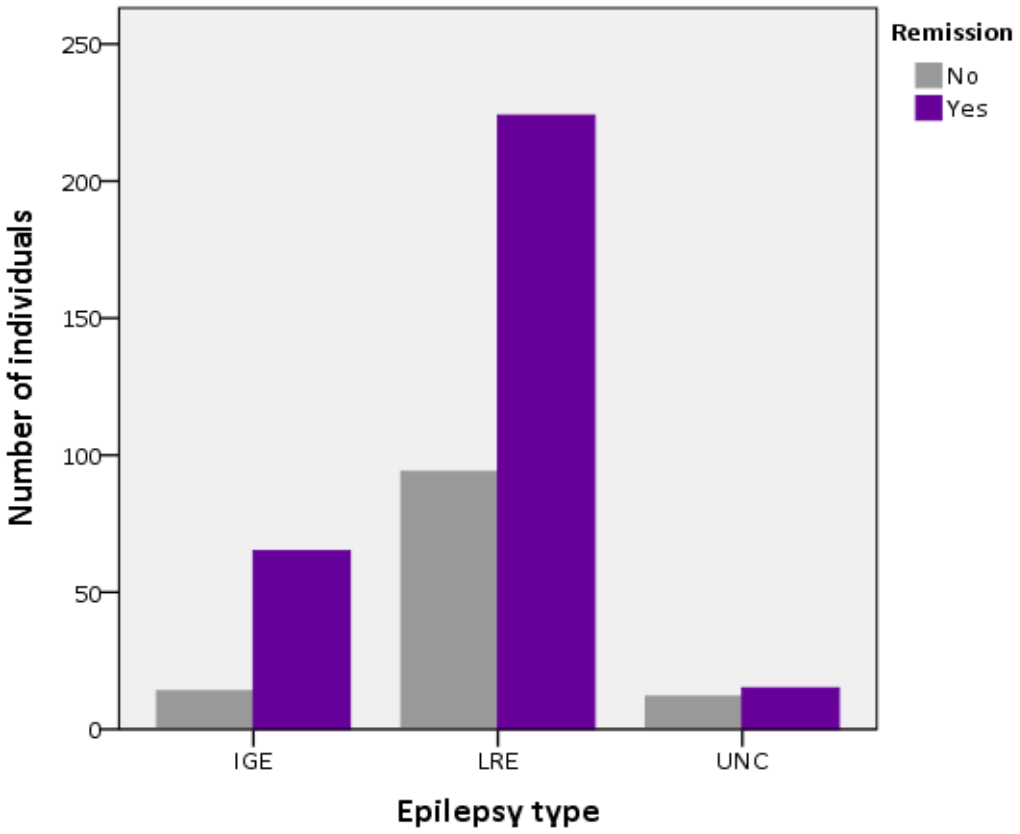


Figure 7.3 Bar chart for association between epilepsy type and response (n=424)

In total 79, 318 and 27 individuals presented with idiopathic generalised epilepsy (IGE), localisation related epilepsy (LRE), and unclassified epilepsy (UNC), of which 65, 224 and 15 individuals achieved 1-year remission respectively.

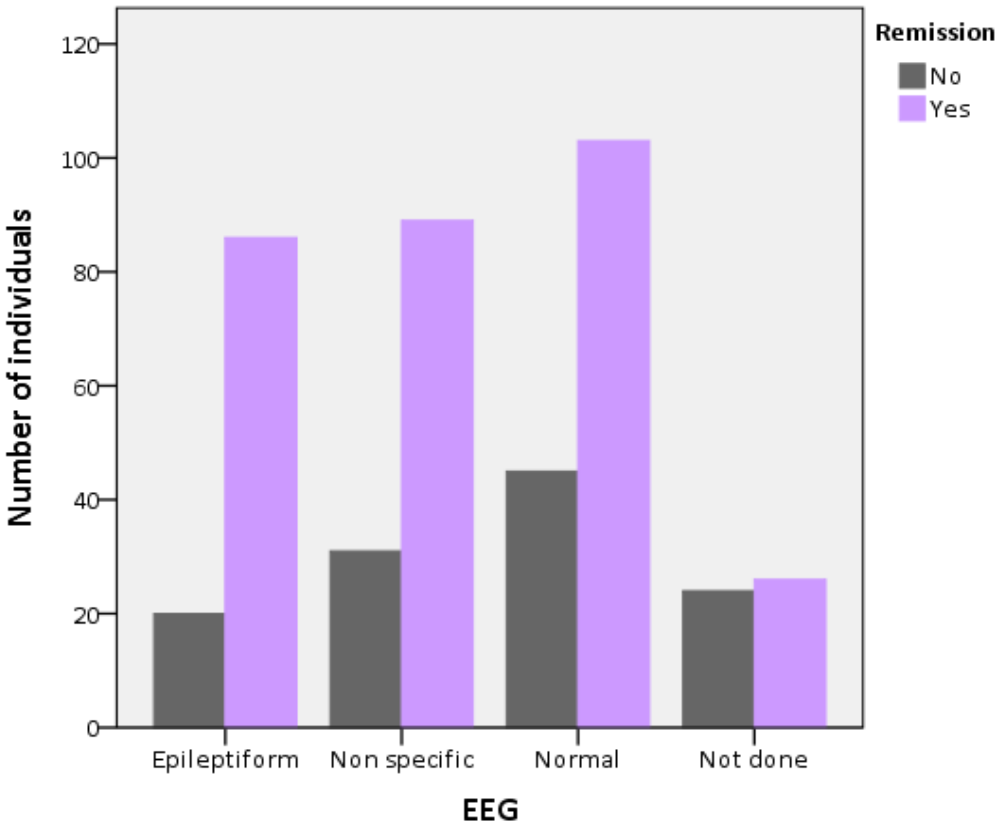


Figure 7.4 Bar chart for association between EEG recording and response (n=424)

In total 106, 120 and 148 individuals presented with epileptiform, non-specific and a normal EEG result respectively. In 50 individuals an EEG (Electroencephalography) was not done. Of these 86, 89, 103 and 26 individuals achieved 1-year remission respectively.

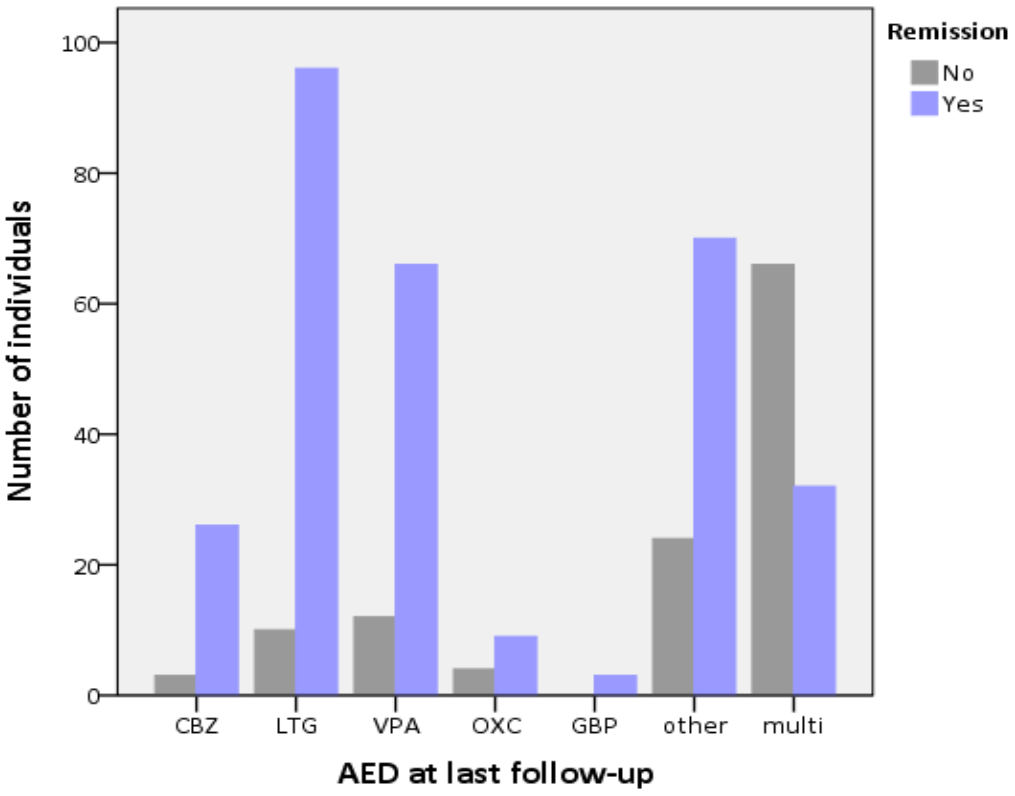


Figure 7.5 Bar chart for association between AED at final follow-up and response

In total 29, 106, 78, 13, 3, 94 and 98 individuals were treated with the antiepileptic drugs (AED) carbamazepine (CBZ), lamotrigine (LTG), valproate (VPA), oxcarbazepine (OXC), gabapentin (GBP), another drug (other), or multiple drugs (multi). Of these 26, 96, 66, 9, 3, 70 and 32 individuals achieved 1-year remission respectively.

7.4.2. Multiple regression models

Results of the LRT from the multivariate binary regression analysis are summarised in Table 7.5. Of the five candidate SNPs, none of the SNPs were significantly associated with treatment outcome (P = < 0.05, before FDR) when the associated clinical covariates of EGG, epilepsy and final follow-up AED were included in the genetic model. Thus none of the genetic variants were predictive of treatment response in our investigation. The genomic information for all five candidate SNPs are summarised in Table 7.6.

Table 7.5 Multivariate logistic regression results for treatment response

Association of each SNP with treatment response in newly treated epilepsy in the presence of associated non-genetic factors

Clinical covariates	**SNP ID (rs)**	**Uncorrected Chi-square *P*-value**
Epilepsy type, EEG, Final follow-up AED	rs17252760	0.356
	rs12919774	0.127
	rs16994558	0.493
	rs316132	0.346
	rs622902	0.093

SNP= single nucleotide polymorphism, rs= reference sequence, EEG= electroencephalography, AED= antiepileptic drug

Table 7.6 Genomic information for the five candidate SNPs investigated

SNP ID (rs)	Chr.	Position (bp)	Closest gene	Allele change	SNP location
rs16994558	23	147259820	*RPL7L1P11, AFF2*	G/A	intergenic
rs622902	6	52954433	*GSTA4*	C/T	intronic
rs316132	6	52955925	*GSTA4*	A/G	intronic
rs12919774	16	8455709	*TMEM114, LOC100131080*	A/G	Intergenic
rs17252760	15	89784049	*MAGEA9B, CXORF6840A*	C/A	intergenic

SNP= single nucleotide polymorphism, Chr= Chromosome, bp= base pair, MAF= minor allele frequency

Table 7.7 Survival analysis univariate Kaplen-Meir and Cox regression analysis of association between clinical covariates and time to 1-year remission

Factor or variable	Uncorrected *P*-value
Age	0.430
Gender	0.047
Epilepsy type	0.448
EEG	0.069
Imaging	0.556
Initial AED treatment	0.058
AED at Remission	$1.4x10^{-5}$

EEG = electroencephalography, AED = antiepileptic drug

7.4.3. Multivariable survival analysis

The results of the Cox regression analyses are presented in Table 7.8. The multivariable Cox regression models including patient gender and AED at remission as covariates, again none of the five SNPs were associated with time to 1-year remission (LRT; *P*= > 0.05, for each SNP before FDR correction, Table 7.8). Therefore although there was some predictive potential of the clinical covariates for time to seizure control there was no influence of any genetic factors on time to remission in our analysis.

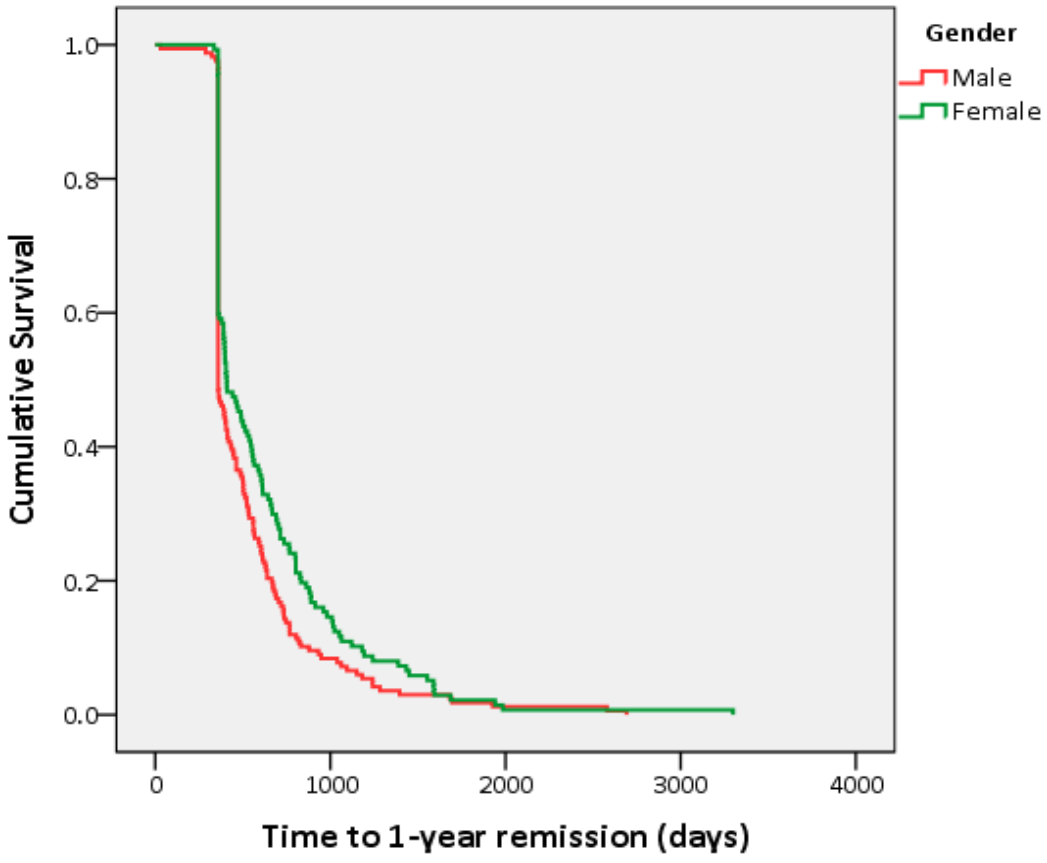


Figure 7.6 Time to 1-year remission analysis for patient gender

Kaplen-Meir curves as generated by SPSS for univariate analysis of time to 1-year remission and patient gender.

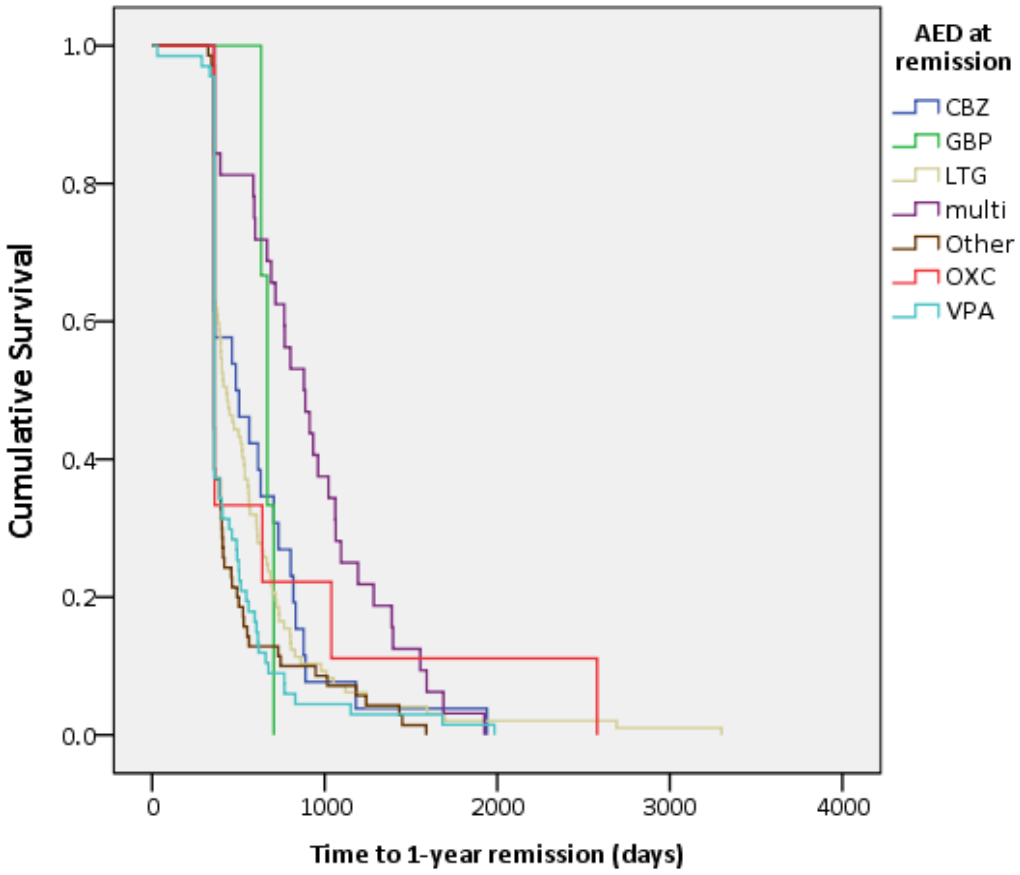


Figure 7.7 Time to 1-year remission analysis for AED at remission

Kaplen-Meir curves as generated by SPSS for univariate analysis of time to 1-year remission and antiepileptic drug (AED) recorded at 1-year remission. Data stratified by AED (antiepileptic drug) recorded at 1-year remission. CBZ = carbamazepine, LTG = lamotrigine, VPA= valproate, OXC= oxcarbazepine, GBP= gabapentin, Other= any of AED administered and Multi= multiple drug treatment.

7.4.4. Multivariable survival analysis

The results of the Cox regression analyses are presented in Table 7.8. The multivariable Cox regression models including patient gender and AED at remission as covariates, again none of the five SNPs were associated with time to 1-year remission (LRT; *P*= > 0.05, for each SNP before FDR correction, Table 7.8). Therefore although there was some predictive potential of the clinical covariates for time to seizure control there was no influence of any genetic factors on time to remission in our analysis.

Table 7.8 Multivariate Cox regression analysis for SNP association with remission
Association between each SNP and time to 1-year remission with adjustment for the associated clinical covariates identified previously.

SNP ID (rs)	Clinical covariates	Chi-square value	Uncorrected Chi-square *P*-value
rs17252760	Gender AED at final follow-up	0.851	0.356
rs12919774		2.328	0.127
rs16994558		0.469	0.493
rs316132		0.888	0.346
rs622902		2.829	0.093

SNP = single nucleotide polymorphism, AED = antiepileptic drug

7.5. Discussion

With a significant amount of individuals (30-40%) presenting non-response and long-term resistance to adequate management with AEDs, the search for markers of drug efficacy in epilepsy has received considerable attention in the challenge to optimise therapeutic drug treatment. Outcomes AED therapy can range from immediate remission to frequent unremitting seizures with multiple treatment failures and so delineating both seizure susceptibility and a predisposition to achieving good seizure control early on remains to be attained Duncan et al., 2006. There is a significant lack of study replication across epilepsy PGx that can be attributed to many limitations presented to the study of complex disease

genetics in general (Tabor et al., 2002, Loscher et al., 2009) and/or that specific to epilepsy PGx research (see section 1.5.6). The candidate gene approach to association study design has furthermore been disappointing in terms of locating influential susceptibility loci and incorporating a multifactorial basis of drug response phenotypes. With this slow progress in predicting AED efficacy including treatment responsiveness, large-scale GWAS with bigger numbers of cases and controls creating more power for genetic detection are being utilised.

Recent GWAS methodologies include two-stage approaches, with a discovery cohort and independent replication cohorts and combined meta-analysis of the two stages. Recent guidelines have been published to standardise the reporting of association studies to facilitate meta-analyses (Little et al., 2009). Moreover inclusion and/or stratification by clinical phenotypes could reveal otherwise undetected linkage (Sanchez et al., 2010, Johnson et al., 2011b, Leu et al., 2012). GWAS meta-analysis powered by international collaboration can create sufficiently sized samples for specific drug response phenotypes in epilepsy populations (Johnson et al., 2011b). Further independent validation is also vital to confirm the initial research findings and for establishing the true genetic causality of any GWAS presented association (McCarthy et al., 2008).

In order to provide an independent population investigation or partial validation of the UK-Australian GWAS for treatment response in newly treated epilepsy, we performed a modest candidate SNP validation study using the Glasgow cohort as an independent UK population of newly treated epilepsy. Five SNPs were selected from a number of new potentially influential loci from the GWAS meta-analysis for response to drug treatment in epilepsy. This included two chromosome 6, *GSTA4* SNPs that were identified to be potentially significant at a genome wide level (p value of at least $5x10^{-6\,-7}$, with and without inclusion of outcome associated clinical covariates) and an additional 3 GWAS associated intergenic SNPs, rs16994558, rs12919774 and rs17252760 (a 5' to *AFF2* variant, a SNP near 3' of *TMEM114* and also the ribosomal protein S14 pseudogene gene and a final SNP near the CXorf40A locus, the iduronate 2-sulfatase pseudogene (*IDSP1*) and the *MAGEA9B* gene respectively) (Table 7.8). None of the five SNPs were found to individually predict treatment outcome (remission status) or time to 1-year remission in our analysis using an independent UK cohort. Our findings therefore suggest little potential significance of the five SNPs to seizure control in newly treated epilepsy, when considered in isolation.

Given the independent influence of patient clinical characteristics on drug response phenotypes in epilepsy (Callaghan et al., 2011) there is a growing appreciation of the incorporation of these factors in current epilepsy PGx research (Cavalleri et al., 2011, Grover et al., 2011). The concept of potential interaction of both genetic and clinical variables may help detect the remaining gene-phenotype association that could exist and/or confirm some of the minor genetic effects already reported (Kasperaviciute and Sisodiya, 2009, Loscher et al.,

2009). With this additional multivariate testing however, we again didn't find associations with the five SNPs in both the time to 1-year remission analysis and treatment response outcome (success of seizure control) in the regression analysis. Our findings demonstrated some predictive potential of several clinical factors in AED drug response which was similarly suggested by the Australian/Australian GWAS analyses and also by previous clinical epidemiological studies (Mohanraj and Brodie, 2006, Callaghan et al., 2007, So, 2011, Brodie et al., 2012) whilst the influential potential of the five candidate SNPs filtered out from the GWAS analysis was nonexistent. Similarly no biological significance of these five SNPs was apparent from the bioinformatics analysis (literature searches and online bioinformatics prediction tools). Our investigation therefore did not confirm the significance of the genetic variants or a role of *GSTA4* on seizure control or treatment success in newly treated epilepsy. Time-to remission, which was not studied in the original Australian-UK GWAS study, though provides a better outcome measure. To some degree time- to event data may better represent ease of seizure control and could more accurately indicate whether a patient is likely to remit. No associations were nonetheless evident in the genetic analysis using time to 1-year remission data. Because of a lack of genotype-phenotype associations and insignificance of any potential biological function of either *GSTA4* SNPs or any of the intergenic variants, the role of *GSTA4* in drug treatment in newly treated epilepsy cannot be confirmed by our study.

A number of clinical predictors for seizure remission have previously been investigated for their association with long-term epilepsy outcome (Hitiris et al., 2007, Callaghan et al., 2011, Brodie et al., 2012). Although the genetic subset of analysis performed for this chapter was largely unremarkable, some clinical covariates also appeared to influence treatment outcome or response in our investigation. A diverse range of clinical predictors have previously been proposed as poor prognostic indicators, including abnormal neurological examination, EEG and brain imaging (Sillanpaa, 1993, Berg et al., 1996, Mattson et al., 1996, Berg et al., 2001), a large number of pre-treatment seizures (Sillanpaa, 1993, Kwan and Brodie, 2000a, Leschziner et al., 2006, Hitiris et al., 2007) (including their number and frequency), presence or absence of a neurological deficit and can also explain some of the variability in remission among patients (Callaghan et al., 2011).

The most consistent determinant associated with long-term epilepsy outcome is however the ease of controlling seizures, which includes i) how soon seizures are controlled by AED, ii) how frequent seizures recur despite treatment initiation and iv) how many AEDs had to be used to control seizures (Mohanraj and Brodie, 2006, So, 2011, Brodie et al., 2012). Although information on pre-treatment seizures and seizure frequency was not available for all patients in our cohort, we found associations with epilepsy type, EEG findings and AED at final follow-up with treatment outcome in the regression analysis.

AED at remission was also associated with time to time to 1-year remission in the

survival analyses, as was patient gender. Better response to treatment appeared evident with the IGE epilepsy type (82%) and slightly better when EEG showed epileptiform abnormality in comparison to non-specific abnormality or no abnormality (81%, 74%, and 70% respectively). Highest response was moreover evident with the use of VPA as the final AED (85%), and VPA also presented with the shortest time to remission in the time-to event analysis (474 days). Moreover poor response was suggested with multiple AED treatment (33%) and poly-therapy also showed the longest time to remission (913 days). Finally a marginally significant difference in time to 1-year remission appeared between the genders. These factors have previously been investigated and in several cases were demonstrated to be of some clinical utility (Mohanraj and Brodie, 2006, Johnson et al., 2011b, So, 2011).

Failure to provide evidence for the significance of the five candidate SNPs and validation for predictive potential of the two *GSTA4* GWAS filtered out SNPs can be attributed first and foremost to the methodological issues that are evident with the methodology of our study and also that of which can be associated with the original GWAS analyses. There are several universally recognised issues associated with whole genome based association studies for PGx (discussed previously in Chapter 1) (Motsinger-Reif et al., 2010). The main recognised disadvantage relates to the large sample size requirement (Nebert et al., 2008a, Motsinger-Reif et al., 2010). Sample size limitations are a challenge in any GWAS study of complex traits that attempt to detect modest effect sizes, but are amplified in many pharmacogenomic studies (McCarthy et al., 2008, Crowley et al., 2009, Khoury et al., 2009) and even more so when performing independent validation studies (Ioannidis et al., 2001). Consortia efforts such as that used for the newly treated epilepsy GWAS meta-analysis are suggested to overcome the limitation of sample size and power (Motsinger-Reif et al., 2010, Johnson et al., 2011b). Although modest sample sizes were provided by the meta-analyses performed for the Australian-UK GWAS study, (discovery cohort, n=831 replication cohort n=260) and that of our validation study cohort (n=424), there remains the possibility of limited study power, for detecting potential effects across all 3 cohorts. In addition to this, the potential effect of population substructure across at least the Australia and the UK cohorts may pose further errors in true effect identification, as there may be variation in patterns of LD and frequency of alleles of interest between the discovery and validation population (Johnson et al., 2011b).

