ABSTRACT

MYERS, RACHEL ANNE. Population Genetic Methods for Detecting Genetic Contributions to Complex Traits. (Under the direction of Jeffrey Thorne and Eric Stone.)

Linking genomic variation to observed differences in phenotype is a major goal in both statistical and population genetics. These two fields have recently been split due to the requirement in population genetics for sequence data while statistical genetics utilizes genome-wide ascertained data. However, more advanced sequencing technologies are now bridging this gap between statistical and population genetics for genome-wide studies by providing faster and more affordable sequencing.

Our focus is to incorporate tests of natural selection from population genetics with genotype-phenotype associations from statistical genetics in order to better understand genomic variation in complex traits. In chapter two, we study natural selection and anti-malarial drug resistance associations in *Plasmodium falciparum*. In chapter three, we expand the study to include both a genome-wide genotyping dataset as well as a greater range of anti-malarial drugs. In chapter four, we explore the hypothesis that *de novo* mutations cause Autism Spectrum Disorder and Schizophrenia. Finally, in chapter five, we test the rare variant hypothesis using resequencing data from unaffected individuals and individuals diagnosed with Autism Spectrum Disorder or Schizophrenia. Through these studies, we demonstrate the importance of understanding how natural selection is linked with phenotype, as well as analyzing the resulting impact on allele frequency and genomic variation.

Population Genetic Methods for Detecting Genetic Contributions to Complex Traits

by
Rachel Anne Myers

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2010

APPROVED BY:

_____	_____
Jeffrey Thorne	Eric Stone
Chair of Advisory Committee	Co-Chair of Advisory Committee
_____	_____
Philip Awadalla	Ignazio Carbone

_____
Jung-Ying Tzeng

DEDICATION

In loving memory of my brother,

David John Myers,

whose short life inspired me to study medical genetics.

~

In loving memory of my grandfather,

John Smith Yates,

who taught me to question everything.

~

In honor of my parents,

Dr. Ronald T. and Linda Y. Myers,

who wholeheartedly supported me in every endeavor on which I have embarked.

BIOGRAPHY

I was born on November 7, 1982 in DuBois, Pennsylvania, to Ronald and Linda Myers. After the birth of my brother, David, we relocated to northwest Ohio, where we shared eight short years as a family of four. At the age of thirteen, I lost my brother to the genetic disorder called Adrenoleukodystropy (ALD). Witnessing and helping in his battle for life, I was inspired to pursue a career in medical research, hoping to one day improve someone else's battle for life.

In 2001, I became closer to my career goal by beginning my undergraduate study at The Pennsylvania State University in State College, PA. In 2005, I graduated with Bachelor of Science degree in Biochemistry and Molecular Biology, a minor in Chemistry, and many lessons learned along the way. During my undergrad, I held many lab assistant positions but the most influential was at The Plant Science Research Center at University of Toledo. Under the direction of Dr. Parani and Dr. Siram, I studied the effect of nitric oxide on gene expression in plants. During this experience, I quickly learned that wet lab work was not my strength; rather, my strength was in analyzing the results of the gene expression study. Drs. Parani and Siram noticed the affinity to data mining and sat down with me to discuss my future as a researcher; recommending I pursue a graduate program in bioinformatics.

Drawn to NSCU by its strengths in statistical and population genetics, I began my graduate career in 2005, studying bioinformatics. While taking a population genetics course, I discovered population genetics was what I wanted to study and I joined the

Awadalla lab in fall of 2006. During this experience, I further pursued my long time goal of medical research, studying malaria, autism spectrum disorder, and schizophrenia.

ACKNOWLEDGEMENTS

I thank my advisor, Philip Awadalla, for the invitation to join his lab and for embracing gchat and skype for our long distance communication. Philip, it has been a great joy to watch the lab expand and you grow as a PI. I thank my 'advisor pro-tempe', Eric Stone, for being available when I needed a face-to-face advisor chat or when I didn't understand Philip's marching orders. I thank Jeffrey Thorne for his thoughtful and critical questions, always ensuring we came to a solution, even if it required a few days for both of us to think. To Jung-Ying Tzeng, I am grateful for your time answering my statistical questions, and to Ignazio Carbone, I enjoyed our discussions and your perspective.

To the former and current Awadalla lab members, Kate, Jon, Julie, Jacki, Martine, Ferran, and Youssef, thank you for the many edits of my manuscripts, your assistance with my research, making my Montreal visits enjoyable, and the moral support I needed from time to time. I wish you all the very best in your future endeavors and hope our paths cross again. To Lisa, Jon, Ben, and Shengdar, while this journey has been rocky from time to time, it has been much easier experiencing it with you. To the remaining BRC students and faculty, thank you for lending me an ear when I had a quick question. To JB, Karen, Tina, and Dr. Zeng, thank you for taking care me administratively, without you I would be lost in a maze of paperwork.