As discussed throughout this thesis, defining phenotypes for drug response remains one of the main challenges in epilepsy PGx and the importance of a universally accepted definition and the collection of phenotype data has become increasingly appreciated in the context of GWAS (Nebert et al., 2008a). The widely accepted 1-year period of seizure freedom definition used for our research dichotomises drug response or treatment outcome into "responder" and "non-responder" categories. Although, this has served extremely useful for

the retrospective approach to epilepsy PGx (does not require lengthy follow-up) and is also appropriate, given that absolute seizure freedom for at least 12 months is the only relevant outcome associated with meaningful improvement of quality of life (Kwan et al., 2010). This dichotomisation however may obscure important information for drug response (Johnson et al., 2011b), thus a potentially quantitative trait may be being transformed into a binary trait and so result in the power of a case-control association study being less than that of a quantitative trait analysis (Yang et al., 2010, Johnson et al., 2011b). In this present study a time to event analysis was additionally undertaken as an attempt to resolve this issue of defining seizure remission.

In addition to these general issues with study design there are other potential confounders more specific to this validation attempt, namely the approach taken for selecting the five candidate SNPs to validate the GWAS findings. These SNPs may not be the most influential to treatment outcome in newly treated epilepsy. The five SNPs were selected from the unreported GWAS meta-analysis findings, though these were not necessarily the top five 'GWAS hits' (GWAS significance level of $P = 5x10^{-7}$ for suggestive phenotype association) (Meyer et al., 2010). Our five candidate SNPs were selected based on results of gene expression data in accordance to strong association with outcome (not all at the suggested level of GWAS significance) and with the additional consideration of clinical covariates for detecting an improved signal. This approach to SNP selection was for the purpose of adding further biological or functional value to the GWAS analysis findings on which this study was based. The initial GWAS meta-analysis report identified 10 *GSTA4* SNPs that associated with treatment outcome at the GWAS significance level, however only the top two in terms of lowest Chi-square p-value for association were studied in this investigation. With the potential role of *GSTA4* gene signified from earlier research and the GWAS analysis, more value may have been achieved from investigating all 10 *GSTA4* SNPs in addition to those with additional potential importance due to gene expression data.

GSTA4 appears to be a good candidate gene for influencing drug response in epilepsy due to its role as a defense enzyme against pharmacologically active electrophilic compounds. Further investigation into *GSTA4* may thus be warranted to confirm the potential of this gene to AED treatment response. Rather than limiting to a handful of SNPs, genotyping a greater number of GWAS associated SNPs with slightly less predictive potential (higher Chi-square association values) would provide a richer dataset and so may yield additional new causal loci associated with treatment outcome. Moreover carrying out a candidate gene validation study for *GSTA4* combining all three independent cohorts would add greater power for detecting influential markers across phenotypically heterogeneous populations.

It is unlikely that a small selection of common variants can successfully predict AED responsiveness alone. Several clinical factors are known to determine drug resistance in

epilepsy, including aetiology, early age at seizure onset, type of epileptic syndrome and seizure, structural brain abnormalities or lesions, high pre-treatment seizure frequency, or abnormal EEG findings (Kwan and Brodie, 2002, Loscher, 2005b, French, 2007). Not all of the above mentioned clinical variables were however included in our study, due to unavailable data. This included pre-treatment seizure frequency (most associated with drug-resistant epilepsy (Callaghan et al., 2011) and also identified from the GWAS meta-analysis) and neurological deficit (Callaghan et al., 2011), which, if significantly associated with outcome or time to remission could have influenced our results.

A more comprehensive analysis with complete clinical data sets of all variables of particular significance may be required in future analyses. The stratification of patient cohorts into subgroups by such demographic and clinical factors has also been suggested as a means of acknowledging both the heterogeneity within patient populations and the concept of revealing any masked genetic influence. Most studies are performed in multiple types of epilepsy and with multiple AEDs, but like many association studies no attempt was made in our analyses to separate children from adults or the different epilepsy types, despite there being important differences in drug-response and type of AEDs used in each subgroup.

Treatment differences between the GWAS meta-analysis and validation cohorts are another obvious confounder. Potential differences in methodology and recruitment of individuals from one study to another may have additionally influenced our results. Significant differences in drug treatment are to be expected with international collaboration studies in PGx in general. The Australian clinical drug treatment regimens are largely based on CBZ, VPA and LTG. Both UK cohorts also differed slightly in AED treatment due to the source populations from which both UK cohorts were derived. As previously described the UK Glasgow newly treated population (from which patients from this validation study were selected) was largely recruited via a drug trial comparing VPA to LTG, whilst the Australian GWAS source population was not recruited as part of a drug-trial, rather a PGx study and were mainly prescribed CBZ or VPA, and the SANAD GWAS source cohort was the two arm randomised SANAD drug trial of established and newer antiepileptic medications (Marson, et al. 2007a, b), where LTG again dominated initial treatment. This would explain the highly significant association identified with AED treatment in both our survival analysis for time to 1-year remission and the general regression analyses for predicting response to treatment (see chapter 5 for a discussion on differences in AED treatment between populations and the potential genetic effects). We performed univariate and regression analyses on the Glasgow UK population using data on initial AED treatment, final AED treatment and/ or AED at remission, but this was not stratified by drug type, rather each treatment variable was treated as a single covariate.

GWASs can not only identify novel associations for further study, but can help counter the selective reporting and pursuit of false positive findings that may occur when PGx studies are limited to candidate genes (Guessous et al., 2009). PGx has however focused mainly on SNPs as a source of common variation. Single and common variants within genes alone have poor predictive value, it is thus unlikely only common polymorphisms and only five variants at that are solely responsible for the different drug response phenotypes. GWAS are generally underpowered, to detect effects of other sources of variation such as rare and novel SNPs that are now becoming appreciated to also contribute to the genetic heterogeneity of drug response.

To summarise future research efforts for predicting seizure remission are likely to benefit from the use of broader definitions of response in larger patient cohorts. This may power detection of modest effects for a common biological pathway for all 3 populations, but a strategy that i) considers underlying epilepsy etiology, ii) incorporates clinical covariates (e.g., seizure/syndrome type, age at onset) or at the very least considering all strongly implicated clinical factors (i.e. pre-treatment seizure frequency and neurological deficit)(Blanca Sanchez et al., 2010, Cavalleri et al., 2011, Johnson et al., 2011b) and iii) classifies patients according to response to a specific AED should help in the identification of potentially larger and more clinically relevant AED-specific genetic effects (Cavalleri et al., 2011).

In addition to large collaborations and consortia meta-analysis, in recent times studies are now taking advantage of large-scale deep genome-sequencing to develop a better understanding of the human genome and these are better powered to detect mutations and rare variants, and it is very likely that such approaches will also be used in the future for pharmacogenomic studies in epilepsy.

CHAPTER EIGHT

FINAL DISCUSSION

CONTENTS

8.1 Recent progress in antiepileptic drug pharmacogenetics

The first AED was PB, discovered to have antiepileptic properties in 1850s and PHT, was then later developed in 1912. Epilepsy is now typically managed with AEDs, of which there are currently over 20 with around a dozen in common use (Rogawski and Porter, 1990, Cavalleri et al., 2011). Although the recognition that genetic factors play a role in individual response to AED therapy came about in the 1960s with the discovery of congenital enzyme deficiency, (Kutt H, 1968) it was not until the late 90s that the first association with genetic polymorphisms in the phase I enzymes CYP2C9 and CYP2C19 with PHT metabolism was reported (Mamiya et al., 1998, Aynacioglu et al., 1999, Nakajima et al., 2005). This was confirmed by several other experimental studies (Aynacioglu et al., 1999, Mamiya et al., 2000, Kerb et al., 2001, Allabi et al., 2005) and initiated numerous searches for additional DME candidate genes for AED response (Saruwatari et al., 2010).

In terms of AED efficacy or response phenotypes, *CYP2C9**2 and *3 polymorphisms were the first variants implicated in altered AED dosing in early research concerning PHT PK (Odani et al., 1996, Odani et al., 1997, Mamiya et al., 1998) and also more recently confirmed by several lines of evidence not only for PHT, but also for the AEDs, PB, VPA and ZNS (Mamiya et al., 2000, Hung et al., 2004, Tate et al., 2005, Chaudhry et al., 2009, Loscher et al., 2009, Saruwatari et al., 2010). Additional PK candidate genes were later investigated and associations with AED dose have over the years have been found for CBZ PKs with *CYP3A5**3 and *EPHX1* and also variants within *UGT1A4* and *UGT2B7* for LTG (Loscher et al., 2009, Saruwatari et al., 2010). The *ABCB1* transporter PK gene was initially implicated in AED dosing in 2001 (Kerb et al., 2001, Ebid et al., 2007, Simon et al., 2007). The original *ABCB1* C3435T variant association with AED responsiveness in individuals with epilepsy (Tishler et al., 1995) was subsequently proposed and demonstrated experimentally in 2003 (Siddiqui et al., 2003). Additional transporters associated with AED response include *RALBP1*, P1 and P2 proteins with drug resistant epilepsy (Awasthi et al., 2005, Leschziner et al., 2007b, Soranzo et al., 2007, Kasperaviciute and Sisodiya, 2009, Loscher et al., 2009)

The search for novel candidate genes and functional variants in these and previously associated genes continues, but it quickly became clear that PK polymorphisms alone do not explain most of the variation in AED dosage or efficacy (Depondt and Shorvon, 2006, Kasperaviciute and Sisodiya, 2009). The potential of mutations in drug targets to AED treatment, namely the *SCN1A* gene was initially discovered through early studies of Mendelian epilepsies (Guerrini et al., 1998), and in 2005 the first drug target association with AED dosage was reported (Tate et al., 2005, Tate et al., 2006).

Most recent candidate genes implicated in influencing AED efficacy include that for the *GST* gene (although so far only concerning adverse effects) (Ueda et al., 2007, Zaccara et

al., 2007, Kasperaviciute and Sisodiya, 2009), *OCTN1* (Szoeke et al., 2006, Loscher et al., 2009) *ABCC2* (Kim et al., 2010, Qu et al., 2012) and *GABARA1* all of which are associated with AED responsiveness (Kumari et al., 2010). Despite these gene implications the first and only clinical impact of genetic variation in epilepsy is the *HLA-B*1502* variant as a strong predictor of CBZ-induced SJS/TEN in patients from Asia and of Asian descent (Loscher et al., 2009, Yip et al., 2012). No such progress has been made in AED efficacy. The limited success in AED PGx studies to date has been associated with the lack of concordance in research findings, the foremost reason being the variation in treatment regime among clinicians, inconsistent phenotype definitions (i.e., definition of resistance versus response to AEDs) and heterogeneity in epilepsy phenotypes among studies.

Drug response phenotype is perhaps the fundamental factor in PGx genetic studies of responsiveness and a lack of consensus in its definition has inevitably resulted in difficulties in making comparisons across studies (Kasperaviciute and Sisodiya, 2009, Cavalleri et al., 2011). Two meta-analyses have been reported assessing the role of the Pgp ABCB1 3435C>T variant in drug response (Siddiqui et al., 2003, Leschziner et al., 2007a) and these clearly demonstrate the importance phenotype definitions when investigating potential genetic effects. Neither of these meta-analyses provided evidence for ABCB1 3435C>T association with multidrug resistance, though this can be attributed to the huge variation in drug-resistance phenotype definitions of the original studies that makes meaningful meta-analysis hardly possible (Kasperaviciute and Sisodiya, 2009).

There are various other potential explanations for the discordant results, including, retrospective design and relatively small sample size and/or short duration of most studies. Additional factors include heterogeneity of the epilepsy syndromes with their variable causes and prognoses and a reduction in power to high-dimensionality of multigenic data sets under investigation (Hirschhorn et al., 2002, McCarthy et al., 2008, Kasperaviciute and Sisodiya, 2009).

The multifactorial nature of AED response necessitates the search for multiple genes or variants (Loscher et al., 2009, Kumari et al., 2011). Indeed several investigations including some of the studies mentioned above have utilised a multigenic approach (Anderson, 2008, Petrovski et al., 2009, Johnson et al., 2011b). Moreover the application of ML methods has now moved to PGx (Hahn et al., 2003, Ferraro and Buono, 2006, Petrovski et al., 2009, Pander et al., 2010, Silva et al., 2011). A growing number observational studies have also recently shown that when genotype is considered alongside other genomic factors and clinical predictors the proportion of variation in response can increase significantly (Franciotta et al., 2009, Makmor-Bakry et al., 2009, Petrovski et al., 2010, Sanchez et al., 2010, Johnson et al., 2011b).

With the completion of the human genome project in 2001 and the advent of genomic technologies, epilepsy PGx research has now begun utilising a more genome-wide approach to research i.e. one that captures greater amount and type of genetic variation (Evans and McLeod, 2003, Goldstein et al., 2003, Grant and Hakonarson, 2007, Crowley et al., 2009). This genome wide approach has previously proven to speed up the discovery of drug response markers in other disease areas and aided the integration of PGx to the clinical practice (Takeuchi et al., 2009, Daly, 2010a, Wu and Reynolds, 2012). In 2010, the first GWAS for focal epilepsy was published (Kasperaviciute et al., 2010) and this was quickly followed by a GWAS meta-analyses concerning PGEs (Leu et al., 2012). The next step for epilepsy PGx thus remains GWAS studies for AED efficacy as to finally utilise whole genome data in the search for novel prognostic gene and/or SNPs for treatment responsiveness in epilepsy (Johnson et al., 2011b).

Prospective epidemiological study of newly-diagnosed epilepsy across all age ranges, countries, and continents is considered the ideal for studies into drug efficacy (Kasperaviciute and Sisodiya, 2009, Cavalleri et al., 2011, Johnson et al., 2011b). As the analysis of already available retrospective data is likely to continue, for the time-being carefully designed long-term follow-up studies would identify the patterns of outcome and delineate the different phenotypes for successfully identifying drug response markers for valid pharmacogenomic investigations. Multicentre GWAS meta-analyses are the way forward for this, and are presently being constructed (Johnson *et* al 2012 unpublished)(Leu et al., 2012).

A recent Australian-UK GWAS meta-analysis effort for newly treated epilepsy presents the first GWAS study for drug response and seizure remission in newly treated epilepsy (Speed et al., 2013). Through the careful consideration and standardisation of patient phenotypes as well as using prospective drug response data, incorporating influential clinical covariates and utilising multi-centre collaborations, this GWAS study is an advancement in AED PGx. Researchers however conclude a lack of limited study power to detect common genetic determinants of weak to modest effects (Johnson *et* al 2012 unpublished).

The growing consensus in the field of pharmacogenomics and disease genomics that all genomic mutations i.e. common, rare, SNP, CNV, insertions or deletions, microsatellites are possible sources of variability with a combination of small and/or large effects on complex disease/traits. There is also an increasing trend towards there complete ascertainment. This concept of whole genome sequencing is moreover on the rise due to advances in whole genome sequencing technology. A GWAS utilising CNV with encouraging findings has already been reported for complex forms of epilepsy (Mefford et al., 2010). Arguments thus exist for the benefits of the candidate gene approach, genome-wide and now whole genome approach to genetic analysis. The former nevertheless provides an increasingly economical method of locating all common variation, and the latter at present, continues to provide information of

markers of interest (Daly and Day, 2001, McCarthy et al., 2008, Guessous et al., 2009). Besides candidate gene studies are clearly a low-cost method of validating larger scale studies quickly and efficiently. And so in combination, high-quality phenotyping across multiple research centres, dense genomic patient profiles from GWAS and whole-genome sequencing and effective PGx validation of the most influential markers could provide a more comprehensive investigation to finally locate genetic factors that can guide epilepsy treatment (Cavalleri et al., 2011).

8.2 Thesis findings and potential future direction

PGx studies in epilepsy and in particular newly treated patients are currently small in number. The available studies however clearly demonstrate the relatively modest and multigenic influence on AED response that is also largely dependent on individual patient's demographical and clinical characteristics. Our findings similarly indicate a complex interplay of multiple genetic factors from well-known pharmacological pathways of AEDs in combination with clinical prognostic factors underlie the treatment path and responsiveness to AED therapy. We also show the potential benefits of new statistical methods for more efficiently capturing these relatively small effects with better efficiency.

8.2.1 Genetic markers for predicting antiepileptic drug dose requirement

PGx studies concerning AED dosing have largely concentrated on PK DME genes and primarily the genes for the CYP2C9 and CYP2C19 enzymes (Loscher et al., 2009). The first PD gene studied was *SCN1A* in a report that found a splicing variant rs3812718 to be related to variable CBZ and PHT doses (Tate et al., 2005). In Chapter 3 and 4 dosing of AEDs was studied using maintenance dose and/or maximum dose data in two independent investigations that either searched for potential markers associated with AED dose in PK candidate genes (CBZ DMEs) or PD genes (AED target; Na_V channel). We failed to find predictive genetic markers for AED maintenance dose in either drug pathway investigation.

In Chapter 3 a comprehensive search for markers within several genes was performed as to capture enough variation across whole genes, yet our search was not successful. A major caveat of the study in Chapter 3 is that we only investigated selected candidate genes that were limited to the PK of CBZ. Previous reports however managed to identify single DME SNP variants that appeared to influence AED dose using a limited candidate SNP search within CYP450 genes namely CYP2C9, 2C19 and 3A5 (Hung et al., 2004, Makmor-Bakry et al., 2009, Park et al., 2009, Saruwatari et al., 2010). Small-scale genetic investigations (i.e. those that assesses only a handful or small selection of candidate variants) is a problem with many

PGx candidate gene efforts as they run the risk of not capturing all potentially influential variants and thus pose a greater chance of reporting false positive or false negative findings (Daly and Day, 2001, Ferraro et al., 2006). Moreover contradictory research findings exist for most if not all studies for AED response (Loscher et al., 2009). Though the predictive power of the CYP2C9 and CYP2C19 variants are proven, these are mainly concerned with PHT metabolism and are yet to demonstrate general clinical utility (Anderson, 2008). The previously reported CYP2C9 and CYP3A5 polymorphisms were not included in our study or the final data analysis after quality control checks, thus we were unable to assess their potential function to CBZ dosing.

Our attempt to replicate the association between maximum dose and the *SCN1A* rs3161362 SNP in Chapter four using retrospective recruited patients with newly treated epilepsy receiving CBZ only was unsuccessful (n=168) (Tate et al., 2005). Similar to our finding two independent research groups using Japanese and Italian patients also failed to associate maximum dose requirements of CBZ and rs3161362 genotype, when investigating drug responsiveness in drug-resistant and responsive patients (Abe et al., 2008, Manna et al., 2011). The primary analysis in Chapter 4 was however for maximum and/or maintenance dose regardless of AED administered. This was performed using all patients thus a larger cohort of individuals (n=586) and proved more successful. Our follow up investigation of the original drug target variant using data for several AEDs in addition to CBZ was thus positive. Regardless of whether drug target or DME gene variation is more influential of AED dose, the most important question is the predictive value of our genetic findings. In Chapter 3 one PK gene, *UGT2B7* was the main gene implicated in influencing CBZ dose, though not statistically proven in the data analysis. The *SCN1A* SNP accounted for 6.5% and 2.5% of the variation in PHT and CBZ dose requirement respectively (Tate et al., 2005), which may not be adequate for it to be considered clinically significant. Moreover our study, which considered a wider selection of drugs similarly only appeared to explain limited dose variation.

The work presented in Chapter 4 also demonstrates the complexity of gene-environment interactions where the type of AED used for treatment was shown to effect maximum dose. When maximum drug dose was stratified by AED type we observed significant association between the *SCN1A* SNP and maximum dose requirement, with some indication of potential specificity to LTG and no effect evident with CBZ (Thompson et al., 2011). This finding was not in line with Tate and colleagues (Tate et al., 2005, Tate et al., 2006) who demonstrated association with CBZ and also with PHT, at a greater extent. Our sample population was considerably bigger than that of Tate *et* al and in addition to this we carried out a non-specific study of AED usage. Using a non-specific drug data analysis provided us with a larger number of patients and thus better power to detect effect size. A non-specific drug analysis moreover removes bias towards a particular AED, which may no longer

be widely prescribed in clinical practice (Glauser et al., 2006, Hakami et al., 2012). Given this drug specific effect that appeared to cause modest dose changes and equally moderate level of dose variability attributable to the rs3812718 SNP, genetic variation in the *SCN1A* gene may have a significant effect on dosing in newly treated epilepsy patients and so necessitates further enquiry (Kasperaviciute and Sisodiya, 2009).

PK is known to be responsible for variable serum concentration of drugs and their doses for therapeutic effect. The *SCN1A* variant genotype can be implicated with higher doses of AED in our analysis. Our data indicates drug-gene interactions may influence drug dose and we also show association between rs3812718 genotype and standardised maximum dose requirement regardless of AED. We did not however take opportunity to explore a multigenic effect of functional DME polymorphisms and the *SCN1A* variant in combination or even multiple variants across SCN1A and/or additional drug target genes, which may have provided more definitive results for the studies in Chapter 3 and 4. With the patient DNA available for the Glasgow samples we will be able to confirm our findings with maximum dose and rs3812718 genotype (investigated in the SANAD cohort) in this independent epilepsy cohort. With the wealth of clinical data available for both cohorts, we could also perform a more comprehensive investigation of the non-genetic contribution to the variability in AED dosage requirement and the multigenic contribution to clinical AED usage.

8.2.2 Utilisation of multigenic machine learning models for AED response

The work presented in both Chapters 5 and 6 demonstrate some advantage in the use of the *k*NN supervised ML approach for analysing complex genomic data. In Chapter 5 we could not independently replicate the predictive potential of Australian multigenic classifier developed by Petrovski *et* al 2007 using UK newly treated epilepsy patients. Through an international collaboration with the study authors, Professor O'Brien and Dr Petrovski in Melbourne, Australia, we did however observe some predictive success of the five-SNP classifier when data was stratified by CBZ and VPA. Similar to Chapter 4 these findings signify a potential drug specific effect in the genetic contribution to AED response and so lend support to the inclusion of non-genetic or clinical covariates in the search for predictive markers in PGx analysis (Sanchez et al., 2010).

The 2007 report by Petrovski *et* al presents the first PGx study utilising the novel statistical method of ML for treatment response (Kasperaviciute and Sisodiya, 2009, Johnson et al., 2011b) thus demonstrating greater power of ML methods for studying the complex phenotype of drug response. In a later study the authors indicated the prognostic value of neuropsychiatric factors for initial 12-month seizure control and additionally reported that the five-SNP classifier presented greater prognostic value when considered alongside this

neuropsychiatric data (ABNAS score) (Petrovski et al., 2010). The consideration of the complexity of epilepsy related phenotypes in genomic research is finally being recognised as fundamental to accurately predicting treatment response. Researchers have now proposed future efforts should also consider etiology as a covariate in analysis of responsiveness to specific AEDs (Kasperaviciute and Sisodiya, 2009, Cavalleri et al., 2011). Great consideration of clinical covariates that associate with treatment response was similarly given in our studies thus we performed extensive data stratification before the development of any prognostic models (Sanchez et al., 2010). Drug responsiveness in epilepsy has previously been studied greatly in the context of drug-resistance or the unresponsive phenotype in individuals with chronic epilepsy. This research moreover lends most focus to Pgp, however not all AEDs are thought to be substrates of Pgp and similar discord is present for the influence of Pgp variants. However the notion is presently that if Pgp does have an effect on AED responsiveness this is likely to only be minute in nature (Kasperaviciute and Sisodiya, 2009).

In Chapter 6 the *k*NN ML was also used for investigating genetic factors influencing complex epilepsy syndromes. The clinical utility of prognostic markers or models can only be considered after independent verification. The 11-SNP and 16-SNP kNN classifiers that were developed for PGE sub-syndrome type (i.e. JME vs. CAE) and PGE respectively in our studies were however unable to classify individuals with any great deal of confidence. Similarly utilisation of other ML approaches we could not confidently identify predictive markers for PGE sub-syndrome. Two widely used ML developed classifiers (NN and Ensemble) were however able to predict PGE patients from non-PGE controls with greater confidence. This was not the primary study aim but demonstrates that ML methods are likely to be more suitable for epilepsy phenotype data.

One cannot over emphasise the significance of sample size for complex phenotypes (Cavalleri et al., 2011, Ferraro et al., 2012). This is moreover exaggerated when considering independent validation, which requires additional patient groups. An advantage of ML approaches in this context is thus the ability to test any results internally in separated data before developing an overall prognostic model, and then independently testing in other cohorts before suggestion of any genomic association. The number of cases in a test cohort was thus crucial for both studies and can be considered a major contributor to *k*NN model failure. To improve on the investigation in Chapter 5, for future investigations the UK cohorts can be combined to form a large cohort of newly treated epilepsy in which new variants with potentially more significance to a UK epilepsy population can be identified. Several ML approaches are available including those that model gene-gene interactions that are likely to occur in complex data sets and these may be more suited to developing models for predicting disease occurrence.

8.2.3 Genome wide association study for newly treated epilepsy and drug responsiveness

Although we could not establish any predictive potential of the five-SNPs proposed to be of particular significance to treatment response in newly treated epilepsy (Speed et al., 2013) in Chapter 7, our findings did confirm the non-genetic influence in epilepsy treatment prognosis. The prognostic potential of several clinical covariates in the ease and early success of AED treatment was demonstrated. With our case control analysis of 12-month remission status, the predictive potential of EEG, epilepsy type and treatment AED was identified. This is in line with what is indicated in literature (So, 2011), where as of yet only clinical predictors of modest effect have been located (Loscher et al., 2009, Callaghan et al., 2011, So, 2011).

In a recent study several factors were associated with a decreased cumulative probability for a 12-month or greater seizure remission including presence of developmental delay, symptomatic generalised epilepsy syndrome, longer duration of intractability, and most notable number of AEDs failed which was also an independent negative predictor of seizure remission (Callaghan et al., 2011). We did not perform such a detailed non-genetic analysis; however there were a greater number of responders in both LRE and IGE patients and the least number of responders with multiple AED treatment (usually an indication of failure of at least 2 AEDs) in our study cohort (Kwan and Brodie, 2001a, Kwan and Sperling, 2009).

Chapter 7 specifies that to develop the most accurate predictive models for treatment outcomes, multiple sources of information should be integrated including clinical characteristics genomic data, historical data and neuropsychiatric data (Johnson et al., 2011b). This is consistent with the concept that the determinants of seizure recurrence are multifactorial; therefore, many different non-genetic covariates as well as genetic variants are likely to provide optimal prediction for numerous individual patients with epilepsy (Bhathena and Spear, 2008).

Because both our results and that of the GWAS findings have validated the value of previously recognised predictors in newly treated epilepsy, the GWAS and the results for this current Chapter appear to be valid with little chance of false positive findings. Chapter 7 also shows that with time to 12-month remission data a more accurate representation of seizure control can be provided than that provided by the binary classification of response (Johnson et al., 2011b). Whilst no genomic advances were made in this chapter, *GSTA4* can be implicated as a candidate gene for seizure control from the original GWAS study report and so would benefit from further analysis. Moreover the study presented in Chapter 7 is the first to analyse data from a GWAS analysis for newly treated epilepsy. Although no genomic significance was identified from the initial GWAS investigation, there appears some speculation of the *GSTA4 gene* and the rs622902 variant in treatment responsiveness in

epilepsy, not necessarily from our data but that of the original GWAS international meta-analysis effort. Further work can include a comprehensive analysis of variation across the *GSTA4* gene in both UK cohorts of newly treated epilepsy available to us, using a tSNP approach to search for novel susceptibility markers for AED responsiveness.

8.2.4 Research conclusions

Several approaches were employed for each study. Genomic variation was considered in drug pharmacokinetic proteins, in drug target proteins and finally across multiple candidate genes, thus providing three broad hypothesis as sources housing potential influential genomic variation. In addition to this several methods were considered for SNP selection and genomic data assessment.

Summary of research findings

- Genetic variation in DME genes alone is unlikely to have a significant effect on the dose required for effective drug treatment.

- Genetic variation in neuronal ion channel proteins as AED targets may be more influential in treatment response than pharmacokinetic pathway variation; the previously implicated Na_V channel gene rs3812718 gene polymorphism may alter maximal drug dose requirement of some AEDs but not others.

- Machine learning demonstrates greater power as a novel approach to complex genomic data analysis; Treatment differences across populations and limited cohort size may dampen the predictive power of an Australian treatment response classifier
- Machine learning may also be a better method for modelling genomic data for the common genetic epilepsies, however the PGE sub-syndromes remain far too complex to characterise using candidate genes and modest mixed population cohorts, requiring larger homogenous patient cohorts and whole genome analysis for future research.

- Initial findings of a multicentre GWA study was not validated in an independent UK cohort of treatment response, however *GSTA4* variation could be influential in treatment outcome in epilepsy and the gene warrants further in depth analysis.

8.2.5 Overall thesis conclusions

AEDs provide the main treatment method for seizures in epilepsy yet characteristically present variable levels of success in terms of their effectiveness for treating the many types of epilepsy syndromes that exist. Changes in AED therapy including successive treatment regimen and drug switching have been reported to influence seizure outcome (Mohanraj and Brodie, 2006, Luciano and Shorvon, 2007, Brodie et al., 2012, Wang et al., 2013). The prescribing practice in the context of epilepsy may therefore be a major indicator for the considerable number of individuals who continue to experience seizures (Lhatoo et al., 2001, Sander, 2004, Luciano and Shorvon, 2007, Brodie et al., 2012). AEDs could therefore benefit from the application of the concept of personalised medicine through PGx study as to maximise pharmacotherapy for epilepsy treatment with minimal complication. Better understanding of the common genetic variation contributing to the individual patient differences in response to AEDs has so far provided greater insight of the genomic basis of AED dosing to progress from a trial and error methodology and improve overall drug efficacy in terms of achieving long-term seizure remission. Clinical prognostic indicators are few in number and genomic influencers are more or less non-existent.

This thesis demonstrates that both the control or treatment of seizures in epilepsy and the effective use of AEDs for this intention is complex, and reliant on many factors including common and rare genetic polymorphisms, clinical covariates and environmental interaction, all of which need to be elucidated before any unknown heritability can be detected. Whether such complexity can ever be incorporated into clinical practice is unclear. The characterisation of this heritability is heavily reliant on robust phenotypes of variability for both AED response and the disease itself, as well as cohort collaboration for greater statistical power for detection. A timely effort to effectively achieve these initial steps alone is expected. Progress in AED PGx has not expanded as rapidly as initially anticipated. PGx for improving seizure control however remains a worthwhile ground of epilepsy research as in the long run can ultimately help provide rapid treatment, reduce mortality and morbidity and decrease medical costs which currently burden this common neurological disorder greatly.

8.2.6 Patient Impact of pharmacogenomics

Pharmacogenomics holds the promise of selecting the right drug at the right dose for the right person for better outcomes in terms of successful seizure control, adverse effects and time to remission (Johnson et al., 2011b). From a clinical perspective, identification of genetic variants either by GWA or sequencing is merely the first step in understanding genetic factors influencing individual response to pharmacotherapy. Any identified factors can only be considered for clinical application after robust assessment of each association to determine its

true clinical utility (Kasperaviciute and Sisodiya, 2009). From previous candidate and whole genome complex disease association studies it is clearly evident that despite successful replication of markers even if of high risk or odds ratios, their performance in terms of predictive accuracy and specificity for a clinical phenotype such as treatment outcome remains poor (Ferraro et al., 2012).

One of the growing trends in complex disease genomics is the establishment of the function of any identified genetic variants (Ferraro et al., 2012). Knowledge of biological function provides a good foundation for subsequently interpreting the potential influence of genetic associations on a disease phenotype and so can greatly support any identified genetic association. Functional studies can include those investigating changes in gene expression, splicing and protein function and should be performed in conjunction with genetic studies to improve data interpretation and strengthen data analysis. Another essential endpoint is clinical phenotype in this case pharmacological phenotypes. Pharmacological phenotypes require careful selection before study design and this can be directed through their potential clinical utility i.e. locating genetic variants that can predict ADRs or drug efficacy. Correlation between pharmacological phenotypes and functional genetic variants remains the biggest challenge for future large-scale genomic studies and is critical for clinical translation.

Eventually pharmacogenomics will lead to the development of rapid high-throughput assays to optimise patient diagnosis, the use of which will additionally create medical, ethical, legal, and regulatory pressures and these should be considered now, before they emerge (Cavalleri et al., 2011).

8.2.7 Impact of pharmacogenomics on drug development

With the advancements in disease genomics, a rational approach to new and better therapies has become a realistic prospect. In terms of new or emerging drugs, PGx can be applied to drug design in several ways. Firstly PGx research can be used to ‘rescue’ any existing drugs that have been withdrawn from the market, most likely due to serious or common AEs in a number of people. Retrospective genotyping of such clinical trial participants could identify the genetic make-up of the often small proportion of patients who suffered these, as to prevent their use in these genetic groups in the future. Subjects’ eligibility to participate in additional clinical trials will be decided by the results of such PGx tests. A subgroup of individuals may also for genetic reasons, in contrast respond well to a drug without side effects. Individuals from both response phenotype groups, will thus benefit from a drug being placed on the market with the provision that specific genetic tests will be administered prior to drug prescription.

In terms of drug research and development, the only way forward would be to couple

new drug trials with PGx studies. Pharmacogenomics approaches would focus on the identification of genetically determined drug targets involved in disease and/ or genetic polymorphisms associated with treatment response. Such research could assist pharmaceutical companies to develop more effective drugs with fewer side effects. In pharmacotherapy for a particular disease, numerous polymorphisms may influence drug metabolism or disease development. These polymorphisms must be identified before PGx and pharmacogenomic products can be developed. The complexity of both the human genome, and human diseases, however will make it difficult and time-consuming to produce this information.

8.2.8 Next generation sequencing and platforms for data analysis

NGS is the next stage in the genetics of complex traits and also likely to impact drug response (di Iulio and Rotger, 2012). This will help unravel the complexity of the human genome in terms of genetic variations that are yet to be discovered and the biological mechanisms that surrounds these variants. The impact of NGS technologies on genomics is expected to be far reaching and will change the field of disease genetics including that which influences disease treatment for years to come (Zhang et al., 2011). Future PGx studies are consequently likely to focus on exome genotyping in search of novel genetic markers associated with AED sensitivity. Whole genome sequencing in patients will allow the detection of underlying rare mutations that in combination with the data for common genomic variation is more appropriate for studying the multigenic nature of treatment responsiveness in epilepsy. This could hopefully identify unresponsive individuals from responsive individuals and also improve the clinical management of patients who require unusually high or low drug doses to control their seizures. NGS technologies will have a striking impact on genomic research and the entire biological field. One problem with next generation sequencing projects is the handling of massive amounts of sequencing data that must be organized, cleaned up, assembled, and analyzed. Sequencing of an entire genome can generate millions of pieces of sequence that must be assembled. Easy to use computing programs are thus desperately needed to make data interpretation manageable and fast. A variety of software tools are under development and many are available online for NGS data analysis. Their functions fit into several general categories each of which poses a challenge to be met for efficient analysis of NGS data:

i) Software packages or applications for the alignment of NGS reads to a reference sequence; The most important step in NGS data analysis is the successful alignment or assembly of short reads to a reference genome and this critical step is further challenged by the emergence of new NGS short-read technology.

ii) Packages for genome annotation and functional prediction of mutations; After the successful alignment and assembly of NGS data, the next challenge in NGS data analysis is the interpretation of data- A large number of presumed 'novel' genetic variants are present by chance in any single human genome and this makes it difficult to identify which of the numerous characterised variants are actually causal

iii) End-user software packages and cloud computing; The former provides a user-friendly interface, easy to use data input and output formats, and integrates multiple computing programs into one software package. It is difficult for many research laboratories to successfully conduct NGS projects due to the high level of information technology support required. A possible solution is cloud computing. In cloud computing, a user can use a virtual operating system (or "cloud") to process data on a computer cluster for high parallel tasks (allows scientists to rent both storage and processing power virtually by accessing servers as they are needed).

8.2.9 Future work in epilepsy pharmacogenetics

Many of the research studies for epilepsy PGx have focused on patients resistant to antiepileptic drug therapy. Classification of this response phenotype requires treatment failure with multiple agents thus is often concerned with long-term or chronic epilepsy. In this thesis we investigated responsiveness to drug therapy in newly treated epilepsy patients. The benefit of scanning the entire genome for new susceptibility loci for AED efficacy has become evident and subsequently more robust genetic influences on drug response in epilepsy are anticipated (Crowley et al., 2009, Loscher et al., 2009, Daly, 2010a). The GWA approach does offer an opportunity to locate associations not previously considered in epilepsy candidate gene studies, whilst also providing the required statistical power to detect multiple modest genetic effects that are assumed to determine AED response (Ferraro et al., 2012). An independent GWAS study for studying the influence of common genetic variation.

Bibliography

Abbott NJ, Khan EU, Rollinson CM, Reichel A, Janigro D, Dombrowski SM, Dobbie MS, Begley DJ (2002) Drug resistance in epilepsy: the role of the blood-brain barrier. Novartis Found Symp 243:38-47; discussion 47-53, 180-185.

Abe T, Seo T, Ishitsu T, Nakagawa T, Hori M, Nakagawa K (2008) Association between SCN1A polymorphism and carbamazepine-resistant epilepsy. Br J Clin Pharmacol 66:304-307.

Ad N, Lee P, Cox JL (2002) Type A aortic dissection with associated anomaly of the carotid and vertebral vessels. J Thorac Cardiovasc Surg 123:570-571.

Ahn H, Moon H, Fazzari MJ, Lim N, Chen JJ, Kodell RL (2007) Classification by ensembles from random partitions of high-dimensional data. Computational Statistics & Data Analysis 51:6166-6179.

Alfirevic A, Jorgensen AL, Williamson PR, Chadwick DW, Park BK, Pirmohamed M (2006) HLA-B locus in Caucasian patients with carbamazepine hypersensitivity. Pharmacogenomics 7:813-818.

Allabi AC, Gala JL, Horsmans Y (2005) CYP2C9, CYP2C19, ABCB1 (MDR1) genetic polymorphisms and phenytoin metabolism in a Black Beninese population. Pharmacogenet Genomics 15:779-786.

Altmann A, Beerenwinkel N, Sing T, Savenkov I, Doumer M, Kaiser R, Rhee SY, Fessel WJ, Shafer RW, Lengauer T (2007) Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. Antivir Ther 12:169-178.

Anderson GD (2008) Pharmacokinetic, pharmacodynamic, and pharmacogenetic targeted therapy of antiepileptic drugs. Ther Drug Monit 30:173-180.

Annegers JF, Hauser WA, Elveback LR (1979) Remission of seizures and relapse in patients with epilepsy. Epilepsia 20:729-737.

Aronica E, Yankaya B, Troost D, van Vliet EA, Lopes da Silva FH, Gorter JA (2001) Induction of neonatal sodium channel II and III alpha-isoform mRNAs in neurons and microglia after status epilepticus in the rat hippocampus. Eur J Neurosci 13:1261-1266.

Arteaga CL, Sliwkowski MX, Osborne CK, Perez EA, Puglisi F, Gianni L (2012) Treatment of HER2-positive breast cancer: current status and future perspectives. Nat Rev Clin Oncol 9:16-32.

Awasthi S, Hallene KL, Fazio V, Singhal SS, Cucullo L, Awasthi YC, Dini G, Janigro D (2005) RLIP76, a non-ABC transporter, and drug resistance in epilepsy. BMC Neurosci 6.

Aynacioglu AS, Brockmoller J, Bauer S, Sachse C, Guzelbey P, Ongen Z, Nacak M, Roots I (1999) Frequency of cytochrome P450 CYP2C9 variants in a Turkish population and functional relevance for phenytoin. Br J Clin Pharmacol 48:409-415.

Baksh MF, Kelly PJ (2007) Statistical methods for examining genetic influences of resistance to anti-epileptic drugs. Expert Review of Clinical Pharmacology 1:137-144.

Ban HJ, Heo JY, Oh KS, Park KJ (2010) Identification of type 2 diabetes-associated combination of SNPs using support vector machine. BMC genetics 11:26.

Barcs G, Walker EB, Elger CE, Scaramelli A, Stefan H, Sturm Y, Moore A, Flesch G, Kramer L, D'Souza J (2000) Oxcarbazepine placebo-controlled, dose-ranging trial in refractory partial epilepsy. Epilepsia 41:1597-1607.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS, Type 1 Diabetes Genetics C (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41:703-707.

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265.

Battino D, Croci D, Rossini A, Messina S, Mamoli D, Perucca E (2003) Serum carbamazepine concentrations in elderly patients: a case-matched pharmacokinetic evaluation based on therapeutic drug monitoring data. Epilepsia 44:923-929.

Baulac S, Huberfeld G, Gourfinkel-An I, Mitropoulou G, Beranger A, Prud'homme JF, Baulac M, Brice A, Bruzzone R, LeGuern E (2001) First genetic evidence of GABA(A) receptor dysfunction in epilepsy: a mutation in the gamma2-subunit gene. Nature genetics 28:46-48.

Beck H, Elger CE (2008) Epilepsy research: a window onto function to and dysfunction of the human brain. Dialogues Clin Neurosci 10:7-15.

Beckmann JS, Estivill X, Antonarakis SE (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 8:639-646.

Beghi E, Perucca E (1995) The management of epilepsy in the 1990s. Acquisitions, uncertainties and priorities for future research. Drugs 49:680-694.

Benarroch EE (2007) GABAA receptor heterogeneity, function, and implications for epilepsy. Neurology 68:612-614.

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125:279-284.

Bentley DR (2000) The Human Genome Project--an overview. Med Res Rev 20:189-196.

Berg AT, Berkovic SF, Brodie MJ, Buchhalter J, Cross JH, van Emde Boas W, Engel J, French J, Glauser TA, Mathern GW, Moshe SL, Nordli D, Plouin P, Scheffer IE (2010) Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005-2009. Epilepsia 51:676-685.

Berg AT, Kelly MM (2006) Defining intractability: comparisons among published definitions. Epilepsia 47:431-436.

Berg AT, Levy SR, Novotny EJ, Shinnar S (1996) Predictors of intractable epilepsy in childhood: a case-control study. Epilepsia 37:24-30.

Berg AT, Shinnar S, Levy SR, Testa FM, Smith-Rapaport S, Beckerman B (2001) Early development of intractable epilepsy in children: a prospective study. Neurology 56:1445-1452.

Bergey GK (2005) Evidence-based Treatment of Idiopathic Generalized Epilepsies with New Antiepileptic Drugs. Epilepsia 46:161-168.

Berkovic SF, Mulley JC, Scheffer IE, Petrou S (2006) Human epilepsies: interaction of genetic and acquired factors. Trends in neurosciences 29:391-397.

Bethmann K, Fritschy JM, Brandt C, Loscher W (2008) Antiepileptic drug resistant rats differ from drug responsive rats in GABA A receptor subunit expression in a model of temporal lobe epilepsy. Neurobiol Dis 31:169-187.

Beydoun A, D'Souza J (2012) Treatment of idiopathic generalized epilepsy - a review of the evidence. Expert opinion on pharmacotherapy 13:1283-1298.

Bhalla D, Godet B, Druet-Cabanac M, Preux PM (2011) Etiologies of epilepsy: a comprehensive review. Expert Rev Neurother 11:861-876.

Bhaskar H, Hoyle DC, Singh S (2006a) Machine learning in bioinformatics: a brief survey and recommendations for practitioners. Comput Biol Med 36:1104-1125.

Bhaskar H, Hoyle DC, Singh S (2006b) Machine learning in bioinformatics: a brief survey and recommendations for practitioners. Computers in biology and medicine 36:1104-1125.

Bhathena A, Spear BB (2008) Pharmacogenetics: improving drug and dose selection. Curr Opin Pharmacol 8:639-646.

Bialer M, Johannessen SI, Kupferberg HJ, Levy RH, Perucca E, Tomson T (2007) Progress report on new antiepileptic drugs: a summary of the Eigth Eilat Conference (EILAT VIII). Epilepsy research 73:1-52.

Bialer M, White HS (2010) Key factors in the discovery and development of new antiepileptic drugs. Nat Rev Drug Discov 9:68-82.

Bianchi MT, Song L, Zhang H, Macdonald RL (2002) Two different mechanisms of disinhibition produced by GABAA receptor mutations linked to epilepsy in humans. J Neurosci 22:5321-5327.

Binder SR, Genovese MC, Merrill JT, Morris RI, Metzger AL (2005) Computer-assisted pattern recognition of autoantibody results. Clin Diagn Lab Immunol 12:1353-1357.

Birbeck GL, Hays RD, Cui X, Vickrey BG (2002) Seizure reduction and quality of life improvements in people with epilepsy. Epilepsia 43:535-538.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J,

Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Xu M, Haidar JN, Yu Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799-816.

Blanca Sanchez M, Herranz JL, Leno C, Arteaga R, Oterino A, Valdizan EM, Nicolas JM, Adin J, Shushtarian M, Armijo JA (2010a) UGT2B7_-161C>T polymorphism is associated with lamotrigine concentration-to-dose ratio in a multivariate study. Ther Drug Monit 32:177-184.

Board PG, Coggan M, Chelvanayagam G, Easteal S, Jermiin LS, Schulte GK, Danley DE, Hoth LR, Griffor MC, Kamath AV, Rosner MH, Chrunyk BA, Perregaux DE, Gabel CA, Geoghegan KF, Pandit J (2000) Identification, characterization, and crystal structure of the Omega class glutathione transferases. J Biol Chem 275:24798-24806.

Bonnett L, Smith CT, Smith D, Williamson P, Chadwick D, Marson AG (2012) Prognostic factors for time to treatment failure and time to 12 months of remission for patients with focal epilepsy: post-hoc, subgroup analyses of data from the SANAD trial. Lancet neurology 11:331-340.

Borst P, Evers R, Kool M, Wijnholds J (2000) A family of drug transporters: the multidrug resistance-associated proteins. J Natl Cancer Inst 92:1295-1302.

Bourdet SV, Gidal BE, Alldredge BK (2001) Pharmacologic management of epilepsy in the elderly. J Am Pharm Assoc (Wash) 41:421-436.
Bournissen FG, Moretti ME, Juurlink DN, Koren G, Walker M, Finkelstein Y (2009) Polymorphism of the MDR1/ABCB1 C3435T drug-transporter and resistance to anticonvulsant drugs: a meta-analysis. Epilepsia 50:898-903.

Brodie MJ (2005a) Diagnosing and predicting refractory epilepsy. Acta Neurol Scand Suppl 181:36-39.

Brodie MJ (2005b) Response to antiepileptic drug therapy: winners and losers. Epilepsia 46 Suppl 10:31-32.

Brodie MJ, Barry SJ, Bamagous GA, Norrie JD, Kwan P (2012) Patterns of treatment response in newly diagnosed epilepsy. Neurology 78:1548-1554.

Brodie MJ, Dichter MA (1997) Established antiepileptic drugs. Seizure 6:159-174.

Brodie MJ, French JA (2000) Management of epilepsy in adolescents and adults. Lancet 356:323-329.

Brodie MJ, Kwan P (2001) The star systems: overview and use in determining antiepileptic drug choice. CNS drugs 15:1-12; discussion 13-15.

Brodie MJ, Kwan P (2002) Staged approach to epilepsy management. Neurology 58:S2-8.

Brodie MJ, Sills GJ (2011) Combining antiepileptic drugs--rational polytherapy? Seizure 20:369-375.

Brookes AJ (1999) The essence of SNPs. Gene 234:177-186.

Browne TR (1998) Pharmacokinetics of antiepileptic drugs. Neurology 51:S2-7.

Browne TR, Holmes GL (2001) Epilepsy. New England Journal of Medicine 344:1145-1151.

Bu HZ, Kang P, Deese AJ, Zhao P, Pool WF (2005) Human in vitro glutathionyl and protein adducts of carbamazepine-10,11-epoxide, a stable and pharmacologically active metabolite of carbamazepine. Drug Metab Dispos 33:1920-1924.

Bu HZ, Zhao P, Dalvie DK, Pool WF (2007) Identification of primary and sequential bioactivation pathways of carbamazepine in human liver microsomes using liquid chromatography/tandem mass spectrometry. Rapid Commun Mass Spectrom 21:3317-3322.

Burchell B (2003) Genetic variation of human UDP-glucuronosyltransferase: implications in disease and drug glucuronidation. Am J Pharmacogenomics 3:37-52.

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P (2005) Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 28:171-182.

Callaghan B, Schlesinger M, Rodemer W, Pollard J, Hesdorffer D, Allen Hauser W, French J (2011) Remission and relapse in a drug-resistant epilepsy population followed prospectively. Epilepsia 52:619-626.

Callaghan BC, Anand K, Hesdorffer D, Hauser WA, French JA (2007) Likelihood of seizure remission in an adult population with refractory epilepsy. Ann Neurol 62:382-389.
Camfield C, Camfield P, Gordon K, Dooley J (1996) Does the number of seizures before treatment influence ease of control or remission of childhood epilepsy? Not if the number is 10 or less. Neurology 46:41-44.

Carder PJ, Hume R, Fryer AA, Strange RC, Lauder J, Bell JE (1990) Glutathione S-transferase in human brain. Neuropathol Appl Neurobiol 16:293-303.

Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91-99.

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446-452.

Castaldi PJ, Cho MH, Cohn M, Langerman F, Moran S, Tarragona N, Moukhachen H, Venugopal R, Hasimja D, Kao E, Wallace B, Hersh CP, Bagade S, Bertram L, Silverman EK, Trikalinos TA (2010) The COPD genetic association compendium: a comprehensive online database of COPD genetic associations. Hum Mol Genet 19:526-534.

Catterall WA (1992) Cellular and molecular biology of voltage-gated sodium channels. Physiol Rev 72:S15-48.

Catterall WA (1999) Molecular properties of brain sodium channels: an important target for anticonvulsant drugs. Adv Neurol 79:441-456.

Catterall WA (2000) From ionic currents to molecular mechanisms: the structure and function of voltage-gated sodium channels. Neuron 26:13-25.

Catterall WA, Perez-Reyes E, Snutch TP, Striessnig J (2005) International Union of Pharmacology. XLVIII. Nomenclature and structure-function relationships of voltage-gated calcium channels. Pharmacological reviews 57:411-425.

Cavalleri GL, McCormack M, Alhusaini S, Chaila E, Delanty N (2011) Pharmacogenomics and epilepsy: the road ahead. Pharmacogenomics 12:1429-1447.

Cavalleri GL, Weale ME, Shianna KV, Singh R, Lynch JM, Grinton B, Szoeke C, Murphy K, Kinirons P, O'Rourke D, Ge D, Depondt C, Claeys KG, Pandolfo M, Gumbs C, Walley N, McNamara J, Mulley JC, Linney KN, Sheffield LJ, Radtke RA, Tate SK, Chissoe SL, Gibson RA, Hosford D, Stanton A, Graves TD, Hanna MG, Eriksson K, Kantanen AM, Kalviainen R, O'Brien TJ, Sander JW, Duncan JS, Scheffer IE, Berkovic SF, Wood NW, Doherty CP, Delanty N, Sisodiya SM, Goldstein DB (2007) Multicentre search for genetic susceptibility loci in sporadic epilepsy syndrome and seizure types: a case-control study. Lancet Neurol 6:970-980.

Chapman AG (1998) Glutamate receptors in epilepsy. Prog Brain Res 116:371-383.

Chapman AG (2000) Glutamate and epilepsy. J Nutr 130:1043S-1045S.

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18-31.

Chaudhary UJ, Duncan JS, Lemieux L (2011) A dialogue with historical concepts of epilepsy from the Babylonians to Hughlings Jackson: persistent beliefs. Epilepsy Behav 21:109-114.
Chaudhry AS, Urban TJ, Lamba JK, Birnbaum AK, Remmel RP, Subramanian M, Strom S, You JH, Kasperaviciute D, Catarino CB, Radtke RA, Sisodiya SM, Goldstein DB, Schuetz EG (2009) CYP2C9*1B promoter polymorphisms, in linkage with CYP2C19*2, affect phenytoin auto-induction of clearance and maintenance dose. J Pharmacol Exp Ther.