To my Wednesday trivia team, your moral support over the last 5 years is priceless. Finally, thank you Dong, for your editorial review of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Complex Trait Models in the Age of Resequencing

Introduction

Understanding the genetic etiology of complex traits or diseases in natural populations of both humans and human pathogens is an important focus in medical and statistical genomics because understanding the genetic causes of complex disease leads to early diagnosis and personalized treatments. One of the main issues concerning this relationship is determining the link between causal allele frequencies and their effects on phenotypic variation. The Common Disease – Common Variant (CDCV) hypothesis proposes that this relationship can be best described by a limited set of common alleles individually contributing to a smaller portion of phenotypic variation in complex disease. In contrast, the Common Disease – Rare Variant (CDRV) hypothesis argues that this relationship is better modeled through a larger set of rare alleles each contributing to a larger portion of phenotypic variation in complex disease. Each hypothesis has relative strengths and weaknesses depending on the complex disease that they are trying to model. For example, diseases that do not affect reproductive fitness are better described by the CDCV while diseases that have a detrimental impact on reproduction are better modeled by the CDRV. The following sections will describe both of these hypotheses in detail, discussing their experimental designs, practical applications, and limitations. This chapter will conclude with an overview of natural selection and its role in both the CDCV and the CDRV hypotheses.

The Common Disease - Common Variant Hypothesis

Overview

The common disease – common variant (CDCV) hypothesis proposes that common or complex diseases having an inherited component are caused by alleles at moderate frequency (minor allele frequency > 0.05). This hypothesis was derived from observations of Mendelian diseases, where genetic markers showed a strong association with phenotype status. For example, the apoliopotien E type 4 allele (APOE-epsilon 4) is associated with increased risk of Alzheimer's Disease [1]. Expanding on these observations, Collins and Lander proposed creating a catalogue of common variants and using those variants to search for susceptibility genes by testing allele frequency differences in affected versus unaffected cohorts [2].

The human leukocyte antigen (HLA) system became the first genomic region to be thoroughly mapped and used to test for autoimmune disease associations [3]. As the catalogue of common variants grew with the HapMap project [4,5], coupled with technologies capable of parallelized genotyping of hundreds of thousands of variants in a single sample [6], systematic genome-wide association studies became a popular tool for studying human complex diseases and traits. Additionally, this strategy of cataloguing common variants and systematically testing associations with different traits and diseases were adapted for model organisms and human pathogens (e.g. *Plasmodium falciparum* [7,8]).

Experimental Design

CDCV is primarily tested using genome-wide association studies (GWAS). The two main types of GWAS are dictated by phenotype; discrete (case-control) and quantitative. A case-control design consists of a cohort of affected samples (cases) and a cohort of unaffected samples (controls). A quantitative trait design consists of a random collection of samples for which a continuous trait is measured (e.g. blood pressure or drug response). For both designs, the ideal sample selection would have a homogenous population; however, this is rarely the case. Typically samples are drawn from multiple populations, which creates population structure in theses study cohorts because allele frequency varies between populations. However, this population structure can result in false positive associations due to variance in allele frequencies within populations and between populations. To address such false positives, several methods for controlling population structure have been devised. For example, one method pairs cases and controls based on ethnicity or sampling location to standardize population representation in each cohort. Such a methodology can mitigate the effects of false positive associations since standardizing population representation also standardizes allele frequency variance between cohorts. Alternatively, evidence of population structure from either principal component analysis [9] or software programs such as structure [10] can be incorporated as cofactors in association tests. This correction is important since researchers have proposed using "common controls" [11], or a group of normal individuals that have been genotyped and available to other researchers, as controls for multiple GWAS.

The genotype data collected from the cohorts in GWAS are extracted utilizing microarray technologies like Affimetrix's SNPchip and Illumina's BeadChip. These tools are commonly used to genotype thousands to millions of previously ascertained single nucleotide polymorphisms (SNPs) in each sample. One of the primary criticisms of these aforementioned tools is that the collection of SNPs for these tools has been ascertained by resequencing a limited set of samples. For example, SNPs ascertained using the HapMap European samples could be detrimental to a study focusing on African-American populations. Also, testing every known common variant is beyond the scope of current array-based technologies. Due to this limitation, a smaller, informative set of SNPs that have strong correlations with their neighbors ($r^2 \geq 0.8$) can be used [12]. These SNPs, called tagSNPs, can capture most of the common variation by utilizing the linkage disequilibrium (LD) patterns observed in HapMap samples [13]. Quality control filters including Hardy Weinberg Equilibrium, call rate, minor allele frequency, and Mendelian error are then applied to the genotype calls to minimize error in the resulting genotype data.