Chen P, Lin JJ, Lu CS, Ong CT, Hsieh PF, Yang CC, Tai CT, Wu SL, Lu CH, Hsu YC, Yu HY, Ro LS, Lu CT, Chu CC, Tsai JJ, Su YH, Lan SH, Sung SF, Lin SY, Chuang HP, Huang LC, Chen YJ, Tsai PJ, Liao HT, Lin YH, Chen CH, Chung WH, Hung SI, Wu JY, Chang CF, Chen L, Chen YT, Shen CY (2011) Carbamazepine-induced toxic effects and HLA-B*1502 screening in Taiwan. N Engl J Med 364:1126-1133.

Chen Y, Lu J, Pan H, Zhang Y, Wu H, Xu K, Liu X, Jiang Y, Bao X, Yao Z, Ding K, Lo WH, Qiang B, Chan P, Shen Y, Wu X (2003) Association between genetic variation of CACNA1H and childhood absence epilepsy. Ann Neurol 54:239-243.
Choi H, Heiman G, Pandis D, Cantero J, Resor SR, Gilliam FG, Hauser WA (2008) Seizure remission and relapse in adults with intractable epilepsy: a cohort study. Epilepsia 49:1440-1445.

Chung JY, Cho JY, Yu KS, Kim JR, Lim KS, Sohn DR, Shin SG, Jang IJ (2008) Pharmacokinetic and pharmacodynamic interaction of lorazepam and valproic acid in relation to UGT2B7 genetic polymorphism in healthy subjects. Clin Pharmacol Ther 83:595-600.
Clare JJ, Tate SN, Nobbs M, Romanos MA (2000) Voltage-gated sodium channels as therapeutic targets. Drug Discov Today 5:506-520.

Clarke W, McMillin G (2006) Application of TDM, pharmacogenomics and biomarkers for neurological disease pharmacotherapy: focus on antiepileptic drugs. Personalized Medicine 3:139-149.

Cockerell OC, Johnson AL, Sander JW, Hart YM, Shorvon SD (1995) Remission of epilepsy: results from the National General Practice Study of Epilepsy. Lancet 346:140-144.

Cockerell OC, Johnson AL, Sander JW, Shorvon SD (1997) Prognosis of epilepsy: a review and further analysis of the first nine years of the British National General Practice Study of Epilepsy, a prospective population-based study. Epilepsia 38:31-46.

Colhoun HM, McKeigue PM, Davey Smith G (2003) Problems of reporting genetic associations with complex outcomes. Lancet 361:865-872.

Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580-1581.

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res 34:W621-625.

Copley RR (2004) Evolutionary convergence of alternative splicing in ion channels. Trends Genet 20:171-176.

Cosgun E, Limdi NA, Duarte CW (2011) High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. Bioinformatics 27:1384-1389.

Cover T, Hart P (1967) Nearest neighbor pattern classification. Information Theory, IEEE Transactions on 13:21-27.

Cox AG (2010) Pharmacogenomics and drug transport/efflux. In: Concepts in pharmacogenomics(Zdanowicz, M. M., ed) Bethesda, MD: American Society of Health-System Pharmacists.

Cramer JA (1994) Quality of life for people with epilepsy. Neurol Clin 12:1-13.

Cramer JA, Ben Menachem E, French J (2001) Review of treatment options for refractory epilepsy: new medications and vagal nerve stimulation. Epilepsy research 47:17-25.

Cramer JA, Fisher R, Ben-Menachem E, French J, Mattson RH (1999) New antiepileptic drugs: comparison of key clinical trials. Epilepsia 40:590-600.

Crews KR, Gaedigk A, Dunnenberger HM, Klein TE, Shen DD, Callaghan JT, Kharasch ED, Skaar TC (2012) Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for codeine therapy in the context of cytochrome P450 2D6 (CYP2D6) genotype. Clin Pharmacol Ther 91:321-326.

Crimins F, Dimitri R, Klein T, Palmer N, Cowen L (2005) Higher Dimensional Approach for Classification of Lung Cancer Microarray Data. In: Methods of Microarray Data Analysis(Shoemaker, J. and Lin, S., eds), pp 191-205: Springer US.

Crowley JJ, Sullivan PF, McLeod HL (2009) Pharmacogenomic genome-wide association studies: lessons learned thus far. Pharmacogenomics 10:161-163.

Crunelli V, Leresche N (2002) Childhood absence epilepsy: Genes, channels, neurons and networks. Nat Rev Neurosci 3:371-382.

Czarnowski I, Jędrzejowicz P (2008) Data Reduction Algorithm for Machine Learning and Data Mining. In: New Frontiers in Applied Artificial Intelligence, vol. 5027 (Nguyen, N. et al., eds), pp 276-285: Springer Berlin Heidelberg.

Daly AK (2003) Pharmacogenetics of the major polymorphic metabolizing enzymes. Fundamental & clinical pharmacology 17:27-41.

Daly AK (2010a) Genome-wide association studies in pharmacogenomics. Nat Rev Genet 11:241-246.

Daly AK (2010b) Pharmacogenetics and human genetic polymorphisms. Biochem J 429:435-449.

Daly AK, Day CP (2001) Candidate gene case-control association studies: advantages and potential pitfalls. Br J Clin Pharmacol 52:489-499.

Davey J, Turner RM, Clarke MJ, Higgins JP (2011) Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC medical research methodology 11:160.

Davila S, Wright VJ, Khor CC, Sim KS, Binder A, Breunis WB, Inwald D, Nadel S, Betts H, Carrol ED, de Groot R, Hermans PW, Hazelzet J, Emonts M, Lim CC, Kuijpers TW, Martinon-Torres F, Salas A, Zenz W, Levin M, Hibberd ML (2010) Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nat Genet 42:772-776.

de Bakker PIW (2009) Selection and Evaluation of Tag-SNPs Using Tagger and HapMap. Cold Spring Harbor protocols 2009:pdb.ip67.

Delgado-Escueta AV (2007) Advances in genetics of juvenile myoclonic epilepsies. Epilepsy Curr 7:61-67.

Depondt C (2006a) Pharmacogenetics in neuropsychiatric diseases: epilepsy as a model. Acta Neurol Belg 106:157-167.

Depondt C (2006b) The potential of pharmacogenetics in the treatment of epilepsy. European journal of paediatric neurology : EJPN : official journal of the European Paediatric Neurology Society 10:57-65.

Depondt C, Godard P, Espel RS, Da Cruz AL, Lienard P, Pandolfo M (2011) A candidate gene study of antiepileptic drug tolerability and efficacy identifies an association of CYP2C9 variants with phenytoin toxicity. Eur J Neurol 18:1159-1164.
Depondt C, Shorvon SD (2006) Genetic association studies in epilepsy pharmacogenomics: lessons learnt and potential applications. Pharmacogenomics 7:731-745.

Derrac J, Triguero I, Garcia S, Herrera F (2012) Integrating Instance Selection, Instance Weighting, and Feature Weighting for Nearest Neighbor Classifiers by Coevolutionary Algorithms. IEEE Trans Syst Man Cybern B Cybern.

Desai AA, Innocenti F, Ratain MJ (2003) UGT pharmacogenomics: implications for cancer risk and cancer therapeutics. Pharmacogenetics 13:517-523.

Desmots F, Rissel M, Loyer P, Turlin B, Guillouzo A (2001) Immunohistological analysis of glutathione transferase A4 distribution in several human tissues using a specific polyclonal antibody. J Histochem Cytochem 49:1573-1580.

di Iulio J, Rotger M (2012) Pharmacogenomics: what is next? Front Pharmacol 2:86.

Dibbens LM, Heron SE, Mulley JC (2007) A polygenic heterogeneity model for common epilepsies with complex genetics. Genes Brain Behav 6:593-597.

Dichter MA, Ayala GF (1987) Cellular mechanisms of epilepsy: a status report. Science 237:157-164.

Dinu V, Zhao H, Miller PL (2007) Integrating domain knowledge with statistical and data mining methods for high-density genomic SNP disease association analysis. Journal of biomedical informatics 40:750-760.

Dlugos DJ, Buono RJ, Ferraro TN (2006) Defining the clinical role of pharmacogenetics in antiepileptic drug therapy. The pharmacogenomics journal 6:357-359.

Dlugos DJ, Sammel MD, Strom BL, Farrar JT (2001) Response to first drug trial predicts outcome in childhood temporal lobe epilepsy. Neurology 57:2259-2264.

Drazen JM, Yandava CN, Dube L, Szczerback N, Hippensteel R, Pillari A, Israel E, Schork N, Silverman ES, Katz DA, Drajesk J (1999) Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. Nature genetics 22:168-170.

Duncan JS, Sander JW, Sisodiya SM, Walker MC (2006) Adult epilepsy. Lancet 367:1087-1100.

Eadie MJ (1998) Therapeutic drug monitoring--antiepileptic drugs. Br J Clin Pharmacol 46:185-193.

Ebid AH, Ahmed MM, Mohammed SA (2007) Therapeutic drug monitoring and clinical outcomes in epileptic Egyptian patients: a gene polymorphism perspective study. Ther Drug Monit 29:305-312.

Egger M, Smith GD (1997) Meta-analysis: Potentials and promise. Bmj 315:1371-1374.

Eichelbaum M, Ingelman-Sundberg M, Evans WE (2006) Pharmacogenomics and individualized drug therapy. Annu Rev Med 57:119-137.

Eichelbaum M, Tomson Tr, Tybring G, Bertilsson L (1985) Carbamazepine Metabolism in Man: Induction and Pharmacogenetic Aspects. Clinical Pharmacokinetics 10:80-90.
Elger CE (2003) Pharmacoresistance: modern concept and basic data derived from human brain tissue. Epilepsia 44 Suppl 5:9-15.
Emmert-Streib F, Dehmer M (2010) Medical biostatistics for complex diseases. Weinheim: Wiley-Blackwell.

Engel J, Jr. (2006a) ILAE classification of epilepsy syndromes. Epilepsy research 70 Suppl 1:S5-10.

Engel J, Jr. (2006b) Report of the ILAE classification core group. Epilepsia 47:1558-1568.

Engel J, Pedley TA (2008) Epilepsy : a comprehensive textbook. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Escayg A, MacDonald BT, Meisler MH, Baulac S, Huberfeld G, An-Gourfinkel I, Brice A, LeGuern E, Moulard B, Chaigne D, Buresi C, Malafosse A (2000) Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+2. Nat Genet 24:343-345.

Evans WE, Johnson JA (2001) Pharmacogenomics: the inherited basis for interindividual differences in drug response. Annu Rev Genomics Hum Genet 2:9-39.

Evans WE, McLeod HL (2003) Pharmacogenomics--drug disposition, drug targets, and side effects. N Engl J Med 348:538-549.

Evans WE, Relling MV (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. Science 286:487-491.

Evans WE, Relling MV (2004) Moving towards individualized medicine with pharmacogenomics. Nature 429:464-468.

Everitt B, Howell DC (2005) Encyclopedia of statistics in behavioral science. Chichester: John Wiley.

Fatovic-Ferencic S, Durrigl MA (2001) The sacred disease and its patron saint. Epilepsy & behavior : E&B 2:370-373.

Feero WG, Guttmacher AE, Collins FS (2010) Genomic medicine--an updated primer. N Engl J Med 362:2001-2011.

Ferraro TN, Buono RJ (2005) The relationship between the pharmacology of antiepileptic drugs and human gene variation: an overview. Epilepsy & behavior : E&B 7:18-36.

Ferraro TN, Buono RJ (2006) Polygenic epilepsy. Adv Neurol 97:389-398.

Ferraro TN, Dlugos DJ, Buono RJ (2006) Challenges and opportunities in the application of pharmacogenetics to antiepileptic drug therapy. Pharmacogenomics 7:89-103.

Ferraro TN, Dlugos DJ, Hakonarson H, Buono RJ (2012) Strategies for Studying the Epilepsy Genome. In: Jasper's Basic Mechanisms of the Epilepsies(Noebels, J. L. et al., eds) Bethesda

MD: Michael A Rogawski, Antonio V Delgado-Escueta, Jeffrey L Noebels, Massimo Avoli and Richard W Olsen.

Ferrell PB, Jr., McLeod HL (2008) Carbamazepine, HLA-B*1502 and risk of Stevens-Johnson syndrome and toxic epidermal necrolysis: US FDA recommendations. Pharmacogenomics 9:1543-1546.

Fisher RS, van Emde Boas W, Blume W, Elger C, Genton P, Lee P, Engel J, Jr. (2005) Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). Epilepsia 46:470-472.

Fix E, Hodges, JL (1951) Discriminatory analysis – non-parametric discrimination: Consistency properties. . Randolph Field, Texas: USAF School of Aviation Medicine.
Franciotta D, Kwan P, Perucca E (2009) Genetic basis for idiosyncratic reactions to antiepileptic drugs. Current opinion in neurology 22:144-149.

Frankel WN (2009) Genetics of complex neurological disease: challenges and opportunities for modeling epilepsy in mice and rats. Trends Genet 25:361-367.

French JA (2002) Response to Early AED Therapy and Its Prognostic Implications. Epilepsy Curr 2:69-71.

French JA (2007) Refractory epilepsy: clinical overview. Epilepsia 48 Suppl 1:3-7.

French JA, Kanner AM, Bautista J, Abou-Khalil B, Browne T, Harden CL, Theodore WH, Bazil C, Stern J, Schachter SC, Bergen D, Hirtz D, Montouris GD, Nespeca M, Gidal B, Marks WJ, Jr., Turk WR, Fischer JH, Bourgeois B, Wilner A, Faught RE, Jr., Sachdeo RC, Beydoun A, Glauser TA (2004) Efficacy and tolerability of the new antiepileptic drugs, II: Treatment of refractory epilepsy: report of the TTA and QSS Subcommittees of the American Academy of Neurology and the American Epilepsy Society. Epilepsia 45:410-423.

French JA, Kugler AR, Robbins JL, Knapp LE, Garofalo EA (2003) Dose-response trial of pregabalin adjunctive therapy in patients with partial seizures. Neurology 60:1631-1637.
Fromm MF (2000) P-glycoprotein: a defense mechanism limiting oral bioavailability and CNS accumulation of drugs. Int J Clin Pharmacol Ther 38:69-74.

Fukushima Y, Seo T, Hashimoto N, Higa Y, Ishitsu T, Nakagawa K (2008a) Glutathione-S-transferase (GST) M1 null genotype and combined GSTM1 and GSTT1 null genotypes are risk factors for increased serum gamma-glutamyltransferase in valproic acid-treated patients. Clin Chim Acta 389:98-102.

Fukushima Y, Seo T, Hashimoto N, Higa Y, Ishitsu T, Nakagawa K (2008b) Glutathione-S-transferase (GST) M1 null genotype and combined GSTM1 and GSTT1 null genotypes are risk factors for increased serum gamma-glutamyltransferase in valproic acid-treated patients. Clin Chim Acta 389:98-102.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16:906-914.

Gabriel S, Ziaugra L (2004) SNP genotyping using Sequenom MassARRAY 7K platform. Current protocols in human genetics / editorial board, Jonathan L Haines [et al] Chapter 2:Unit 2 12.

Gabriel S, Ziaugra L, Tabbaa D (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet Chapter 2:Unit 2 12.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Gardiner M (2005) Genetics of idiopathic generalized epilepsies. Epilepsia 46 Suppl 9:15-20.
Gardiner SJ, Begg EJ (2006) Pharmacogenetics, drug-metabolizing enzymes, and clinical practice. Pharmacol Rev 58:521-590.

Gastaldi M, Bartolomei F, Massacrier A, Planells R, Robaglia-Schlupp A, Cau P (1997) Increase in mRNAs encoding neonatal II and III sodium channel alpha-isoforms during kainate-induced seizures in adult rat hippocampus. Brain Res Mol Brain Res 44:179-190.
George AL, Jr. (2005) Inherited disorders of voltage-gated sodium channels. J Clin Invest 115:1990-1999.

Gillham RA, Williams N, Wiedmann KD, Butler E, Larkin JG, Brodie MJ (1990) Cognitive function in adult epileptic patients established on anticonvulsant monotherapy. Epilepsy research 7:219-225.

Glauser T, Ben-Menachem E, Bourgeois B, Cnaan A, Chadwick D, Guerreiro C, Kalviainen R, Mattson R, Perucca E, Tomson T (2006) ILAE treatment guidelines: evidence-based analysis of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes. Epilepsia 47:1094-1120.

Goldstein BA, Hubbard AE, Cutler A, Barcellos LF (2010) An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC genetics 11:49.

Goldstein DB (2005) The genetics of human drug response. Philos Trans R Soc Lond B Biol Sci 360:1571-1572.

Goldstein DB (2009) Common genetic variation and human traits. N Engl J Med 360:1696-1698.

Goldstein DB, Tate SK, Sisodiya SM (2003) Pharmacogenetics goes genomic. Nat Rev Genet 4:937-947.

Goldstein DB, Weale ME (2001) Population genomics: linkage disequilibrium holds the key. Curr Biol 11:R576-579.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-537.

Goto S, Seo T, Murata T, Nakada N, Ueda N, Ishitsu T, Nakagawa K (2007) Population estimation of the effects of cytochrome P450 2C9 and 2C19 polymorphisms on phenobarbital clearance in Japanese. Ther Drug Monit 29:118-121.
Grant SF, Hakonarson H (2007) Recent development in pharmacogenomics: from candidate genes to genome-wide association studies. Expert Rev Mol Diagn 7:371-393.

Graves TD (2006) Ion channels and epilepsy. Qjm 99:201-217.

Gray IC, Campbell DA, Spurr NK (2000) Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet 9:2403-2408.

Greenberg DA, Durner M, Delgado-Escueta AV (1992) Evidence for multiple gene loci in the expression of the common generalized epilepsies. Neurology 42:56-62.

Greenberg DA, Subaran R (2011) Blinders, phenotype, and fashionable genetic analysis: a critical examination of the current state of epilepsy genetic studies. Epilepsia 52:1-9.

Grover S, Gupta M, Kukreti R (2011) Challenges and recommendations for conducting epidemiological studies in the field of epilepsy pharmacogenetics. Indian J Hum Genet 17:S4-S11.

Guerrini R, Dravet C, Genton P, Belmonte A, Kaminska A, Dulac O (1998) Lamotrigine and seizure aggravation in severe myoclonic epilepsy. Epilepsia 39:508-512.

Guessous I, Gwinn M, Khoury MJ (2009) Genome-wide association studies in pharmacogenomics: untapped potential for translation. Genome Med 1:46.

Guillemette C (2003) Pharmacogenomics of human UDP-glucuronosyltransferase enzymes. The pharmacogenomics journal 3:136-158.

Guo Y, Baum LW, Sham PC, Wong V, Ng PW, Lui CH, Sin NC, Tsoi TH, Tang CS, Kwan JS, Yip BH, Xiao SM, Thomas GN, Lau YL, Yang W, Cherny SS, Kwan P (2012) Two-stage genome-wide association study identifies variants in CAMSAP1L1 as susceptibility loci for epilepsy in Chinese. Hum Mol Genet 21:1184-1189.

Gurwitz D, McLeod HL (2009) Genome-wide association studies: powerful tools for improving drug safety and efficacy. Pharmacogenomics 10:157-159.

Haerian BS, Lim KS, Tan CT, Raymond AA, Mohamed Z (2011) Association of ABCB1 gene polymorphisms and their haplotypes with response to antiepileptic drugs: a systematic review and meta-analysis. Pharmacogenomics 12:713-725.

Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19:376-382.

Hakami T, Todaro M, Petrovski S, Macgregor L, Velakoulis D, Tan M, Matkovic Z, Gorelik A, Liew D, Yerra R, O'Brien TJ (2012) Substitution Monotherapy With Levetiracetam vs Older Antiepileptic Drugs: A Randomized Comparative Trial. Arch Neurol 1-9.

Hardy J, Singleton A (2009) Genomewide association studies and human disease. N Engl J Med 360:1759-1768.

Harley IJ, Narod SA (2009) Single nucleotide polymorphisms - variation on a theme. Bjog 116:1556-1557.

Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. New York: Springer.
Hauser WA, Rich SS, Lee JR, Annegers JF, Anderson VE (1998) Risk of recurrent seizures after two unprovoked seizures. N Engl J Med 338:429-434.

Hayes JD, Flanagan JU, Jowsey IR (2005) Glutathione transferases. Annual review of pharmacology and toxicology 45:51-88.

Heinzen EL, Yoon W, Tate SK, Sen A, Wood NW, Sisodiya SM, Goldstein DB (2007) Nova2 interacts with a cis-acting polymorphism to influence the proportions of drug-responsive splice variants of SCN1A. American journal of human genetics 80:876-883.

Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, Franke A, Muhle H, de Kovel C, Baker C, von Spiczak S, Kron KL, Steinich I, Kleefuss-Lie AA, Leu C, Gaus V, Schmitz B, Klein KM, Reif PS, Rosenow F, Weber Y, Lerche H, Zimprich F, Urak L, Fuchs K, Feucht M, Genton P, Thomas P, Visscher F, de Haan GJ, Moller RS, Hjalgrim H, Luciano D, Wittig M, Nothnagel M, Elger CE, Nurnberg P, Romano C, Malafosse A, Koeleman BP, Lindhout D, Stephani U, Schreiber S, Eichler EE, Sander T (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet 41:160-162.

Heron SE, Khosravani H, Varela D, Bladen C, Williams TC, Newman MR, Scheffer IE, Berkovic SF, Mulley JC, Zamponi GW (2007) Extended spectrum of idiopathic generalized epilepsies associated with CACNA1H functional variants. Ann Neurol 62:560-568.

Heron SE, Phillips HA, Mulley JC, Mazarib A, Neufeld MY, Berkovic SF, Scheffer IE (2004) Genetic variation of CACNA1H in idiopathic generalized epilepsy. Ann Neurol 55:595-596.

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95-108.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45-61.

Hitiris N, Brodie MJ (2006) Modern antiepileptic drugs: guidelines and beyond. Current opinion in neurology 19:175-180.

Hitiris N, Mohanraj R, Norrie J, Sills GJ, Brodie MJ (2007) Predictors of pharmacoresistant epilepsy. Epilepsy research 75:192-196.

Hoppe C (2005) Bioinformatics: computers or clinicians for complex disease risk assessment? Eur J Hum Genet 13:893-894.

Huang W, Lin YS, McConn DJ, 2nd, Calamia JC, Totah RA, Isoherranen N, Glodowski M, Thummel KE (2004) Evidence of significant contribution from CYP3A5 to hepatic drug metabolism. Drug Metab Dispos 32:1434-1445.

Hung CC, Lin CJ, Chen CC, Chang CJ, Liou HH (2004) Dosage recommendation of phenytoin for patients with epilepsy with different CYP2C9/CYP2C19 polymorphisms. Ther Drug Monit 26:534-540.

Hunt R, Sauna ZE, Ambudkar SV, Gottesman MM, Kimchi-Sarfaty C (2008) Silent (Synonymous) SNPs: Should We Care About Them? , vol. 578, pp 23-39.

Hustert E, Haberl M, Burk O, Wolbold R, He YQ, Klein K, Nuessler AC, Neuhaus P, Klattig J, Eiselt R, Koch I, Zibat A, Brockmoller J, Halpert JR, Zanger UM, Wojnowski L (2001) The genetic determinants of the CYP3A5 polymorphism. Pharmacogenetics 11:773-779.
Ikediobi ON, Shin J, Nussbaum RL, Phillips KA, Walsh JM, Ladabaum U, Marshall D (2009) Addressing the challenges of the clinical application of pharmacogenetic testing. Clin Pharmacol Ther 86:28-31.

Inaba T, Nebert DW, Burchell B, Watkins PB, Goldstein JA, Bertilsson L, Tucker GT (1995) Pharmacogenetics in clinical pharmacology and toxicology. Can J Physiol Pharmacol 73:331-338.

Ingelman-Sundberg M (2004a) Human drug metabolising cytochrome P450 enzymes: properties and polymorphisms. Naunyn Schmiedebergs Arch Pharmacol 369:89-104.

Ingelman-Sundberg M (2004b) Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. Trends Pharmacol Sci 25:193-200.

Ingelman-Sundberg M, Oscarson M, McLellan RA (1999) Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. Trends Pharmacol Sci 20:342-349.

Ioannidis JP (2003) Genetic associations: false or true? Trends in molecular medicine 9:135-138.

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29:306-309.

Iyengar SK, Elston RC (2007) The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods Mol Biol 376:71-84.

Jang SY, Kim MK, Lee KR, Park MS, Kim BC, Cho KH, Lee MC, Kim YS (2009) Gene-to-gene interaction between sodium channel-related genes in determining the risk of antiepileptic drug resistance. J Korean Med Sci 24:62-68.

Janz D, Beck-Mannagetta G, Sander T (1992) Do idiopathic generalized epilepsies share a common susceptibility gene? Neurology 42:48-55.

Jazwinska EC (2001) Exploiting human genetic variation in drug discovery and development. Drug Discov Today 6:198-205.

Joaqu, n D, Salvador G, Francisco H (2010) IFS-CoCo: Instance and feature selection based on cooperative coevolution with nearest neighbor rule. Pattern Recogn 43:2082-2105.

Johannessen SI, Battino D, Berry DJ, Bialer M, Kramer G, Tomson T, Patsalos PN (2003) Therapeutic drug monitoring of the newer antiepileptic drugs. Ther Drug Monit 25:347-363.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nature genetics 29:233-237.

Johnson JA (2001) Drug target pharmacogenomics: an overview. Am J Pharmacogenomics 1:271-281.

Johnson JA (2003) Pharmacogenetics: potential for individualized drug therapy through genetics. Trends Genet 19:660-666.

Johnson JA, Evans WE (2002) Molecular diagnostics as a predictive tool: genetics of drug efficacy and toxicity. Trends in molecular medicine 8:300-305.

Johnson JA, Gong L, Whirl-Carrillo M, Gage BF, Scott SA, Stein CM, Anderson JL, Kimmel SE, Lee MT, Pirmohamed M, Wadelius M, Klein TE, Altman RB (2011a) Clinical

Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. Clin Pharmacol Ther 90:625-629.

Johnson MR, Tan NC, Kwan P, Brodie MJ (2011b) Newly diagnosed epilepsy and pharmacogenomics research: a step in the right direction? Epilepsy & behavior : E&B 22:3-8.
Johnston JA, Rees MI, Smith PE (2009) Epilepsy genetics: clinical beginnings and social consequences. Qjm 102:497-499.

Judson R, Stephens JC, Windemuth A (2000) The predictive power of haplotypes in clinical response. Pharmacogenomics 1:15-26.

Kaneko S, Yoshida S, Kanai K, Yasui-Furukori N, Iwasa H (2008) Development of individualized medicine for epilepsy based on genetic information. Expert Review of Clinical Pharmacology 1:661-681.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong MY, Karczewski KJ, Huber W, Weissman SM, Gerstein

MB, Korbel JO, Snyder M (2010) Variation in transcription factor binding among humans. Science 328:232-235.

Kasperaviciute D, Catarino CB, Heinzen EL, Depondt C, Cavalleri GL, Caboclo LO, Tate SK, Jamnadas-Khoda J, Chinthapalli K, Clayton LM, Shianna KV, Radtke RA, Mikati MA, Gallentine WB, Husain AM, Alhusaini S, Leppert D, Middleton LT, Gibson RA, Johnson MR, Matthews PM, Hosford D, Heuser K, Amos L, Ortega M, Zumsteg D, Wieser HG, Steinhoff BJ, Kramer G, Hansen J, Dorn T, Kantanen AM, Gjerstad L, Peuralinna T, Hernandez DG, Eriksson KJ, Kalviainen RK, Doherty CP, Wood NW, Pandolfo M, Duncan JS, Sander JW, Delanty N, Goldstein DB, Sisodiya SM (2010) Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study. Brain : a journal of neurology 133:2136-2147.

Kasperaviciute D, Sisodiya SM (2009) Epilepsy pharmacogenetics. Pharmacogenomics 10:817-836.

Kass RS (2005) The channelopathies: novel insights into molecular and genetic mechanisms of human disease. J Clin Invest 115:1986-1989.

Kearney JA (2012) Advances in epilepsy genetics and genomics. Epilepsy Curr 12:143-146.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12:996-1006.

Kerb R, Aynacioglu AS, Brockmoller J, Schlagenhaufer R, Bauer S, Szekeres T, Hamwi A, Fritzer-Szekeres M, Baumgartner C, Ongen HZ, Guzelbey P, Roots I, Brinkmann U (2001)

The predictive value of MDR1, CYP2C9, and CYP2C19 polymorphisms for phenytoin plasma levels. The pharmacogenomics journal 1:204-210.

Kerr BM, Thummel KE, Wurden CJ, Klein SM, Kroetz DL, Gonzalez FJ, Levy RH (1994) Human liver carbamazepine metabolism. Role of CYP3A4 and CYP2C8 in 10,11-epoxide formation. Biochem Pharmacol 47:1969-1979.

Ketter TA, Frye MA, Cora-Locatelli G, Kimbrell TA, Post RM (1999) Metabolism and excretion of mood stabilizers and new anticonvulsants. Cell Mol Neurobiol 19:511-532.

Khoury MJ, Bertram L, Boffetta P, Butterworth AS, Chanock SJ, Dolan SM, Fortier I, Garcia-Closas M, Gwinn M, Higgins JP, Janssens AC, Ostell J, Owen RP, Pagon RA, Rebbeck TR,

Rothman N, Bernstein JL, Burton PR, Campbell H, Chockalingam A, Furberg H, Little J, O'Brien TR, Seminara D, Vineis P, Winn DM, Yu W, Ioannidis JP (2009) Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. Am J Epidemiol 170:269-279.

Kim DU, Kim MK, Cho YW, Kim YS, Kim WJ, Lee MG, Kim SE, Nam TS, Cho KH, Kim YO, Lee MC (2011a) Association of a synonymous GAT3 polymorphism with antiepileptic drug pharmacoresistance. J Hum Genet 56:640-646.

Kim DW, Lee SK, Chu K, Jang IJ, Yu KS, Cho JY, Kim SJ (2009) Lack of association between ABCB1, ABCG2, and ABCC2 genetic polymorphisms and multidrug resistance in partial epilepsy. Epilepsy research 84:86-90.

Kim KY, Kim BJ, Yi GS (2004) Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics 5:160.

Kim MK, Moore JH, Kim JK, Cho KH, Cho YW, Kim YS, Lee MC, Kim YO, Shin MH (2011b) Evidence for epistatic interactions in antiepileptic drug resistance. J Hum Genet 56:71-76.

Kim WJ, Lee JH, Yi J, Cho YJ, Heo K, Lee SH, Kim SW, Kim MK, Kim KH, In Lee B, Lee MG (2010) A nonsynonymous variation in MRP2/ABCC2 is associated with neurological adverse drug reactions of carbamazepine in patients with epilepsy. Pharmacogenet Genomics 20:249-256.

Kirchheiner J, Brockmoller J (2005) Clinical Consequences of Cytochrome P450 2C9 Polymorphisms. Clin Pharmacol Ther 77:1-16.

Kirchheiner J, Seeringer A (2007) Clinical implications of pharmacogenetics of cytochrome P450 drug metabolizing enzymes. Biochimica et biophysica acta 1770:489-494.

Kitteringham NR, Pirmohamed M, Park BK (1998) 3 The pharmacology of the cytochrome P450 enzyme system. Bailli√®re's Clinical Anaesthesiology 12:191-211.

Klotz U (2007) The role of pharmacogenetics in the metabolism of antiepileptic drugs - Pharmacokinetic and therapeutic implications. Clinical Pharmacokinetics 46:271-279.

Koster ES, Rodin AS, Raaijmakers JA, Maitland-van der Zee AH (2009) Systems biology in pharmacogenomic research: the way to personalized prescribing? Pharmacogenomics 10:971-981.

Kotsiantis SB (2007) Supervised Machine Learning: A Review of Classification Techniques. In: Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, pp 3-24: IOS Press.

Krauss GL, Serratosa JM, Villanueva V, Endziniene M, Hong Z, French J, Yang H, Squillacote D, Edwards HB, Zhu J, Laurenza A (2012) Randomized phase III study 306: adjunctive perampanel for refractory partial-onset seizures. Neurology 78:1408-1415.

Krikova EV, Val'dman EA, Avakian GN, Andreev Ia A, Denisov EV, Rider FK, Biktimerov RR, Chukanova AS, Burd SG (2009) [Association study of the SCN1 gene polymorphism and effective dose of lamotrigine]. Zh Nevrol Psikhiatr Im S S Korsakova 109:57-62.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature genetics 22:139-144.

Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, Schuetz J, Watkins PB, Daly A, Wrighton SA, Hall SD, Maurel P, Relling M, Brimer C, Yasuda K, Venkataramanan R, Strom S, Thummel

K, Boguski MS, Schuetz E (2001) Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. Nature genetics 27:383-391.

Kumari R, Lakhan R, Garg RK, Kalita J, Misra UK, Mittal B (2011) Pharmacogenomic association study on the role of drug metabolizing, drug transporters and drug target gene polymorphisms in drug-resistant epilepsy in a north Indian population. Indian journal of human genetics 17 Suppl 1:S32-40.

Kumari R, Lakhan R, Kalita J, Misra UK, Mittal B (2010) Association of alpha subunit of GABAA receptor subtype gene polymorphisms with epilepsy susceptibility and drug resistance in north Indian population. Seizure 19:237-241.

Kuo CC (1998) A common anticonvulsant binding site for phenytoin, carbamazepine, and lamotrigine in neuronal Na+ channels. Mol Pharmacol 54:712-721.

Kupferberg H (2001) Animal models used in the screening of antiepileptic drugs. Epilepsia 42 Suppl 4:7-12.

Kutt H MF (1968) Management of epilepsy with diphenylhydantoin sodium: Dosage regulation for problem patients. JAMA: The Journal of the American Medical Association 203:969-972.

Kwan P, Arzimanoglou A, Berg AT, Brodie MJ, Allen Hauser W, Mathern G, Moshe SL, Perucca E, Wiebe S, French J (2010) Definition of drug resistant epilepsy: consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies. Epilepsia 51:1069-1077.

Kwan P, Brodie MJ (2000a) Early identification of refractory epilepsy. N Engl J Med 342:314-319.

Kwan P, Brodie MJ (2000b) Epilepsy after the first drug fails: substitution or add-on? Seizure 9:464-468.

Kwan P, Brodie MJ (2001a) Effectiveness of first antiepileptic drug. Epilepsia 42:1255-1260.

Kwan P, Brodie MJ (2001b) Neuropsychological effects of epilepsy and antiepileptic drugs. Lancet 357:216-222.

Kwan P, Brodie MJ (2002) Refractory epilepsy: a progressive, intractable but preventable condition? Seizure 11:77-84.

Kwan P, Brodie MJ (2004) Drug treatment of epilepsy: when does it fail and how to optimize its use? CNS Spectr 9:110-119.

Kwan P, Brodie MJ (2005) Potential role of drug transporters in the pathogenesis of medically intractable epilepsy. Epilepsia 46:224-235.

Kwan P, Brodie MJ (2010) Definition of refractory epilepsy: defining the indefinable? Lancet Neurol 9:27-29.

Kwan P, Poon WS, Ng HK, Kang DE, Wong V, Ng PW, Lui CH, Sin NC, Wong KS, Baum L (2008) Multidrug resistance in epilepsy and polymorphisms in the voltage-gated sodium channel genes SCN1A, SCN2A, and SCN3A: correlation among phenotype, genotype, and mRNA expression. Pharmacogenet Genomics 18:989-998.

Kwan P, Schachter SC, Brodie MJ (2011) Drug-resistant epilepsy. N Engl J Med 365:919-926.

Kwan P, Sills GJ, Brodie MJ (2001) The mechanisms of action of commonly used antiepileptic drugs. Pharmacol Ther 90:21-34.

Kwan P, Sperling MR (2009) Refractory seizures: try additional antiepileptic drugs (after two have failed) or go directly to early surgery evaluation? Epilepsia 50 Suppl 8:57-62.

Kwan P, Wong V, Ng PW, Lui CH, Sin NC, Poon WS, Ng HK, Wong KS, Baum L (2009) Gene-wide tagging study of association between ABCB1 polymorphisms and multidrug resistance in epilepsy in Han Chinese. Pharmacogenomics 10:723-732.

Lakhan R, Kumari R, Misra UK, Kalita J, Pradhan S, Mittal B (2009) Differential role of sodium channels SCN1A and SCN2A gene polymorphisms with epilepsy and multiple drug resistance in the north Indian population. Br J Clin Pharmacol 68:214-220.

Lamba JK, Lin YS, Schuetz EG, Thummel KE (2002) Genetic contribution to variable human CYP3A-mediated metabolism. Adv Drug Deliv Rev 54:1271-1294.

Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, Robles V (2006) Machine learning in bioinformatics. Brief Bioinform 7:86-112.

Leabman MK, Huang CC, DeYoung J, Carlson EJ, Taylor TR, de la Cruz M, Johns SJ, Stryke D, Kawamoto M, Urban TJ, Kroetz DL, Ferrin TE, Clark AG, Risch N, Herskowitz I, Giacomini KM (2003) Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proc Natl Acad Sci U S A 100:5896-5901.

Lee JK, Williams PD, Cheon S (2008) Data mining in genomics. Clin Lab Med 28:145-166, viii.

Lee Sj Fau - Goldstein JA, Goldstein JA Functionally defective or altered CYP3A4 and CYP3A5 single nucleotide polymorphisms and their detection with genotyping tests.

Lee SY, Lee ST, Kim JW (2007) Contributions of CYP2C9/CYP2C19 genotypes and drug interaction to the phenytoin treatment in the Korean epileptic patients in the clinical setting. J Biochem Mol Biol 40:448-452.

Leppik IE (2000) Monotherapy and polypharmacy. Neurology 55:S25-29.

Leschziner G, Jorgensen AL, Andrew T, Pirmohamed M, Williamson PR, Marson AG, Coffey AJ, Middleditch C, Rogers J, Bentley DR, Chadwick DW, Balding DJ, Johnson MR (2006) Clinical factors and ABCB1 polymorphisms in prediction of antiepileptic drug response: a prospective cohort study. Lancet Neurol 5:668-676.

Leschziner GD, Andrew T, Leach JP, Chadwick D, Coffey AJ, Balding DJ, Bentley DR, Pirmohamed M, Johnson MR (2007a) Common ABCB1 polymorphisms are not associated

with multidrug resistance in epilepsy using a gene-wide tagging approach. Pharmacogenet Genomics 17:217-220.

Leschziner GD, Andrew T, Pirmohamed M, Johnson MR (2007b) ABCB1 genotype and PGP expression, function and therapeutic drug response: a critical review and recommendations for future research. The pharmacogenomics journal 7:154-179.

Leterrier C, Brachet A, Fache MP, Dargent B (2010) Voltage-gated sodium channel organization in neurons: protein interactions and trafficking pathways. Neuroscience letters 486:92-100.

Lettre G (2012) Using height association studies to gain insights into human idiopathic short and syndromic stature phenotypes. Pediatr Nephrol.

Leu C, de Kovel CG, Zara F, Striano P, Pezzella M, Robbiano A, Bianchi A, Bisulli F, Coppola A, Giallonardo AT, Beccaria F, Trenite DK, Lindhout D, Gaus V, Schmitz B, Janz D, Weber YG, Becker F, Lerche H, Kleefuss-Lie AA, Hallman K, Kunz WS, Elger CE, Muhle H, Stephani U, Moller RS, Hjalgrim H, Mullen S, Scheffer IE, Berkovic SF, Everett KV, Gardiner MR, Marini C, Guerrini R, Lehesjoki AE, Siren A, Nabbout R, Baulac S, Leguern E, Serratosa JM, Rosenow F, Feucht M, Unterberger I, Covanis A, Suls A, Weckhuysen S, Kaneva R, Caglayan H, Turkdogan D, Baykan B, Bebek N, Ozbek U, Hempelmann A, Schulz H, Ruschendorf F, Trucks H, Nurnberg P, Avanzini G, Koeleman BP, Sander T (2012) Genome-wide linkage meta-analysis identifies susceptibility loci at 2q34 and 13q31.3 for genetic generalized epilepsies. Epilepsia 53:308-318.

Levy RH (1995) Cytochrome P450 isozymes and antiepileptic drug interactions. Epilepsia 36 Suppl 5:S8-13.

Levy RH (2002) Antiepileptic drugs. Philadelphia ; London: Lippincott Williams & Wilkins.
Lhatoo SD, Sander JW, Shorvon SD (2001) The dynamics of drug treatment in epilepsy: an observational study in an unselected population based cohort with newly diagnosed epilepsy followed up prospectively over 11-14 years. Journal of neurology, neurosurgery, and psychiatry 71:632-637.

Lin YS, Dowling AL, Quigley SD, Farin FM, Zhang J, Lamba J, Schuetz EG, Thummel KE (2002) Co-regulation of CYP3A4 and CYP3A5 and contribution to hepatic and intestinal midazolam metabolism. Mol Pharmacol 62:162-172.

Listowsky I (2005) A subclass of mu glutathione S-transferases selectively expressed in testis and brain. Methods Enzymol 401:278-287.

Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, Khoury MJ, Cohen B, Davey-Smith G, Grimshaw J, Scheet P, Gwinn M, Williamson RE, Zou GY, Hutchings K, Johnson CY, Tait V, Wiens M, Golding J, van Duijn C, McLaughlin J, Paterson A, Wells G, Fortier I, Freedman M, Zecevic M, King R, Infante-Rivard C, Stewart A, Birkett N (2009) STrengthening the REporting of Genetic Association studies (STREGA)--an extension of the STROBE statement. European journal of clinical investigation 39:247-266.

Liu Z, Sokka T, Maas K, Olsen NJ, Aune TM (2009) Prediction of disease severity in patients with early rheumatoid arthritis by gene expression profiling. Hum Genomics Proteomics 2009.
Lo Monte AI, Damiano G, Mularo A, Palumbo VD, Alessi R, Gioviale MC, Spinelli G, Buscemi G (2011) Comparison between local and regional anesthesia in arteriovenous fistula creation. J Vasc Access 12:331-335.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177-182.

Loscher W (2002) Current status and future directions in the pharmacotherapy of epilepsy. Trends Pharmacol Sci 23:113-118.

Loscher W (2005a) How to explain multidrug resistance in epilepsy? Epilepsy Curr 5:107-112.

Loscher W (2005c) Mechanisms of drug resistance. Epileptic disorders : international epilepsy journal with videotape 7 Suppl 1:S3-9.

Loscher W, Brandt C (2009) High seizure frequency prior to antiepileptic treatment is a predictor of pharmacoresistant epilepsy in a rat model of temporal lobe epilepsy. Epilepsia.

Loscher W, Delanty N (2009) MDR1/ABCB1 polymorphisms and multidrug resistance in epilepsy: in and out of fashion. Pharmacogenomics 10:711-713.

Loscher W, Klotz U, Zimprich F, Schmidt D (2009) The clinical impact of pharmacogenetics on the treatment of epilepsy. Epilepsia 50:1-23.

Loscher W, Luna-Tortos C, Romermann K, Fedrowitz M (2011) Do ATP-Binding Cassette Transporters Cause Pharmacoresistance in Epilepsy? Problems and Approaches in Determining which Antiepileptic Drugs are Affected. Curr Pharm Des 17:2808-2828.

Loscher W, Potschka H (2002) Role of multidrug transporters in pharmacoresistance to antiepileptic drugs. J Pharmacol Exp Ther 301:7-14.

Loscher W, Potschka H (2005a) Drug resistance in brain diseases and the role of drug efflux transporters. Nat Rev Neurosci 6:591-602.

Loscher W, Potschka H (2005c) Role of drug efflux transporters in the brain for drug disposition and treatment of brain diseases. Prog Neurobiol 76:22-76.

Loscher W, Schmidt D (2011) Modern antiepileptic drug development has failed to deliver: ways out of the current dilemma. Epilepsia 52:657-678.

Lossin C (2009) A catalog of SCN1A variants. Brain & development 31:114-130.

Lossin C, Rhodes TH, Desai RR, Vanoye CG, Wang D, Carniciu S, Devinsky O, George AL, Jr. (2003) Epilepsy-associated dysfunction in the voltage-gated neuronal sodium channel SCN1A. J Neurosci 23:11289-11295.

Lossin C, Wang DW, Rhodes TH, Vanoye CG, George AL, Jr. (2002) Molecular basis of an inherited epilepsy. Neuron 34:877-884.

Loup F, Picard F, Yonekawa Y, Wieser HG, Fritschy JM (2009) Selective changes in GABAA receptor subtypes in white matter neurons of patients with focal epilepsy. Brain : a journal of neurology 132:2449-2463.

Lucas PT, Meadows LS, Nicholls J, Ragsdale DS (2005) An epilepsy mutation in the beta1 subunit of the voltage-gated sodium channel results in reduced channel sensitivity to phenytoin. Epilepsy research 64:77-84.

Lucek PR, Ott J (1997) Neural network analysis of complex traits. Genet Epidemiol 14:1101-1106.

Luciano AL, Shorvon SD (2007) Results of treatment changes in patients with apparently drug-resistant chronic epilepsy. Ann Neurol 62:375-381.

Luna-Tortos C, Fedrowitz M, Loscher W (2008) Several major antiepileptic drugs are substrates for human P-glycoprotein. Neuropharmacology 55:1364-1375.

Lynch BA, Lambeng N, Nocka K, Kensel-Hammes P, Bajjalieh SM, Matagne A, Fuks B (2004) The synaptic vesicle protein SV2A is the binding site for the antiepileptic drug levetiracetam. Proc Natl Acad Sci U S A 101:9861-9866.

Ma MK, Woo MH, McLeod HL (2002) Genetic basis of drug metabolism. American journal of health-system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists 59:2061-2069.

MacDonald BK, Johnson AL, Goodridge DM, Cockerell OC, Sander JW, Shorvon SD (2000) Factors predicting prognosis of epilepsy after presentation with seizures. Ann Neurol 48:833-841.

Macdonald RL, Gallagher MJ, Feng HJ, Kang J (2004) GABA(A) receptor epilepsy mutations. Biochem Pharmacol 68:1497-1506.

Macdonald RL, Kelly KM (1995) Antiepileptic drug mechanisms of action. Epilepsia 36 Suppl 2:S2-12.

Madden S, Maggs JL, Park BK (1996) Bioactivation of carbamazepine in the rat in vivo. Evidence for the formation of reactive arene oxide(s). Drug Metab Dispos 24:469-479.

Magiorkinis E, Sidiropoulou K, Diamantis A (2010) Hallmarks in the history of epilepsy: epilepsy in antiquity. Epilepsy & behavior : E&B 17:103-108.

Makmor-Bakry M, Sills GJ, Hitiris N, Butler E, Wilson EA, Brodie MJ (2009) Genetic variants in microsomal epoxide hydrolase influence carbamazepine dosing. Clin Neuropharmacol 32:205-212.

Malo MS, Blanchard BJ, Andresen JM, Srivastava K, Chen XN, Li X, Jabs EW, Korenberg JR, Ingram VM (1994) Localization of a putative human brain sodium channel gene (SCN1A) to chromosome band 2q24. Cytogenet Cell Genet 67:178-186.

Mamiya K, Hadama A, Yukawa E, Ieiri I, Otsubo K, Ninomiya H, Tashiro N, Higuchi S (2000) CYP2C19 polymorphism effect on phenobarbitone. Pharmacokinetics in Japanese patients with epilepsy: analysis by population pharmacokinetics. Eur J Clin Pharmacol 55:821-825.

Mamiya K, Ieiri I, Shimamoto J, Yukawa E, Imai J, Ninomiya H, Yamada H, Otsubo K, Higuchi S, Tashiro N (1998) The effects of genetic polymorphisms of CYP2C9 and CYP2C19 on phenytoin metabolism in Japanese adult patients with epilepsy: studies in stereoselective hydroxylation and population pharmacokinetics. Epilepsia 39:1317-1323.

Mancinelli L, Cronin M, Sadee W (2000) Pharmacogenomics: the promise of personalized medicine. AAPS PharmSci 2:E4.

Mann MW, Pons G (2007) Various pharmacogenetic aspects of antiepileptic drug therapy: a review. CNS Drugs 21:143-164.

Manna I, Gambardella A, Bianchi A, Striano P, Tozzi R, Aguglia U, Beccaria F, Benna P, Campostrini R, Canevini MP, Condino F, Durisotti C, Elia M, Giallonardo AT, Iudice A, Labate A, La Neve A, Michelucci R, Muscas GC, Paravidino R, Zaccara G, Zucca C, Zara F, Perucca E (2011) A functional polymorphism in the SCN1A gene does not influence antiepileptic drug responsiveness in Italian patients with focal epilepsy. Epilepsia 52:e40-44.

Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M, Donnelly P, Faraone SV, Frazer K, Gabriel S, Gejman P, Guttmacher A, Harris EL, Insel T, Kelsoe JR, Lander E, McCowin N, Mailman MD, Nabel E, Ostell J, Pugh E, Sherry S, Sullivan PF, Thompson JF, Warram J, Wholley D, Milos PM, Collins FS (2007) New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nat Genet 39:1045-1051.

Mansmann U, Winkelmann BR (2002) Classification and prediction in pharmacogenetics--context, construction and validation. Pharmacogenomics 3:157-160.

Marban E, Yamagishi T, Tomaselli GF (1998) Structure and function of voltage-gated sodium channels. J Physiol 508 (Pt 3):647-657.

March PA (1998) Seizures: classification, etiologies, and pathophysiology. Clin Tech Small Anim Pract 13:119-131.

Marson A, Smith CT, Williamson P, Smith D, Jacoby A, Chadwick D (2006) Multicentre randomised controlled trial comparing Standard and New Antiepileptic Drugs (SANAD). Epilepsia 47:1-1.

Marson AG, Al-Kharusi AM, Alwaidh M, Appleton R, Baker GA, Chadwick DW, Cramp C, Cockerell OC, Cooper PN, Doughty J, Eaton B, Gamble C, Goulding PJ, Howell SJ, Hughes A, Jackson M, Jacoby A, Kellett M, Lawson GR, Leach JP, Nicolaides P, Roberts R, Shackley P, Shen J, Smith DF, Smith PE, Smith CT, Vanoli A, Williamson PR (2007a) The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. Lancet 369:1000-1015.

Marson AG, Al-Kharusi AM, Alwaidh M, Appleton R, Baker GA, Chadwick DW, Cramp C, Cockerell OC, Cooper PN, Doughty J, Eaton B, Gamble C, Goulding PJ, Howell SJ, Hughes A, Jackson M, Jacoby A, Kellett M, Lawson GR, Leach JP, Nicolaides P, Roberts R, Shackley P, Shen J, Smith DF, Smith PE, Smith CT, Vanoli A, Williamson PR (2007b) The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. Lancet 369:1016-1026.

Martin R, Rose D, Yu K, Barros S (2006) Toxicogenomics strategies for predicting drug toxicity. Pharmacogenomics 7:1003-1016.

Mattson RH, Cramer JA, Collins JF (1996) Prognosis for total control of complex partial and secondarily generalized tonic clonic seizures. Department of Veterans Affairs Epilepsy Cooperative Studies No. 118 and No. 264 Group. Neurology 47:68-76.

Mattson RH, Cramer JA, Collins JF, Smith DB, Delgado-Escueta AV, Browne TR, Williamson PD, Treiman DM, McNamara JO, McCutchen CB, et al. (1985) Comparison of carbamazepine, phenobarbital, phenytoin, and primidone in partial and secondarily generalized tonic-clonic seizures. N Engl J Med 313:145-151.

McCarthy JJ, Hilfiker R (2000) The use of single-nucleotide polymorphism maps in pharmacogenomics. Nat Biotechnol 18:505-508.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356-369.

McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: past, present and future. Hum Mol Genet 17:R100-101.

McCormack M, Alfirevic A, Bourgeois S, Farrell JJ, Kasperaviciute D, Carrington M, Sills GJ, Marson T, Jia X, de Bakker PI, Chinthapalli K, Molokhia M, Johnson MR, O'Connor GD, Chaila E, Alhusaini S, Shianna KV, Radtke RA, Heinzen EL, Walley N, Pandolfo M, Pichler W, Park BK, Depondt C, Sisodiya SM, Goldstein DB, Deloukas P, Delanty N, Cavalleri GL, Pirmohamed M (2011) HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. N Engl J Med 364:1134-1143.

McCormick DA, Contreras D (2001) On the cellular and network bases of epileptic seizures. Annu Rev Physiol 63:815-846.

McCorry D, Chadwick D, Marson A (2004) Current drug treatment of epilepsy in adults. Lancet neurology 3:729-735.

McKinney BA, Reif DM, Ritchie MD, Moore JH (2006) Machine learning for detecting gene-gene interactions: a review. Appl Bioinformatics 5:77-88.

McLeod HL (2005) Pharmacogenetic analysis of clinically relevant genetic polymorphisms. Clin Infect Dis 41 Suppl 7:S449-452.

McLeod HL, Evans WE (2001) Pharmacogenomics: unlocking the human genome for better drug therapy. Annu Rev Pharmacol Toxicol 41:101-121.

McNamara JO (1994) Cellular and molecular basis of epilepsy. J Neurosci 14:3413-3425.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, Fulton R, Kucaba TA, Wagner-McPherson C, Barbazuk WB, Gregory SG, Humphray SJ, French L, Evans RS, Bethel G, Whittaker A, Holden JL, McCann OT, Dunham A, Soderlund C, Scott CE, Bentley DR, Schuler G, Chen HC, Jang W, Green ED, Idol JR, Maduro VV, Montgomery KT, Lee E, Miller A, Emerling S, Kucherlapati, Gibbs R, Scherer S, Gorrell JH, Sodergren E, Clerc-Blankenburg K, Tabor P, Naylor S, Garcia D, de Jong PJ, Catanese JJ, Nowak N, Osoegawa K, Qin S, Rowen L, Madan A, Dors M, Hood L, Trask B, Friedman C, Massa H, Cheung VG, Kirsch IR, Reid T, Yonescu R, Weissenbach J, Bruls T, Heilig R, Branscomb E, Olsen A, Doggett N, Cheng JF, Hawkins T, Myers RM, Shang J, Ramirez L, Schmutz J, Velasquez O, Dixon K, Stone NE, Cox DR, Haussler D, Kent WJ, Furey T, Rogic S, Kennedy S, Jones S, Rosenthal A, Wen G, Schilhabel M, Gloeckner G, Nyakatura G, Siebert R, Schlegelberger B, Korenberg J, Chen XN, Fujiyama A, Hattori M, Toyoda A, Yada T, Park HS, Sakaki Y, Shimizu N, Asakawa S, Kawasaki K, Sasaki T, Shintani A, Shimizu A, Shibuya K, Kudoh J, Minoshima S, Ramser J, Seranski P, Hoff C, Poustka A, Reinhardt R, Lehrach H (2001) A physical map of the human genome. Nature 409:934-941.

Mefford HC, Eichler EE (2009) Duplication hotspots, rare genomic disorders, and common disease. Curr Opin Genet Dev 19:196-204.

Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, Baker C, Franke A, Malafosse A, Genton P, Thomas P, Gurnett CA, Schreiber S, Bassuk AG, Guipponi M, Stephani U, Helbig I, Eichler EE (2010) Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. PLoS Genet 6:e1000962.

Meisel C, Roots I, Cascorbi I, Brinkmann U, Brockmoller J (2000) How to manage individualized drug therapy: application of pharmacogenetic knowledge of drug metabolism and transport. Clin Chem Lab Med 38:869-876.

Meisler MH, Kearney J, Ottman R, Escayg A (2001) Identification of epilepsy genes in human and mouse. Annu Rev Genet 35:567-588.

Meisler MH, Kearney JA (2005) Sodium channel mutations in epilepsy and other neurological disorders. J Clin Invest 115:2010-2017.

Meisler MH, O'Brien JE, Sharkey LM (2010) Sodium channel gene family: epilepsy mutations, gene interactions and modifier effects. The Journal of Physiology 588:1841-1848.

Meldrum BS (1995) Excitatory amino acid receptors and their role in epilepsy and cerebral ischemia. Ann N Y Acad Sci 757:492-505.

Meldrum BS (1996) Update on the mechanism of action of antiepileptic drugs. Epilepsia 37 Suppl 6:S4-11.

Meldrum BS (2000) Glutamate as a neurotransmitter in the brain: review of physiology and pathology. J Nutr 130:1007S-1015S.

Meldrum BS, Rogawski MA (2007) Molecular targets for antiepileptic drug development. Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics 4:18-61.

Meng HM, Ren JY, Lv YD, Lin WH, Guo YJ (2011) Association study of CYP3A5 genetic polymorphism with serum concentrations of carbamazepine in Chinese epilepsy patients. Neurol Asia 16:39-45.

Meyer TE, Verwoert GC, Hwang SJ, Glazer NL, Smith AV, van Rooij FJ, Ehret GB, Boerwinkle E, Felix JF, Leak TS, Harris TB, Yang Q, Dehghan A, Aspelund T, Katz R, Homuth G, Kocher T, Rettig R, Ried JS, Gieger C, Prucha H, Pfeufer A, Meitinger T, Coresh J, Hofman A, Sarnak MJ, Chen YD, Uitterlinden AG, Chakravarti A, Psaty BM, van Duijn CM, Kao WH, Witteman JC, Gudnason V, Siscovick DS, Fox CS, Kottgen A (2010) Genome-wide association studies of serum magnesium, potassium, and sodium concentrations identify six Loci influencing serum magnesium levels. PLoS Genet 6.

Miners JO, McKinnon RA, Mackenzie PI (2002) Genetic polymorphisms of UDP-glucuronosyltransferases and their functional significance. Toxicology 181-182:453-456.

Moeller JJ, Rahey SR, Sadler RM (2009) Lamotrigine-valproic acid combination therapy for medically refractory epilepsy. Epilepsia 50:475-479.

Mohanraj R, Brodie MJ (2005) Outcomes in newly diagnosed localization-related epilepsies. Seizure 14:318-323.

Mohanraj R, Brodie MJ (2006) Diagnosing refractory epilepsy: response to sequential treatment schedules. Eur J Neurol 13:277-282.

Mohanraj R, Brodie MJ (2007) Outcomes of newly diagnosed idiopathic generalized epilepsy syndromes in a non-pediatric setting. Acta Neurol Scand 115:204-208.

Mohanraj R, Norrie J, Stephen LJ, Kelly K, Hitiris N, Brodie MJ (2006) Mortality in adults with newly diagnosed and chronic epilepsy: a retrospective comparative study. Lancet Neurol 5:481-487.

Moon H, Ahn H, Kodell RL, Baek S, Lin C-J, Chen JJ (2007) Ensemble methods for classification of patients for personalized medicine with high-dimensional data. Artif Intell Med 41:197-207.

Moore JH (2004) Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn 4:795-803.

Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics 26:445-455.

Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 241:252-261.

Moore JH, Hahn LW, Ritchie MD, Thornton TA, White BC (2004) Routine Discovery of Complex Genetic Models using Genetic Algorithms. Appl Soft Comput 4:79-86.

Moore JH, Ritchie MD (2004) STUDENTJAMA. The challenges of whole-genome approaches to common diseases. Jama 291:1642-1643.

Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD (2008) Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol 32:325-340.

Motsinger-Reif AA, Jorgenson E, Relling MV, Kroetz DL, Weinshilboum R, Cox NJ, Roden DM (2010) Genome-wide association studies in pharmacogenomics: successes and lessons. Pharmacogenetics and genomics.

Motsinger AA, Lee SL, Mellick G, Ritchie MD (2006) GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. BMC Bioinformatics 7:39.

Mullen SA, Crompton DE, Carney PW, Helbig I, Berkovic SF (2009) A neurologist's guide to genome-wide association studies. Neurology 72:558-565.

Mulley JC, Dibbens LM (2009) Chipping away at the common epilepsies with complex genetics: the 15q13.3 microdeletion shows the way. Genome Med 1:33.

Mulley JC, Scheffer IE, Harkin LA, Berkovic SF, Dibbens LM (2005) Susceptibility genes for complex epilepsy. Hum Mol Genet 14 Spec No. 2:R243-249.

Mulley JC, Scheffer IE, Petrou S, Berkovic SF (2003) Channelopathies as a genetic cause of epilepsy. Current opinion in neurology 16:171-176.

Nagar S, Blanchard RL (2006) Pharmacogenetics of uridine diphosphoglucuronosyltransferase (UGT) 1A family members and its role in patient response to irinotecan. Drug Metab Rev 38:393-409.

Nagasawa K, Nakahara Y (1992) [Clinical application of therapeutic drug monitoring (TDM)]. Nihon Ika Daigaku Zasshi 59:2-8.

Nakajima Y, Saito Y, Shiseki K, Fukushima-Uesaka H, Hasegawa R, Ozawa S, Sugai K, Katoh M, Saitoh O, Ohnuma T, Kawai M, Ohtsuki T, Suzuki C, Minami N, Kimura H, Goto Y, Kamatani N, Kaniwa N, Sawada J (2005) Haplotype structures of EPHX1 and their effects on the metabolism of carbamazepine-10,11-epoxide in Japanese epileptic patients. Eur J Clin Pharmacol 61:25-34.

Nebert DW (1999) Pharmacogenetics and pharmacogenomics: why is this relevant to the clinical geneticist? Clin Genet 56:247-258.

Nebert DW (2008) Pharmacogenetics and Pharmacogenomics. In: Encyclopedia of Life Sciences (ELS) Chichester: John Wiley & Sons, Ltd.

Nebert DW, Dieter MZ (2000) The evolution of drug metabolism. Pharmacology 61:124-135.

Nebert DW, Zhang G, Vesell ES (2008a) From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. Drug Metab Rev 40:187-224.

Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11:863-874.

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812-3814.

Ng PC, Kumar P, Henikoff S (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073-1082.

Odani A, Hashimoto Y, Otsuki Y, Uwai Y, Hattori H, Furusho K, Inui K (1997) Genetic polymorphism of the CYP2C subfamily and its effect on the pharmacokinetics of phenytoin in Japanese patients with epilepsy. Clin Pharmacol Ther 62:287-292.

Odani A, Hashimoto Y, Takayanagi K, Otsuki Y, Koue T, Takano M, Yasuhara M, Hattori H, Furusho K, Inui K (1996) Population pharmacokinetics of phenytoin in Japanese patients with epilepsy: analysis with a dose-dependent clearance model. Biological & pharmaceutical bulletin 19:444-448.

Okada Y, Seo T, Ishitsu T, Wanibuchi A, Hashimoto N, Higa Y, Nakagawa K (2008) Population estimation regarding the effects of cytochrome P450 2C19 and 3A5 polymorphisms on zonisamide clearance. Ther Drug Monit 30:540-543.

Olsen RW, Avoli M (1997) GABA and epileptogenesis. Epilepsia 38:399-407.

Orr N, Chanock S (2008) Common genetic variation and human disease. Adv Genet 62:1-32.

Ouahchi K, Lindeman N, Lee C (2006) Copy number variants and pharmacogenomics. Pharmacogenomics 7:25-29.

Ozeki T, Mushiroda T, Yowang A, Takahashi A, Kubo M, Shirakata Y, Ikezawa Z, Iijima M, Shiohara T, Hashimoto K, Kamatani N, Nakamura Y (2011a) Genome-wide association study identifies HLA-A*3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. Human molecular genetics 20:1034-1041.

Ozeki T, Mushiroda T, Yowang A, Takahashi A, Kubo M, Shirakata Y, Ikezawa Z, Iijima M, Shiohara T, Hashimoto K, Kamatani N, Nakamura Y (2011b) Genome-wide association study identifies HLA-A*3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. Hum Mol Genet 20:1034-1041.

Pagani F, Baralle FE (2004) Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet 5:389-396.

Panayiotopoulos CP, International League against Epilepsy. (2005) The epilepsies : seizures, syndromes and management : based on the ILAE classifications and practice parameter guidelines. Chipping Norton, Oxon.: Bladon Medical.
Pander J, Wessels JA, Mathijssen RH, Gelderblom H, Guchelaar HJ (2010) Pharmacogenetics of tomorrow: the 1 + 1 = 3 principle. Pharmacogenomics 11:1011-1017.

Pang GS, Wang J, Wang Z, Lee CG (2009) Predicting potentially functional SNPs in drug-response genes. Pharmacogenomics 10:639-653.

Park BK, Pirmohamed M, Kitteringham NR (1995) The role of cytochrome P450 enzymes in hepatic and extrahepatic human drug toxicity. Pharmacology & therapeutics 68:385-424.

Park PW, Seo YH, Ahn JY, Kim KA, Park JY (2009) Effect of CYP3A5*3 genotype on serum carbamazepine concentrations at steady-state in Korean epileptic patients. J Clin Pharm Ther 34:569-574.

Patsalos PN, Bourgeois BFD (2010) The epilepsy prescriber's guide to antiepileptic drugs. Cambridge: Cambridge University Press.

Patsalos PN, Perucca E (2003) Clinically important drug interactions in epilepsy: interactions between antiepileptic drugs and other drugs. Lancet neurology 2:473-481.

Pearce RE, Lu W, Wang Y, Uetrecht JP, Correia MA, Leeder JS (2008) Pathways of Carbamazepine Bioactivation in Vitro. III. The Role of Human Cytochrome P450 Enzymes in the Formation of 2,3-Dihydroxycarbamazepine. Drug Metabolism and Disposition 36:1637-1649.

Perucca E (1999) The spectrum of the new antiepileptic drugs. Acta Neurol Belg 99:231-238.

Perucca E (2001a) Clinical pharmacology and therapeutic use of the new antiepileptic drugs. Fundam Clin Pharmacol 15:405-417.

Perucca E (2002a) Marketed new antiepileptic drugs: are they better than old-generation agents? Ther Drug Monit 24:74-80.

Perucca E (2002b) Patient-tailored antiepileptic drug therapy: predicting response to antiepileptic drugs. Int Congr Ser 1244:93-103.

Perucca E (2006) Clinically relevant drug interactions with antiepileptic drugs. Br J Clin Pharmacol 61:246-255.

Perucca E, Dulac O, Shorvon S, Tomson T (2001) Harnessing the clinical potential of antiepileptic drug therapy: dosage optimisation. CNS drugs 15:609-621.

Perucca E, Meador KJ (2005) Adverse effects of antiepileptic drugs. Acta Neurol Scand Suppl 181:30-35.

Petrovski S, Szoeke CE, Jones NC, Salzberg MR, Sheffield LJ, Huggins RM, O'Brien TJ (2010) Neuropsychiatric symptomatology predicts seizure recurrence in newly treated patients. Neurology 75:1015-1021.

Petrovski S, Szoeke CE, Sheffield LJ, D'Souza W, Huggins RM, O'Brien T J (2009) Multi-SNP pharmacogenomic classifier is superior to single-SNP models for predicting drug outcome in complex diseases. Pharmacogenet Genomics 19:147-152.

Petsche H, Brazier MAB, ©*sterreichische Akademie der Wissenschaften. (1972) Synchronization of EEG activity in epilepsies : a symposium organized by the Austrian Academy of Sciences, Vienna, Austria, September 12-13, 1971. New York: Springer-Verlag.

Picard F, Bertrand S, Steinlein OK, Bertrand D (1999) Mutated nicotinic receptors responsible for autosomal dominant nocturnal frontal lobe epilepsy are more sensitive to carbamazepine. Epilepsia 40:1198-1209.

Poirier J, Delisle MC, Quirion R, Aubert I, Farlow M, Lahiri D, Hui S, Bertrand P, Nalbantoglu J, Gilfix BM, Gauthier S (1995) Apolipoprotein E4 allele as a predictor of cholinergic deficits and treatment outcome in Alzheimer disease. Proc Natl Acad Sci U S A 92:12260-12264.

Poolos NP, Warner LN, Humphreys SZ, Williams S (2012) Comparative efficacy of combination drug therapy in refractory epilepsy. Neurology 78:62-68.

Porter RJ, Rogawski MA (1992) New antiepileptic drugs: from serendipity to rational discovery. Epilepsia 33 Suppl 1:S1-6.

Prokunina L, Alarcon-Riquelme ME (2004) Regulatory SNPs in complex diseases: their identification and functional validation. Expert reviews in molecular medicine 6:1-15.

Qu J, Zhou BT, Yin JY, Xu XJ, Zhao YC, Lei GH, Tang Q, Zhou HH, Liu ZQ (2012) ABCC2 polymorphisms and haplotype are associated with drug resistance in Chinese epileptic patients. CNS neuroscience & therapeutics 18:647-651.
R Development Core Team (2010). R: A language and environment for statistical computing.
Ragsdale DS, Avoli M (1998) Sodium channels as molecular targets for antiepileptic drugs. Brain Res Brain Res Rev 26:16-28.
Ragsdale DS, McPhee JC, Scheuer T, Catterall WA (1996) Common molecular determinants of local anesthetic, antiarrhythmic, and anticonvulsant block of voltage-gated Na+ channels. Proc Natl Acad Sci U S A 93:9270-9275.
Ramachandran V, Shorvon SD (2003) Clues to the genetic influences of drug responsiveness in epilepsy. Epilepsia 44 Suppl 1:33-37.
Rang HP (2003) Pharmacology. Edinburgh: Churchill Livingstone.
Rees MI (2010) The genetics of epilepsy--the past, the present and future. Seizure 19:680-683.

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502-510.
Remy S, Beck H (2006) Molecular and cellular mechanisms of pharmacoresistance in epilepsy. Brain : a journal of neurology 129:18-35.
Remy S, Gabriel S, Urban BW, Dietrich D, Lehmann TN, Elger CE, Heinemann U, Beck H (2003) A novel mechanism underlying drug resistance in chronic epilepsy. Ann Neurol 53:469-479.

Reutens DC, Berkovic SF (1995) Idiopathic generalized epilepsy of adolescence: are the syndromes clinically distinct? Neurology 45:1469-1476.

Risch N (2000a) Searching for genes in complex diseases: lessons from systemic lupus erythematosus. J Clin Invest 105:1503-1506.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517.

Risch NJ (2000b) Searching for genetic determinants in the new millennium. Nature 405:847-856.

Ritchie MD (2012) The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era. Hum Genet 131:1615-1626.

Ritchie MD, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24:150-157.

Ritchie MD, Motsinger AA (2005) Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies. Pharmacogenomics 6:823-834.

Robey RW, Lazarowski A, Bates SE (2008) P-glycoprotein--a clinical target in drug-refractory epilepsy? Mol Pharmacol 73:1343-1346.

Robinson R, Gardiner M (2004) Molecular basis of Mendelian idiopathic epilepsies. Ann Med 36:89-97.

Roden DM, George AL, Jr. (2002) The genetic basis of variability in drug responses. Nat Rev Drug Discov 1:37-44.

Rodin AS, Gogoshin G, Boerwinkle E (2011) Systems biology data analysis methodology in pharmacogenomics. Pharmacogenomics 12:1349-1360.

Rogawski MA (2011) Revisiting AMPA receptors as an antiepileptic drug target. Epilepsy Curr 11:56-63.

Rogawski MA, Loscher W (2004) The neurobiology of antiepileptic drugs. Nat Rev Neurosci 5:553-564.

Rogawski MA, Porter RJ (1990) Antiepileptic drugs: pharmacological mechanisms and clinical efficacy with consideration of promising developmental stage compounds. Pharmacological reviews 42:223-286.

Roses AD (2000) Pharmacogenetics and the practice of medicine. Nature 405:857-865.

Rowland A, Elliot DJ, Williams JA, Mackenzie PI, Dickinson RG, Miners JO (2006) In vitro characterization of lamotrigine N2-glucuronidation and the lamotrigine-valproic acid interaction. Drug Metab Dispos 34:1055-1062.

Sabbagh A, Genin E, Darlu P (2008) Selecting predictive markers for pharmacogenetic traits: tagging vs. data-mining approaches. Hum Hered 66:10-18.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante

M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-933.

Salinas AE, Wong MG (1999) Glutathione S-transferases--a review. Curr Med Chem 6:279-309.

Sanchez MB, Herranz JL, Leno C, Arteaga R, Oterino A, Valdizan EM, Nicolas JM, Adin J, Armijo JA (2010) Genetic factors associated with drug-resistance of epilepsy: relevance of stratification by patient age and aetiology of epilepsy. Seizure 19:93-101.

Sander JW (1993) Some aspects of prognosis in the epilepsies: a review. Epilepsia 34:1007-1016.

Sander JW (2003a) The epidemiology of epilepsy revisited. Current opinion in neurology 16:165-170.

Sander JW (2003b) Idiopathic generalised epilepsies: not only for the paediatrician. J Neurol Neurosurg Psychiatry 74:147.

Sander JW (2004) The use of antiepileptic drugs--principles and practice. Epilepsia 45 Suppl 6:28-34.

Sander JW, Bell GS (2004) Reducing mortality: an important aim of epilepsy management. J Neurol Neurosurg Psychiatry 75:349-351.

Sander T, Schulz H, Saar K, Gennaro E, Riggio MC, Bianchi A, Zara F, Luna D, Bulteau C, Kaminska A, Ville D, Cieuta C, Picard F, Prud'homme JF, Bate L, Sundquist A, Gardiner RM, Janssen GA, de Haan GJ, Kasteleijn-Nolst-Trenite DG, Bader A, Lindhout D, Riess O, Wienker TF, Janz D, Reis A (2000) Genome search for susceptibility loci of common idiopathic generalised epilepsies. Hum Mol Genet 9:1465-1472.

Saruwatari J, Ishitsu T, Nakagawa K (2010) Update on the Genetic Polymorphisms of Drug-Metabolizing Enzymes in Antiepileptic Drug Therapy. Pharmaceuticals 3:2709-2732.

Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM (2007) Silent polymorphisms speak: how they affect pharmacogenomics and the treatment of cancer. Cancer Res 67:9609-9612.

Schachter SC (2007) Currently available antiepileptic drugs. Neurotherapeutics 4:4-11.

Scharfman HE (2007) The neurobiology of epilepsy. Curr Neurol Neurosci Rep 7:348-354.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American journal of human genetics 78:629-644.

Scherrmann JM (2002) Drug delivery to brain via the blood-brain barrier. Vascul Pharmacol 38:349-354.

Schinkel AH (1997) The physiological function of drug-transporting P-glycoproteins. Semin Cancer Biol 8:161-170.

Schmidt D, Loscher W (2005) Drug resistance in epilepsy: putative neurobiologic and clinical mechanisms. Epilepsia 46:858-877.

Schmidt D, Loscher W (2009) New developments in antiepileptic drug resistance: an integrative view. Epilepsy Curr 9:47-52.

Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19:212-219.

Schug J (2008) Using TESS to predict transcription factor binding sites in DNA sequence. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] Chapter 2:Unit 2 6.

Selim N, Branum GD, Liu X, Whalen R, Boyer TD (2000) Differential lobular induction in rat liver of glutathione S-transferase A1/A2 by phenobarbital. Am J Physiol Gastrointest Liver Physiol 278:G542-550.

Semah F, Picot MC, Adam C, Broglin D, Arzimanoglou A, Bazin B, Cavalcanti D, Baulac M (1998) Is the underlying cause of epilepsy a major prognostic factor for recurrence? Neurology 51:1256-1262.

Seo T, Nagata R, Ishitsu T, Murata T, Takaishi C, Hori M, Nakagawa K (2008a) Impact of CYP2C19 polymorphisms on the efficacy of clobazam therapy. Pharmacogenomics 9:527-537.

Seo T, Nakada N, Ueda N, Hagiwara T, Hashimoto N, Nakagawa K, Ishitsu T (2006) Effect of CYP3A5*3 on carbamazepine pharmacokinetics in Japanese patients with epilepsy. Clin Pharmacol Ther 79:509-510.

Seo T, Pahwa P, McDuffie HH, Yurube K, Egoshi M, Umemoto Y, Ghosh S, Fukushima Y, Nakagawa K (2008b) Association between cytochrome P450 3A5 polymorphism and the lung function in Saskatchewan grain workers. Pharmacogenet Genomics 18:487-493.

Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114.

Shah SC, Kusiak A (2004) Data mining and genetic algorithm based gene/SNP selection. Artif Intell Med 31:183-196.

Shang W, Liu WH, Zhao XH, Sun QJ, Bi JZ, Chi ZF (2008) Expressions of glutathione S-transferase alpha, mu, and pi in brains of medically intractable epileptic patients. BMC Neurosci 9:67.

Shastry BS (2002) SNP alleles in human disease and evolution. J Hum Genet 47:561-566.

Shastry BS (2003) SNPs and haplotypes: genetic markers for disease and drug response (review). Int J Mol Med 11:379-382.

Shastry BS (2004) Role of SNP/haplotype map in gene discovery and drug development: An overview. Drug Development Research 62:143-150.

Shastry BS (2006) Pharmacogenetics and the concept of individualized medicine. The pharmacogenomics journal 6:16-21.

Shi MM, Bleavins MR, de la iglesia FA (1999) Technologies for detecting genetic polymorphisms in pharmacogenomics. Mol Diagn 4:343-351.

Shneker BF, Fountain NB (2003) Epilepsy. Dis Mon 49:426-478.

Shorvon SD (2004) The treatment of epilepsy. Malden ; Oxford: Blackwell Science.

Shorvon SD (2009) Epilepsy. Oxford: Oxford University Press.

Shorvon SD (2010) Handbook of epilepsy treatment. Oxford: Wiley-Blackwell.

Siddiqui A, Kerb R, Weale ME, Brinkmann U, Smith A, Goldstein DB, Wood NW, Sisodiya SM (2003) Association of multidrug resistance in epilepsy with a polymorphism in the drug-transporter gene ABCB1. N Engl J Med 348:1442-1448.

Sillanpaa M (1993) Remission of seizures and predictors of intractability in long-term follow-up. Epilepsia 34:930-936.

Sillanpaa M, Schmidt D (2006) Natural history of treated childhood-onset epilepsy: prospective, long-term population-based study. Brain : a journal of neurology 129:617-624.
Sills GJ (2004) Changing channels: mechanisms and responsiveness to antiepileptic drugs in chronic epilepsy. Epilepsy Curr 4:98-99.

Sills GJ, Kwan P, Butler E, de Lange EC, van den Berg DJ, Brodie MJ (2002) P-glycoprotein-mediated efflux of antiepileptic drugs: preliminary studies in mdr1a knockout mice. Epilepsy & behavior : E&B 3:427-432.

Silva S, Anunciação O, Lotz M (2011) A Comparison of Machine Learning Methods for the Prediction of Breast Cancer. In: Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, vol. 6623 (Pizzuti, C. et al., eds), pp 159-170: Springer Berlin Heidelberg.

Silverman BW, Jones MC (1989) E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). International Statistical Review / Revue Internationale de Statistique 57:233-238.
Simon C, Stieger B, Kullak-Ublick GA, Fried M, Mueller S, Fritschy JM, Wieser HG, Pauli-Magnus C (2007) Intestinal expression of cytochrome P450 enzymes and ABC transporters and carbamazepine and phenytoin disposition. Acta Neurol Scand 115:232-242.

Simon R (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 23:7332-7341.

Simon R, Wang SJ (2006) Use of genomic signatures in therapeutics development in oncology and other diseases. The pharmacogenomics journal 6:166-173.

Sisodiya S, Duncan J (2004) Epilepsy: epidemiology, clinical assessment, investigation and natural history. Medicine 32:47-51.

Sisodiya SM (2003) Mechanisms of antiepileptic drug resistance. Current opinion in neurology 16:197-201.

Sisodiya SM (2005) Genetics of drug resistance. Epilepsia 46 Suppl 10:33-38.

Sisodiya SM, Mefford HC (2011) Genetic contribution to common epilepsies. Current opinion in neurology 24.

Smith M, Wilcox KS, White HS (2007) Discovery of antiepileptic drugs. Neurotherapeutics 4:12-17.

So E (2011) Predictors of outcome in newly diagnosed epilepsy: Clinical, EEG and MRI. Neurol Asia 16:27.

Soranzo N, Kelly L, Martinian L, Burley MW, Thom M, Sali A, Kroetz DL, Goldstein DB, Sisodiya SM (2007) Lack of support for a role for RLIP76 (RALBP1) in response to treatment or predisposition to epilepsy. Epilepsia 48:674-683.

Spear BB (2001) Pharmacogenetics and antiepileptic drugs. Epilepsia 42 Suppl 5:31-34.

Spear BB, Heath-Chiozzi M, Huff J (2001) Clinical application of pharmacogenetics. Trends Mol Med 7:201-204.

Speed D, Hoggart C, Petrovski S, Tachmazidou I, Coffey A, Jorgensen A, Eleftherohorinou H, De Iorio M, Todaro M, De T, Smith D, Smith PE, Jackson M, Cooper P, Kellett M, Howell S, Newton M, Yerra R, Tan M, French C, Reuber M, Sills GE, Chadwick D, Pirmohamed M, Bentley D, Scheffer I, Berkovic S, Balding D, Palotie A, Marson A, O'Brien TJ, Johnson MR (2013) A genome-wide association study and biological pathway analysis of epilepsy prognosis in a prospective cohort of newly treated epilepsy. Hum Mol Genet.

Staines AG, Coughtrie MW, Burchell B (2004) N-glucuronidation of carbamazepine in human tissues is mediated by UGT2B7. J Pharmacol Exp Ther 311:1131-1137.

Steinlein OK (2004) Genetic mechanisms that underlie epilepsy. Nat Rev Neurosci 5:400-408.
Steinlein OK (2008) Genetics and epilepsy. Dialogues Clin Neurosci 10:29-38.

Stephen LJ, Brodie MJ (2002) Seizure freedom with more than one antiepileptic drug. Seizure 11:349-351.

Stephen LJ, Kwan P, Brodie MJ (2001) Does the cause of localisation-related epilepsy influence the response to antiepileptic drug treatment? Epilepsia 42:357-362.

Stephen LJ, Sills GJ, Leach JP, Butler E, Parker P, Hitiris N, Leach VM, Wilson EA, Brodie MJ (2007) Sodium valproate versus lamotrigine: a randomised comparison of efficacy, tolerability and effects on circulating androgenic hormones in newly diagnosed epilepsy. Epilepsy research 75:122-129.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. American journal of human genetics 73:1162-1169.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. American journal of human genetics 68:978-989.

Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF, Jr., Hampton GM (2001) Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 61:7388-7393.

Swen JJ, Huizinga TW, Gelderblom H, de Vries EG, Assendelft WJ, Kirchheiner J, Guchelaar HJ (2007) Translating pharmacogenomics: challenges on the road to the clinic. PLoS Med 4:e209.

Szoeke C, Sills GJ, Kwan P, Petrovski S, Newton M, Hitiris N, Baum L, Berkovic SF, Brodie MJ, Sheffield LJ, O'Brien TJ (2009) Multidrug-resistant genotype (ABCB1) and seizure recurrence in newly treated epilepsy: Data from International Pharmacogenetic Cohorts. Epilepsia.

Szoeke CE, Newton M, Wood JM, Goldstein D, Berkovic SF, TJ OB, Sheffield LJ (2006a) Update on pharmacogenetics in epilepsy: a brief review. Lancet Neurol 5:189-196.

Szymczak S, Biernacka JM, Cordell HJ, Gonzalez-Recio O, Konig IR, Zhang H, Sun YV (2009) Machine learning in genome-wide association studies. Genet Epidemiol 33 Suppl 1:S51-57.

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3:391-397.

Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, Whittaker P, Ranganath V, Kumanduri V, McLaren W, Holm L, Lindh J, Rane A, Wadelius M, Deloukas P (2009) A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet 5:e1000433.

Talati R, Scholle JM, Phung OJ, Baker WL, Baker EL, Ashaye A, Kluger J, Quercia R, Mather J, Giovenale S, Coleman CI, White CM (2011) In: Effectiveness and Safety of Antiepileptic Medications in Patients With Epilepsy Rockville (MD).

Tan NC, Berkovic SF (2010) The Epilepsy Genetic Association Database (epiGAD): analysis of 165 genetic association studies, 1996-2008. Epilepsia 51:686-689.

Tan NC, Mulley JC, Berkovic SF (2004) Genetic association studies in epilepsy: "the truth is out there". Epilepsia 45:1429-1442.

Tang K, Fu DJ, Julien D, Braun A, Cantor CR, Koster H (1999) Chip-based genotyping by mass spectrometry. Proc Natl Acad Sci U S A 96:10016-10020.
Tang W, Borel AG, Abbott FS (1996) Conjugation of glutathione with a toxic metabolite of valproic acid, (E)-2-propyl-2,4-pentadienoic acid, catalyzed by rat hepatic glutathione-S-transferases. Drug Metab Dispos 24:436-446.

Tate SK, Depondt C, Sisodiya SM, Cavalleri GL, Schorge S, Soranzo N, Thom M, Sen A, Shorvon SD, Sander JW, Wood NW, Goldstein DB (2005) Genetic predictors of the maximum doses patients receive during clinical use of the anti-epileptic drugs carbamazepine and phenytoin. Proc Natl Acad Sci U S A 102:5507-5512.

Tate SK, Singh R, Hung CC, Tai JJ, Depondt C, Cavalleri GL, Sisodiya SM, Goldstein DB, Liou HH (2006) A common polymorphism in the SCN1A gene associates with phenytoin serum levels at maintenance dose. Pharmacogenet Genomics 16:721-726.

Tate SK, Sisodiya SM (2007) Multidrug resistance in epilepsy: a pharmacogenomic update. Expert opinion on pharmacotherapy 8:1441-1449.

Tauser (2012) Pharmacogenetics: Matching the Right Foundation at Personalized Medicine in the Right Genomic Era, Clinical Applications of Pharmacogenetics.

Thompson CH, Kahlig KM, George Jr AL (2011) SCN1A splice variants exhibit divergent sensitivity to commonly used antiepileptic drugs. Epilepsia.

Thyagaraju K, Hemavathi B, Vasundhara K, Rao AD, Devi KN (2005) Comparative study on glutathione transferases of rat brain and testis under the stress of phenobarbitol and beta-methylcholanthrene. J Zhejiang Univ Sci B 6:759-769.

Tishler DM, Weinberg KI, Hinton DR, Barbaro N, Annett GM, Raffel C (1995) MDR1 gene expression in brain of patients with medically intractable epilepsy. Epilepsia 36:1-6.

Tomson T, Tybring G, Bertilsson L (1983) Single-dose kinetics and metabolism of carbamazepine-10,11-epoxide. Clin Pharmacol Ther 33:58-65.

Torta R, Keller R (1999) Behavioral, psychotic, and anxiety disorders in epilepsy: etiology, clinical features, and therapeutic implications. Epilepsia 40 Suppl 10:S2-20.

Ueda K, Ishitsu T, Seo T, Ueda N, Murata T, Hori M, Nakagawa K (2007) Glutathione S-transferase M1 null genotype as a risk factor for carbamazepine-induced mild hepatotoxicity. Pharmacogenomics 8:435-442.

Upton N (1994) Mechanisms of action of new antiepileptic drugs: rational design and serendipitous findings. Trends Pharmacol Sci 15:456-463.
Vajda FJ (2007) Pharmacotherapy of epilepsy: new armamentarium, new issues. J Clin Neurosci 14:813-823.

van der Weide J, Steijns LS, van Weelden MJ, de Haan K (2001) [Maintenance dose requirement for phenytoin is lowered in genetically impaired drug metabolism independent of concommitant use of other antiepileptics]. Ned Tijdschr Geneeskd 145:312-315.

Vanneschi L, Farinaccio A, Mauri G, Antoniotti M, Provero P, Giacobini M (2011) A comparison of machine learning techniques for survival prediction in breast cancer. BioData Min 4:12.

Vermeulen J, Aldenkamp AP (1995) Cognitive side-effects of chronic antiepileptic drug treatment: a review of 25 years of research. Epilepsy research 22:65-95.

Vinken PJ, Bruyn GW, Meinardi H (1999) The epilepsies. Amsterdam ; Oxford: Elsevier.

Volk HA, Arabadzisz D, Fritschy JM, Brandt C, Bethmann K, Loscher W (2006) Antiepileptic drug-resistant rats differ from drug-responsive rats in hippocampal neurodegeneration and GABA(A) receptor ligand binding in a model of temporal lobe epilepsy. Neurobiol Dis 21:633-646.

Waldegger S, Fakler B, Bleich M, Barth P, Hopf A, Schulte U, Busch AE, Aller SG, Forrest JN, Jr., Greger R, Lang F (1999) Molecular and functional characterization of s-KCNQ1 potassium channel from rectal gland of Squalus acanthias. Pflugers Arch 437:298-304.

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4:587-597.

Wallace RH, Marini C, Petrou S, Harkin LA, Bowser DN, Panchal RG, Williams DA, Sutherland GR, Mulley JC, Scheffer IE, Berkovic SF (2001a) Mutant GABA(A) receptor gamma2-subunit in childhood absence epilepsy and febrile seizures. Nature genetics 28:49-52.

Wallace RH, Scheffer IE, Barnett S, Richards M, Dibbens L, Desai RR, Lerman-Sagie T, Lev D, Mazarib A, Brand N, Ben-Zeev B, Goikhman I, Singh R, Kremmidiotis G, Gardner A, Sutherland GR, George AL, Jr., Mulley JC, Berkovic SF (2001b) Neuronal sodium-channel alpha1-subunit mutations in generalized epilepsy with febrile seizures plus. American journal of human genetics 68:859-865.

Wang L (2010) Pharmacogenomics: a systems approach. Wiley Interdiscip Rev Syst Biol Med 2:3-22.

Wang L, McLeod HL, Weinshilboum RM (2011) Genomics and drug response. N Engl J Med 364:1144-1153.

Wang SJ (2007) Biomarker as a classifier in pharmacogenomics clinical trials: a tribute to 30th anniversary of PSI. Pharm Stat 6:283-296.

Wang SP, Mintzer S, Skidmore CT, Zhan T, Stuckert E, Nei M, Sperling MR (2013) Seizure recurrence and remission after switching antiepileptic drugs. Epilepsia 54:187-193.

Weber WW (2001) The legacy of pharmacogenetics and potential applications. Mutat Res 479:1-18.

Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS, Samani NJ, Shields B, Prokopenko I, Farrall M, Dominiczak A, Johnson T, Bergmann S, Beckmann JS, Vollenweider P, Waterworth DM, Mooser V, Palmer CN, Morris AD, Ouwehand WH, Zhao JH, Li S, Loos RJ, Barroso I, Deloukas P, Sandhu MS, Wheeler E, Soranzo N, Inouye M, Wareham NJ, Caulfield M, Munroe PB, Hattersley AT, McCarthy MI, Frayling TM (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 40:575-583.

Weinshilboum R (2003) Inheritance and drug response. N Engl J Med 348:529-537.
Weinshilboum RM, Wang L (2006) Pharmacogenetics and pharmacogenomics: development, science, and translation. Annual review of genomics and human genetics 7:223-245.

Weller AE, Dahl JP, Lohoff FW, Kampman KM, Oslin DW, Dackis C, Ferraro TN, O'Brien CP, Berrettini WH (2006) No association between polymorphisms in the prostate apoptosis factor-4 gene and cocaine dependence. Psychiatric genetics 16:193-196.

West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Jr., Marks JR, Nevins JR (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A 98:11462-11467.

Whalen R, Boyer TD (1998) Human glutathione S-transferases. Semin Liver Dis 18:345-358.

White HS (1999) Comparative anticonvulsant and mechanistic profile of the established and newer antiepileptic drugs. Epilepsia 40 Suppl 5:S2-10.

White HS, Smith MD, Wilcox KS (2007) Mechanisms of action of antiepileptic drugs. Int Rev Neurobiol 81:85-110.

Wilke RA, Reif DM, Moore JH (2005) Combinatorial Pharmacogenetics. Nat Rev Drug Discov 4:911-918.

Wilkinson GR (2005) Drug metabolism and variability among patients in drug response. N Engl J Med 352:2211-2221.

Wolf CR, Smith G, Smith RL (2000) Science, medicine, and the future: Pharmacogenetics. Bmj 320:987-990.

Wu K, Reynolds NJ (2012) Pharmacogenetic screening to prevent carbamazepine-induced toxic epidermal necrolysis and Stevens-Johnson syndrome: a critical appraisal. Br J Dermatol 166:7-11; discussion 11-14.

Wu P, Jiang L, Chen H (2010) Sodium valproate at the therapeutic concentration inhibits the induction but not the maintenance phase of long-term potentiation in rat hippocampal CA1 area. Biochem Biophys Res Commun 391:582-586.

Yamaori S, Yamazaki H, Iwano S, Kiyotani K, Matsumura K, Honda G, Nakagawa K, Ishizaki T, Kamataki T (2004) CYP3A5 Contributes significantly to CYP3A-mediated drug oxidations in liver microsomes from Japanese subjects. Drug Metab Pharmacokinet 19:120-129.

Yang J, Wray NR, Visscher PM (2010) Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. Genet Epidemiol 34:254-257.

Yarov-Yarovoy V, DeCaen PG, Westenbroek RE, Pan CY, Scheuer T, Baker D, Catterall WA (2012) Structural basis for gating charge movement in the voltage sensor of a sodium channel. Proc Natl Acad Sci U S A 109:E93-102.

Yi HG, Kim HJ, Kim YJ, Han SW, Oh DY, Lee SH, Kim DW, Im SA, Kim TY, Kim CS, Heo DS, Bang YJ (2009) Epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (TKIs) are effective for leptomeningeal metastasis from non-small cell lung cancer patients with sensitive EGFR mutation or other predictive factors of good response for EGFR TKI. Lung Cancer 65:80-84.

Yip VL, Marson AG, Jorgensen AL, Pirmohamed M, Alfirevic A (2012) HLA Genotype and Carbamazepine-Induced Cutaneous Adverse Drug Reactions: A Systematic Review. Clin Pharmacol Ther 92:757-765.

Yoon Y, Song J, Hong SH, Kim JQ (2003) Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines. Clin Chem Lab Med 41:529-534.

Yu F, Catterall W (2003) Overview of the voltage-gated sodium channel family. Genome Biology 4:207.

Yu FH, Mantegazza M, Westenbroek RE, Robbins CA, Kalume F, Burton KA, Spain WJ, McKnight GS, Scheuer T, Catterall WA (2006) Reduced sodium current in GABAergic interneurons in a mouse model of severe myoclonic epilepsy in infancy. Nat Neurosci 9:1142-1149.

Yu FH, Westenbroek RE, Silos-Santiago I, McCormick KA, Lawson D, Ge P, Ferriera H, Lilly J, DiStefano PS, Catterall WA, Scheuer T, Curtis R (2003) Sodium channel beta4, a new disulfide-linked auxiliary subunit with similarity to beta2. J Neurosci 23:7577-7585.

Yu R, Shete S (2005) Analysis of alcoholism data using support vector machines. BMC genetics 6 Suppl 1:S136.

Yuan HY, Chiou JJ, Tseng WH, Liu CH, Liu CK, Lin YJ, Wang HH, Yao A, Chen YT, Hsu CN (2006) FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Res 34:W635-641.

Yukawa E, Mamiya K (2006) Effect of CYP2C19 genetic polymorphism on pharmacokinetics of phenytoin and phenobarbital in Japanese epileptic patients using Non-linear Mixed Effects Model approach. J Clin Pharm Ther 31:275-282.

Zaccara G, Franciotta D, Perucca E (2007) Idiosyncratic adverse reactions to antiepileptic drugs. Epilepsia 48:1223-1244.

Zanger UM (2010) Pharmacogenetics - challenges and opportunities ahead. Front Pharmacol 1:112.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PIW, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding C-J, Doney ASF, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jorgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CNA, Payne F, Perry JRB, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjogren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40:638-645.

Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. J Genet Genomics 38:95-109.
Zhang K, Calabrese P, Nordborg M, Sun F (2002) Haplotype block structure and its applications to association studies: power and study designs. American journal of human genetics 71:1386-1394.

Zhang K, Sun F (2005) Assessing the power of tag SNPs in the mapping of quantitative trait loci (QTL) with extremal and random samples. BMC Genetics 6:51.

Zhang Y-Q, Rajapakse JC (2009) Machine learning in bioinformatics. Hoboken, N.J.: Wiley.

Zhao H, Pfeiffer R, Gail MH (2003) Haplotype analysis in population genetics and association studies. Pharmacogenomics 4:171-178.

Zhao Y, Clark WT, Mort M, Cooper DN, Radivojac P, Mooney SD (2011) Prediction of functional regulatory SNPs in monogenic and complex disease. Hum Mutat 32:1183-1190.

Zhou BT, Zhou QH, Yin JY, Li GL, Qu J, Xu XJ, Liu D, Zhou HH, Liu ZQ (2012) Effects of SCN1A and GABA receptor genetic polymorphisms on carbamazepine tolerability and efficacy in Chinese patients with partial seizures: 2-year longitudinal clinical follow-up. CNS Neurosci Ther 18:566-572.

Zimprich F, Stogmann E, Bonelli S, Baumgartner C, Mueller JC, Meitinger T, Zimprich A, Strom TM (2008) A functional polymorphism in the SCN1A gene is not associated with carbamazepine dosages in Austrian patients with epilepsy. Epilepsia 49:1108-1109.

APPENDIX

CONTENTS

Appendix 1.1 Carbamazepine Drug Metabolising enzyme MALDI-TOF MS PCR and Extension Primers

SNP	Assay	Chromosomal position	Forward Amplification Primer Sequence (5'>3')	Reverse Amplification Primer Sequence (5'>3')	Extension Primer Sequence (5'>3')
rs35073925	1	226027630	ACG TTG GAT GAC AGC GTT TCG GGA GGT TTC	ACG TTG GAT GCT CTC CCT CAT CAG GCT GTA	CTC CAC ATC CCT CTC AG
rs11572081	1	96826966	ACG TTG GAT GGT GTT CAA GAG GAA GCT CAC	ACG TTG GAT GGT CAA TGA CGC AGA GTA GAG	GAG TAG AGT CAC CCA CC
rs35796837	1	75043593	ACG TTG GAT GGT TCA AGC ACA GCA AGA AGG	ACG TTG GAT GTG ACA ATC TTC TCC TGT GGG	CTG TGG GAT GAG GTT GC
rs11572126	1	96814915	ACG TTG GAT GTG TGC AAA AAT GGA AAA GCC	ACG TTG GAT GTG GAA ATT GAG TCC TCT CCC	GGT CCT CTC CCT GTA GTT
rs10799326	1	226009918	ACG TTG GAT GCC TCT GAG CTC AGT ATC TTG	ACG TTG GAT GGT TGC AAA CCA GCA TGA TTT	TAA ACG TGA CTG GAA GAT
rs1058930	1	96818119	ACG TTG GAT GGC TAA TAT CTT ACC TGC TCC	ACG TTG GAT GAA GAA CAC CAA GCA TCA CTG	ACA ATC CTC GGG ACT TTA T
rs10915884	1	226023875	ACG TTG GAT GTT CTG TTC CAG GAT CCC ATC	ACG TTG GAT GAA CTG TCA CAG CCA AGA AGG	AGG GTC TAA AGA GAC ATG A
rs4292394	1	69972949	ACG TTG GAT GAG AGT CTT ACC TAG AAG GTC	ACG TTG GAT GTT CTG TGG AGA TTT GAT GGG	TTA GGT CTC AAT ACT CGG CT
rs2292566	1	226019653	ACG TTG GAT GTG ACA TAC ATC CCT CTC TGG	ACG TTG GAT GCA GGT GGA GAT TCT CAA CAG	CCA CCC TCA CTT CAA GAC TAA
rs1536430	1	96817776	ACG TTG GAT GGG CAT ACA GGA AGC CCA TTT	ACG TTG GAT GAA TAT CCT ACC ACA AAC TG	CAC TAC CAC AAA CTG AAG ATG
rs10264272	1	99262835	ACG TTG GAT GGC CCA CAT ACT TAT TGA GAG	ACG TTG GAT GTC AAC AAT CCA CAA GAC CCC	TCT CAC CCT TTG TGG AGA GCA CTA A
rs11572079	1	96827118	ACG TTG GAT GAA GGT TGT GAG GGA GAA ACG	ACG TTG GAT GAA TTC TCC CAG TTT CTG CCC	CCC CCA GTT TCT GCC CCT TTT TTT TA

SNP	Assay	Chromosomal position	Forward Amplification Primer Sequence (5'>3')	Reverse Amplification Primer Sequence (5'>3')	Extension Primer Sequence (5'>3')
rs17861157	3	75043592	ACG TTG GAT GTG ACA ATC TTC TCC TGT GGG	ACG TTG GAT GGT TCA AGC ACA GCA AGA AGG	GGG GCG AAG GGG CCT AGA GCC AG
rs11572082	3	96826922	ACG TTG GAT GGA CTC TAC TCT GCG TCA TTG	ACG TTG GAT GGC CAC CCC TGA AAT GTT TCC	CCC CCA GGA CGT CAC TAG TGA AGA
rs4987161	3	99366081	ACG TTG GAT GGG GTC TTG TGG ATT GTT GAG	ACG TTG GAT GGC ATG GAT GTG ATC ACT AGC	CCC CCT GTG ATC ACT AGC ACA TCA T
rs2671272	3	226015116	ACG TTG GAT GGC CCA GCA TTG TTA TCT AGC	ACG TTG GAT GTG CAG GTT ACT CTG AAC AAG	CGG TTA CTC TGA ACA AGA ACA GTC T
rs2234922	3	226026406	ACG TTG GAT GAC TTC ATC CAC GTG AAG CCC	ACG TTG GAT GAA AAC TCG TAG AAA GAG CCG	GAA AGT CAG CAA GGG CTT CGG GGT A
rs11572127	3	96814689	ACG TTG GAT GTA GGG TAC ATG TGC ACA ATG	ACG TTG GAT GAT AAA TGG CAA ACC ATG TC	AAA TGG CAA ACC ATG TCA TTT TAA AG
rs35407132	3	75042301	ACG TTG GAT GAC CTG GCA CTG TCA AGG ATG	ACG TTG GAT GTG GAG CCA ATG CGG ATC TG	AGG GAG CCA ATG CGG ATC TGC AGG AC
rs28365062	3	69964271	ACG TTG GAT GCC AGG AGT TTC GAA TAA GCC	ACG TTG GAT GCT ATT CCT GTC AGG AAG ACC	TCT ACT CCT GTC AGG AAG ACC CAC TAC
rs2234700	4	226032896	ACG TTG GAT GGA ACC TCA CCC ACT TTT CAG	ACG TTG GAT GCA GGA TGA AGG TCT ATG TGC	CCT TCC CTT TTG AGC TA
rs3738042	4	226013388	ACG TTG GAT GAC TGC CTT GAC CCA CAG TGC	ACG TTG GAT GGT GCA TAA AAT ATT GGT GGA G	TAT TGG TGG AGC TCT TC
rs1051740	4	226019633	ACG TTG GAT GCT GGC GTT TTG CAA ACA TAC	ACG TTG GAT GAC TGG AAG AAG CAG GTG GAG	GTG GAG ATT CTC AAC AGA
rs11572103	4	96818106	ACG TTG GAT GAA GAA CAC CAA GCA TCA CTG	ACG TTG GAT GGC TAA TAT CTT ACC TGC TCC	CTT ACC TGC TCC ATT TTG A
rs4653695	4	226033083	ACG TTG GAT GAC ATC CGC AAG TTC CTG TC	ACG TTG GAT GCC AAG AAA AGC CTG GAG GG	GGA GCC TGG AGG GCA CTT G
rs28365095	4	99277605	ACG TTG GAT GTT TCA GCA GCT TGG CTG AAG	ACG TTG GAT GTA GCT GAG TGC TGC TGT TTG	GGC TGT TTG CCT GGA GCT TC

SNP	Assay	Chromosomal position	Forward Amplification Primer Sequence (5'>3')	Reverse Amplification Primer Sequence (5'>3')	Extension Primer Sequence (5'>3')
rs2234698	4	226019500	ACG TTG GAT GTT TGC TCC AGG ACT TAC ACC	ACG TTG GAT GTG AAG CCA TAG TGG AAG CAG	GGT GGG GTG AAA CGG AAC TT
rs762551	4	75041917	ACG TTG GAT GTC TGT GAT GCT CAA AGG GTG	ACG TTG GAT GCA GCT GGA TAC CAG AAA GAC	CTC AAT CTA CCA TGC GTC CTG
rs1934980	4	96808973	ACG TTG GAT GAA CTG ATG TCT TTG CTT GGG	ACG TTG GAT GTA CAA ATG GGA GAG TGG AGC	CCT CGA GTG GAG CAA GAT GAC
rs6600894	4	69983092	ACG TTG GAT GCA TCC ATT TTC ACA ATA GCT G	ACG TTG GAT GGT ATT TTT CTT TGT AGA GAC C	CTT TGT AGA GAC CTT TCA CAT T
rs7439366	4	69964338	ACG TTG GAT GGC TGA CGT ATG GCT TAT TCG	ACG TTG GAT GTG GAG TCC TCC AAC AAA ATC	TCA ACA TTT GGT AAG AGT GGA T
rs776746	4	99270539	ACG TTG GAT GGT AAT GTG GTC CAA ACA GGG	ACG TTG GAT GAC CCA GCT TAA CGA ATG CTC	TTC CAG AGC TCT TTT GTC TTT CA
rs2069522	4	75039233	ACG TTG GAT GTT CTC CCA TTC ATG GCC TTC	ACG TTG GAT GTC AGC AGA GCT TAG CCT ATC	GGT GTC CTA TCT GCA TGG CTG CC
rs1934952	4	96797500	ACG TTG GAT GCC AAG CCT GAT ATT CCA TGA	ACG TTG GAT GGA TGA AGA GAG TGT ATG ACC	GGA GCG TGT ATG ACC AGA GCT GA
rs3753663	4	226035289	ACG TTG GAT GTT AGA ACG CTG CCC TGG GAC	ACG TTG GAT GAG CCT GGG ATT GGG AGG AAA	AGA CGC AAA ATG AGA CTC ACA CAG
rs45468096	4	75043539	ACG TTG GAT GTC TTG CTG TGC TTG AAC AGG	ACG TTG GAT GTT TGA CCT TGG AAG TGC CAG	CCT TGT GCC CCC TCA GAA CAG TGT C
rs2292568	4	226027659	ACG TTG GAT GAT GTG CAT GTA GCC GCT CTC	ACG TTG GAT GTG AGA GGG ATG TGG AGC TG	CGA GGG ATG TGG AGC TGC TGT ACC C
rs11572172	4	96797752	ACG TTG GAT GGC ACA GAT TAC CAG GAA TCG	ACG TTG GAT GGA CAG AGA CCT TCC TTC AAG	GCA CAT TTT ACC ACA ATA GAT AAA TA
rs34143170	4	226027548	ACG TTG GAT GTG CCT TCA GCC ACG TGA AAG	ACG TTG GAT GGG GTC AGG GTA GAG AAG TTG	GGG TGT AAA ACC AAA GCC ATG TTC AA

SNP	Assay	Chromosomal position	Forward Amplification Primer Sequence (5'>3')	Reverse Amplification Primer Sequence (5'>3')	Extension Primer Sequence (5'>3')
rs2069524	4	75040276	ACG TTG GAT GTA GAG ACG GAG TTT CAC CAG	ACG TTG GAT GAA TCC CAG CAC TTT GAG AGG	GGG AGC AGC ACT TTG AGA GGC CGA GA
rs4646450	5	99266318	ACG TTG GAT GTA ACA AAG AGC GAG AGG ACG	ACG TTG GAT GGC CTT GTC CAG AAT ACA CAC	ATT CAC TTC ACG TGG CA
rs3738040	5	226013041	ACG TTG GAT GCT GTG CAA TTG TCA GAA GGC	ACG TTG GAT GTC TAA GGG CCT GTG AGA GAG	CTG TGA GAG AGG CAG GG
rs35694136	5	75039613	ACG TTG GAT GGA TTG TTT GAG CTC AGG AGG	ACG TTG GAT GAC AGA GTC TTG CTC TGT CAC	TCA CCC AGG TTG GGG TTC
rs1058932	5	96796861	ACG TTG GAT GCT GAA GAA TGC TAG CCC ATC	ACG TTG GAT GTA ATA GTG GGA ATG TCC TTG	TTG CAG GTG ATA GCA GAT C
rs45550332	5	226032979	ACG TTG GAT GTA TTC CTA CAT GGT TCG TGG	ACG TTG GAT GAG GAA CTT GCG GAT GTC CTG	AGC TCC GGC TCC TCA AAG GC
rs7435335	5	69971335	ACG TTG GAT GCA GTT AAC CAA ATT CAG CAA G	ACG TTG GAT GAT GAA GAA TCT GTT GGT GTC	TTG GTG TCA TGA ATA AAA ACA

SNP = single nucleotide polymorphism, PCR = polymerase change reaction, MALDI-TOF MS = matrix-assisted laser desorption/ ionisation-time of flight mass spectrometer, Chromosomal positions are provided from HapMap Data release 23, March 2008, NCBI B36, dbSNP b126.

Sequenom MALTI-TOF was performed with 5 different multi-plex assays consisting of 23, 21, 21, 20, 6 SNPs respectively.

Appendix 1.2 Details and allele frequencies of 91 tagging SNPs from carbamazepine Drug Metabolising Enzymes genotyped in 159 epilepsy patients

Gene	SNP	Chromosomal position	SNP Alleles	SNP location and function	HWE *P*-Value	% Call rate	MAF (this study)	MAF (public database)	Ref.
EPHX1	rs35073925	226027630	A>G	Exon 5	1.00	99	0.00	0.01	NCBI B36
CYP2C8	rs11572081	96826966	G>A	Exon 7	0.00	96	0.02	0.06	NCBI B36
CYP1A2	rs35796837	75043593	G>A	Exon 2	1.00	91	0.00	0.01	NCBI B36
CYP2C8	rs11572126	96814915	G>A	Intron 4	1.00	99	0.10	0.11	NCBI B36
EPHX1	rs10799326	226009918	T>C	Intron	1.00	84	0.12	0.12	NCBI B36
CYP2C8	rs1058930	96818119	C>G	Exon 5	1.00	98	0.04	0.03	NCBI B36
EPHX1	rs10915884	226023875	C>T	Intron 3	0.37	80	0.18	0.19	NCBI B36
UGT2B7	rs4292394	69972949	G>C	Exon 4	0.75	99	0.40	0.35	NCBI B36
EPHX1	rs2292566	226019653	G>A	Exon 2	1.00	99	0.15	0.13	NCBI B36
CYP2C8	rs1536430	96817776	C>T	Intron 4	1.00	98	0.02	0.02	NCBI B36
CYP3A4	rs12721617	99359911	A>C	Intron 3	1.00	97	0.01	0.01	NCBI B36
CYP2C8	rs2275622	96827178	C>T	Intron 7	0.79	97	0.38	0.32	NCBI B36
CYP2C8	rs2275620	96802598	A>T	Intron 2	0.55	100	0.40	0.39	NCBI B36
EPHX1	rs2740168	226020988	G>A	Intron 3	0.08	96	0.39	0.37	NCBI B36
CYP2C8	rs2071426	96828323	A>G	Intron 8	0.89	96	0.32	0.29	NCBI B36
UGT2B7	rs10028494	69970937	A>C	Intron 4	0.40	83	0.23	0.11	NCBI B36
CYP1A2	rs34067076	75042389	G>A	Exon 2	1.00	97	0.00	0.02	NCBI B36
CYP3A4	rs2246709	99365719	A>G	Intron 6	0.43	99	0.27	0.35	NCBI B36
CYP3A5	rs10264272	99262835	C>T	Exon 7	1.00	100	0.00	0.04	NCBI B36
CYP2C8	rs11572079	96827118	T>C	Intron 7	1.00	99	0.00	0.03	NCBI B36
TMEM63A	rs2292558	226037318	G>C	Intron 3	0.96	98	0.10	0.14	NCBI B36
EPHX1	rs2260863	226019774	C>G	Intron 3	0.45	98	0.26	0.28	NCBI B36

Gene	SNP	Chromosomal position	SNP Alleles	SNP location and function	HWE *P*-Value	% Call rate	MAF (this study)	MAF (public database)	Ref.
CYP2C8	rs1341159	96815619	C>G	Intron 4	1.00	98	0.00	0.27	NCBI B36
CYP3A4	rs4646440	99360870	C>T	Intron 3	1.00	98	0.00	0.01	NCBI B36
CYP3A5	rs28371764	99277593	C>T	5' UTR	1.00	99	0.03	0.07	NCBI B36
UGT2B7	rs4348159	69972952	C>T	Intron 3	0.01	86	0.06	0.06	NCBI B36
CYP3A5	rs6976017	99249999	G>A	Intron 2	1.00	99	0.04	0.04	NCBI B36
EPHX1	rs6965	226033476	T>C	3' near gene	0.00	91	0.49	0.36	NCBI B36
CYP1A2	rs11636419	75047600	A>G	3' UTR	1.00	86	0.00	0.03	NCBI B36
EPHX1	rs3753660	226012776	T>C	Intron 1	0.91	98	0.13	0.13	NCBI B36
CYP3A5	rs1419745	99260092	A>G	Intron 4	1.00	99	0.04	0.02	NCBI B36
CYP1A2	rs2069525	75040372	T>A/C/G	5' near gene	1.00	99	0.03	0.08	NCBI B36
EPHX1	rs3753658	226012686	G>T	Intron 1	0.51	86	0.18	0.20	NCBI B36
UGT2B7	rs3924194	69971092	C>G	Intron 3	0.26	99	0.17	0.17	NCBI B36
CYP2C8	rs1934956	96828160	C>T	Intron 7	1.00	98	0.12	0.16	NCBI B36
CYP3A5	rs28365083	99250236	C>A	Exon 3	1.00	98	0.01	0.02	NCBI B36
UGT2B7	rs10050146	69971576	C>T	5'near gene	0.17	99	0.03	0.03	NCBI B36
TMEM63A	rs360063	226036309	G>A	Intron 4	-	7	-	0.40	NCBI B36
CYP1A2	rs2470890	75047426	T>C	Exon 6	0.54	96	0.35	0.36	NCBI B36
CYP3A5	rs15524	99245914	T>C	3' UTR	1.00	92	0.09	0.07	NCBI B36
CYP3A4	rs2242480	99361466	C>T	Intron 3	0.56	92	0.09	0.17	NCBI B36
CYP2C8	rs11188150	96802737	C>T	Exon 3	1.00	96	0.00	0.01	NCBI B36
EPHX1	rs1877724	226013355	C>T	Intron 1	0.25	97	0.28	0.19	NCBI B36
CYP3A4	rs12333983	99354114	T>A	3' near gene	1.00	99	0.11	0.12	NCBI B36

Gene	SNP	Chromosomal position	SNP Alleles	SNP location and function	HWE *P*-Value	% Call rate	MAF (this study)	MAF (public database)	Ref.
CYP3A4	rs1851426	99382936	C>T	5' near gene	1.00	96	0.04	0.03	NCBI B36
EPHX1	rs4149229	226032928	G>A	Exon 7	1.00	98	0.01	0.04	NCBI B36
CYP3A4	rs4646437	99365083	C>T	Intron 5	0.84	98	0.11	0.13	NCBI B36
CYP2C8	rs11572080	96827030	G>A	Exon 7	0.61	98	0.15	0.05	NCBI B36
CYP3A4	rs4986910	99358524	C>T	Exon 2	1.00	97	0.00	0.01	NCBI B36
CYP1A2	rs17861162	75048753	C>G	3' UTR	1.00	98	0.00	0.18	NCBI B36
EPHX1	rs4149230	226033030	G>C	Exon 7	1.00	97	0.03	0.02	NCBI B36
EPHX1	rs2740170	226024797	C>T	Intron 3	0.82	96	0.20	0.23	NCBI B36
UGT2B7	rs7375178	69969679	C>A	Exon 2	0.42	98	0.39	0.48	NCBI B36
EPHX1	rs2854461	226011644	C>A	Intron 1	0.90	96	0.35	0.38	NCBI B36
EPHX1	rs35561387	226027569	A>G	Exon 5	1.00	96	0.00	0.03	NCBI B36
CYP3A4	rs2740574	99382096	A>G	5' near gene	1.00	98	0.04	0.03	NCBI B36
UGT2B7	rs4356975	69972463	C>T	Intron 3	0.81	95	0.30	0.31	NCBI B36
CYP1A2	rs17861157	75043592	C>A	Exon 2	1.00	95	0.01	0.04	NCBI B36
CYP2C8	rs11572082	96826922	G>C	Intron 6	0.66	98	0.14	0.12	NCBI B36
CYP3A4	rs4987161	99366081	T>C	Exon 7	1.00	97	0.00	0.02	NCBI B36
EPHX1	rs2671272	226015116	C>T	Intron 1	0.91	98	0.21	0.23	NCBI B36
EPHX1	rs2234922	226026406	A>G	Exon 3	0.82	90	0.16	0.18	NCBI B36
CYP2C8	rs11572127	96814689	G>C	Intron 4	-	0	-	0.07	NCBI B36
CYP1A2	rs35407132	75042301	C>T	Exon 1	1.00	98	0.00	0.03	NCBI B36
UGT2B7	rs28365062	69964271	A>G	Exon 3	0.90	96	0.11	0.18	NCBI B36
EPHX1	rs2234700	226032896	T>C	Intron 7	1.00	98	0.00	0.03	NCBI B36

Gene	SNP	Chromosomal position	SNP Alleles	SNP location and function	HWE *P*-Value	% Call rate	MAF (this study)	MAF (public database)	Ref.
EPHX1	rs3738042	226013388	G>A	Intron 1	-	7	-	0.28	NCBI B36
EPHX1	rs1051740	226019633	T>C	Exon 2	0.59	97	0.30	0.33	NCBI B36
CYP2C8	rs11572103	96818106	A>T	Intron 4	0.01	98	0.01	0.03	NCBI B36
EPHX1	rs4653695	226033083	A>C	3'UTR	0.68	82	0.16	0.15	NCBI B36
CYP3A5	rs28365095	99277605	G>A	5' UTR	1.00	98	0.01	0.02	NCBI B36
EPHX1	rs2234698	226019500	T>C	Exon 2	1.00	98	0.04	0.03	NCBI B36
CYP1A2	rs762551	75041917	A>C	Intron 1	0.70	97	0.28	0.31	NCBI B36
CYP2C8	rs1934980	96808973	T>C	Intron 4	0.95	98	0.13	0.19	NCBI B36
UGT2B7	rs6600894	69983092	G>A	3' near gene	0.38	97	0.16	0.22	NCBI B36
UGT2B7	rs7439366	69964338	C:T	Intron 3	0.55	98	0.40	0.50	NCBI B36
CYP3A5	rs776746	99270539	G>A	Intron 10	0.00	97	0.50	0.06	NCBI B36
CYP1A2	rs2069522	75039233	T>C	5' near gene	-	49	-	0.08	NCBI B36
CYP2C8	rs1934952	96797500	G>A	Intron 1	0.78	98	0.35	0.37	NCBI B36
TMEM63A	rs3753663	226035289	T>A	Intron 3	-	19	-	0.03	NCBI B36
CYP1A2	rs45468096	75043539	C>T	Intron 2	-	2	-	0.02	NCBI B36
EPHX1	rs2292568	226027659	G>C	Exon 5	1.00	97	0.03	0.03	NCBI B36
CYP2C8	rs11572172	96797752	A>C	Intron 1	1.00	95	0.03	0.06	NCBI B36
EPHX1	rs34143170	226027548	C>T	Exon 5	1.00	98	0.06	0.08	NCBI B36
CYP1A2	rs2069524	75040276	A>G	5' near gene	-	0	-	0.08	NCBI B36
CYP3A5	rs4646450	99266318	G>A	Intron 9	0.47	99	0.18	0.18	NCBI B36
EPHX1	rs3738040	226013041	G>A	Intron 1	0.91	98	0.07	0.12	NCBI B36
CYP1A2	rs35694136	75039613	T>N	5' near gene	-	0	-	0.24	NCBI B36
CYP2C8	rs1058932	96796861	C>T	3' UTR	0.95	99	0.13	0.19	NCBI B36
EPHX1	rs45550332	226032979	G>A	Exon 7	1.00	100	0.00	0.01	NCBI B36
UGT2B7	rs7435335	69971335	G>A	Exon 7	0.98	100	0.10	0.18	NCBI B36

SNP = single nucleotide polymorphism, PCR = polymerase change reaction, SNP frequency data was compiled from HapMap and NCBI dbSNP databases. Chromosomal positions are given in base pairs as per NCBI B36 assembly, dbSNP b126.

Appendix 1.3 SNPs captured by the 91 tagging SNPs across carbamazepine Drug Metabolising Enzymes

Tagging SNP	Tagged SNPs	Gene	Chromosome location
rs2071426	rs2071426	*CYP2C8*	96828323
	rs1934982	*CYP2C8*	96802124
	rs6583967	*CYP2C8*	96814475
	rs1934957	*CYP2C8*	96815114
	rs11572139	*CYP2C8*	96808886
	rs1341164	*CYP2C8*	96800873
	rs2185571	*CYP2C8*	96824975
	rs1934983	*CYP2C8*	96801929
	rs11572093	*CYP2C8*	96824406
rs1341159	rs1341159	*CYP2C8*	96815619
	rs11572082	*CYP2C8*	11572082
	rs1934951	*CYP2C8*	1934951
	rs2275622	*CYP2C8*	2275622
	rs2275620	*CYP2C8*	2275620
	rs1934952	*CYP2C8*	96797500
	rs1536430	*CYP2C8*	96817776
	rs11572126	*CYP2C8*	96814915
	rs11572127	*CYP2C8*	96814689
	rs11572079	*CYP2C8*	96827118
	rs11572172	*CYP2C8*	96797752
	rs1934956	*CYP2C8*	96828160

Tagging SNP	Tagged SNPs	Gene	Chromosome location
rs11572082	rs11572082	*CYP2C8*	96826922
	rs11572150	*CYP2C8*	96807128
	rs11572174	*CYP2C8*	96797571
	rs10509681	*CYP2C8*	96798749
	rs11188153	*CYP2C8*	96805090
	rs11572169	*CYP2C8*	96799774
	rs11572107	*CYP2C8*	96817233
rs1934980	rs1934980	*CYP2C8*	96808973
	rs1934951	*CYP2C8*	96798548
	rs1058932	*CYP2C8*	96796861
	rs1934980	*CYP2C8*	96808973
	rs1341162	*CYP2C8*	96810612
	rs1113129	*CYP2C8*	96811045
	rs10882520	*CYP2C8*	96799688
	rs11572101	*CYP2C8*	96818362
rs2275622	rs2275622	*CYP2C8*	96827178
	rs7095531	*CYP2C8*	96811841
	rs1891073	*CYP2C8*	96804911
	rs6583968	*CYP2C8*	96816357
	rs1934953	*CYP2C8*	96797470
	rs1934984	*CYP2C8*	96801805

Tagging SNP	Tagged SNPs	Gene	Chromosome location
rs2275620	rs2275620	*CYP2C8*	96802598
	rs1934985	*CYP2C8*	96801753
	rs1891071	*CYP2C8*	96805371
	rs7910936	*CYP2C8*	96804451
	rs1341163	*CYP2C8*	96810552
rs1934952	rs1934952	*CYP2C8*	96797500
	rs11572177	*CYP2C8*	96797270
rs1536430	rs1536430	*CYP2C8*	96817776
rs11572126	rs11572126	*CYP2C8*	96814915
rs11572127	rs11572127	*CYP2C8*	96814689
rs11572079	rs11572079	*CYP2C8*	96827118
rs11572172	rs11572172	*CYP2C8*	96797752
rs12333983	rs12333983	*CYP3A4*	99354114
	rs2404955	*CYP3A4*	99353279
rs1851426	rs1851426	*CYP3A4*	99382936
	rs2687105	*CYP3A4*	99376946
rs4646440	rs4646440	*CYP3A4*	99360870
rs4646437	rs4646437	*CYP3A4*	99365083
rs2242480	rs2242480	*CYP3A4*	99361466
rs12721617	rs12721617	*CYP3A4*	99359911
rs2246709	rs2246709	*CYP3A4*	99365719

Tagging SNP	Tagged SNPs	Gene	Chromosome location
rs11773597	rs11773597	*CYP3A4*	99382451
rs1419745	rs1419745	*CYP3A5*	99260092
	rs776741	*CYP3A5*	99279136
	rs4646447	*CYP3A5*	99268390
	rs4646453	*CYP3A5*	99260362
	rs4646449	*CYP3A5*	99266443
	rs4646456	*CYP3A5*	99245275
	rs4646458	*CYP3A5*	99245013
	rs4646446	*CYP3A5*	99275083
rs6976017	rs6977165	*CYP3A5*	99269397
	rs6976017	*CYP3A5*	99249999
	rs6956305	*CYP3A5*	99241310
rs15524	rs4646457	*CYP3A5*	99245080
	rs15524	*CYP3A5*	99245914
	rs776746	*CYP3A5*	99270539
rs4646450	rs4646450	*CYP3A5*	99266318
	rs3924192	*UGT2B7*	69970964
	rs6858558	*UGT2B7*	69969543
	rs7698645	*UGT2B7*	69971910
	rs4541594	*UGT2B7*	69972272
	rs6600884	*UGT2B7*	69968066

Tagging SNP	Tagged SNPs	Gene	Chromosome location
	rs7439152	*UGT2B7*	69969006
	rs6600891	*UGT2B7*	69971596
	rs12642938	*UGT2B7*	69976217
	rs12513195	*UGT2B7*	69972086
	rs7375178	*UGT2B7*	69969679
	rs4351080	*UGT2B7*	69972319
	rs7442453	*UGT2B7*	69969180
	rs6600893	*UGT2B7*	69978901
	rs4521414	*UGT2B7*	69973525
rs7375178	rs9995928	*UGT2B7*	69976663
	rs10050146	*UGT2B7*	69971576
rs10050146	rs10050146	*UGT2B7*	69971576
rs7435335	rs7435335	*UGT2B7*	69971335
rs3924194	rs3924194	*UGT2B7*	69971092
rs6600894	rs6600894	*UGT2B7*	69983092
rs4356975	rs4356975	*UGT2B7*	69972463
rs4348159	rs4348159	*UGT2B7*	69972952
rs10028494	rs10028494	*UGT2B7*	69970937
rs2470890	rs11854147	*CYP1A2*	75052771
	rs2470890	*CYP1A2*	75047426

Tagging SNP	Tagged SNPs	Gene	Chromosome location
rs10799326	rs10799326	*EPHX1*	226009918
	rs3738043	*EPHX1*	226015299
	rs10753410	*EPHX1*	226008101
	rs3766934	*EPHX1*	226015017
	rs3753661	*EPHX1*	226014342
rs2292558	rs2740174	*EPHX1*	226033969
	rs2671267	*EPHX1*	226025690
	rs2292558	*EPHX1*	226037318
	rs1051741	*EPHX1*	226032229
rs2671272	rs2854450	*EPHX1*	226012577
	rs2671272	*EPHX1*	226015116
rs2740170	rs2740171	*EPHX1*	226025528
	rs2740170	*EPHX1*	226024797
rs2292568	rs2292568	*EPHX1*	226027659
rs6965	rs6965	*EPHX1*	226033476
rs3753663	rs3753663	*EPHX1*	226035289
rs2234698	rs2234698	*EPHX1*	226019500
rs2260863	rs2260863	*EPHX1*	226019774
rs2292566	rs2292566	*EPHX1*	226019653

Tagging SNP	Tagged SNPs	Gene	Chromosome location
rs1051740	rs1051740	*EPHX1*	226019633
rs10915884	rs10915884	*EPHX1*	226023875
rs2234922	rs2234922	*EPHX1*	226026406
rs360063	rs360063	*EPHX1*	226036309
rs2740168	rs2740168	*EPHX1*	226020988
rs1877724	rs1877724	*EPHX1*	226013355
rs3753658	rs3753658	*EPHX1*	226012686
rs3738042	rs3738042	*EPHX1*	226013388

SNP = single nucleotide polymorphism

Appendix 1.4 Australian five-SNP classifier PCR and extension primers for MALTI-TOF MS

SNP	Assay	Chromosomal position	Forward Amplification Primer Sequence (5'>3')	Reverse Amplification Primer Sequence (5'>3')	Extension Primer Sequence (5'>3')
rs658624	1	118018767	ACGTTGGATGTAAGGT CTGGCTCATGACAC	ACGTTGGATGTAAGTCA TCCACATAGGTGC	CACAAACCAGGCAGAAA
rs2808526	1	101326687	ACGTTGGATGACTGCC TGTCACACAGTATC	ACGTTGGATGACAGGC CTAACTGGGACAAC	CTGCTCTTCAACCCCAAG
rs678262	1	118021740	ACGTTGGATGCCCAA AGGGTAGCTCAGAAA	ACGTTGGATGGACTGTT CAGCTGTATAGAC	AGCTGTATAGACCAGGTA
rs4869682	1	36656718	ACGTTGGATGACCAG GGCTGCAATGCAAAT	ACGTTGGATGGAGAAT CTGACTTGTCTAGC	ACTCCTTGAGAAGAGGA GC
rs2283170	1	2583141	ACGTTGGATGCCTCAG GAGGGACACAGAG	ACGTTGGATGATCCTTC TGCTCGGCTGCTT	CCCATGGAACGTGCAGCCCG

Chromosomal positions are provided from HapMap Data release 23, March 2008, NCBI B36, dbSNP b126, Sequenom MALTI-TOF was performed with 1 multi-plex assay.

Appendix 1.5 Custom and predesigned primer assays from TaqMan for five GWAS candidate SNP genotyping

SNP	TaqMan SNP Genotyping Assay	Chromosomal position	SNP	Forward Amplification Primer Sequence (5'>3')	Reverse Amplification Primer Sequence (5'>3')	Reporter 1 (VIC Sequence	Reporter 2 (FAM) Sequence
rs17252760	Custom	148653916	rs17252760	TGCCATCAGTTACCTT TAAAACTACATGT	GGATTCATTTGTCC TGTGAGAG AGAA	TCCCACAAAC CCC	CTCCCATAA ACCCC

SNP	TaqMan SNP Genotyping Assay	Chromosomal position	SNP	Context sequence
rs12919774	Pre-designed	8515708	rs12919774	AGGAGAAAATTTCCTCTACTCTGAG[A/G]TCAAGCCATTCTACCAAAAAATAAG
rs16994558	Pre-designed	147452128	rs16994558	AGGAGAAAATTTCCTCTACTCTGAG[A/G]TCAAGCCATTCTACCAAAAAATAAG
rs316132	Pre-designed	52847966	rs316132	GGCCCACTCTTATTTCCCAGTTCTG[C/T]TGCTAGAACATCAAGAGGTGTAGTC
rs622902	Pre-designed	52846474	rs622902	AGGTGGCAGGCCAGGTTTGGCCCAG[A/G]AGTTACAGTCTGCACATTAGACTTG

Chromosomal positions are provided from NCBI B36, dbSNP b126.
The context sequence refers to the nucleotide sequence surrounding the SNP site, where SNP alleles are in brackets and the order of the alleles corresponds to the association with reporter dyes, where Allele 1 = VIC and Allele 2 = FAM