Tests of genotype – phenotype association range from a χ^2 to likelihood ratio tests. The χ^2, extensions of χ^2 like the Armitage-Cochran trend test [14], and logistic regression are used to test in case-control studies. Linear regressions, correlation tests, and likelihood tests are used to evaluate genotype – phenotype associations for quantitative traits. Covariates, such as gender, population structure, and environment that may have an effect on phenotype can be incorporated as factors in both logistic and linear

regression-based testing. SNPs with *P* values meeting a cutoff (e.g. *P* value > 0.05) after multiple testing correction (e.g. Bonferroni) or a False Discovery Rate (FDR) [15] cutoff (e.g. 0.01) are considered candidate variants. Follow-up studies aim to either replicate these associations in a second and larger population, resequence the candidate region to find causal SNPs, or conduct functional studies in model organisms. While *P* values indicate the statistical significance of a SNP association with phenotype, the impact of the allele on the phenotype is summarized in one of three ways: an odds ratio (OR), an effect size, or a relative risk (RR).

Examples

GWAS has been widely used for studying complex human diseases, including breast cancer[16,17,18,19], autism [20,21], schizophrenia [22,23,24], bipolar disorder [24], coronary artery disease, types 1 and 2 diabetes, Crohn's disease [11], and human response to HIV [25] or malaria [26]. In 2007, the Wellcome Trust Case Control Consortium (WTCCC) reported findings for the one of the largest GWAS ever completed. Two thousand cases and three thousand shared controls were used to test genetic associations for seven common diseases – bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes [11]. They reported twenty-four association signals across six of the seven diseases, and these signals have been the focus of several follow-up studies. For example, WTCCC reported 7 loci associated with type 1 diabetes, and a follow up study using

independent samples of similar ethnicity confirmed 4 of the 7 loci [27]. This study-to-study variability in association results and limited replication is a common occurrence in human GWAS.

In addition to humans, GWAS has been used to map variants involved in plants to identify genetic determinates of pathogen response [28] and response to environmental stressors [29]. As efforts to catalogue common variants in human pathogens are completed (e.g. *Plasmodium falciparum* [7,8]), researchers aim to apply lessons learned from human GWAS to human pathogens.

Limitations

While the CDCV hypothesis has been somewhat successful, researchers have found both technical and theoretical limitations with using GWAS to study complex diseases. Technical limitations include sample size requirements, variation ascertainment, and the narrow focus of current genotyping technology, while the primary theoretical limitation is the assumption the phenotypes are selectively neutral.

The most limiting factor in GWAS is the sample size required to detect loci with modest effect sizes. Power studies have shown 3000-5000 samples are required to detect SNPs with effect sizes of 1.3 with 80% power [30]. Such large sample size requirements make GWAS a collaborative endeavor rather than a single-lab investigation. This reliance on collaboration and cooperation between multiple labs can create tangible hurdles such as lab-to-lab variation and greater potential for experimental error. Also, due to the

inverse exponential relationship between the required sample size and effect size [30], some larger studies may be under-powered and those associations found in larger studies maybe of strong effect rather than moderate effect.

Another technical limitation in GWAS is due to the genotyping technology; most microarrays are biased towards European ancestry populations and reduced coverage for other populations, particularly African populations. For example, genome coverage may range from 64-93% for the Caucasian (CEU) HapMap samples and 29-68% for the Yoruba (YRI) HapMap samples when $r^2 \geq 0.8$ between included SNPs and neighboring SNPs [31]. This bias arises from the populations selected for the ascertainment SNPs. Coverage for non-European populations will improve with ongoing efforts to catalogue common variation and LD structure in other populations.

The major theoretical limitation of the CDCV hypothesis is the assumption that phenotypes are selectively neutral. The basis of this assumption is that the complex diseases and traits studied tend to be late onset and/or do not affect reproduction. Additionally, due to the complex nature of a disease and its genetic components, selection is assumed to be negligible at causal sites. This neutrality assumption is often violated, for example, individuals affected with autism are less likely to reproduce. The resulting selection reduces the frequency of susceptibility alleles [32], which decreases power to detect causal variants.

The final issue with GWAS is that many studies have been completed, but few have successful follow-up studies. This suggests that results may be population specific,

spurious, or that the follow-up studies may be under powered. Some of these issues have convinced researchers to consider alternative genetic models such as CDRV for describing the genetic components of phenotype variability.

The Common Disease - Rare Variant Hypothesis

Overview

For several human traits, the CDCV hypothesis has explained a small proportion of genetic etiology. This came as a surprise when these traits (e.g. neurological disorders) were highly heritable. However, complex disease studies focused on small families showed rare inherited and/or *de novo* variants associated with disease [33,34]. These findings altered how researchers viewed the cause of genetic diseases, emphasizing the role of rare variation in common disease. The common disease – rare variant (CDRV) hypothesis proposes that the genetic etiology of complex disease is heterogeneous; many different variants found at low frequency can be causal.

Historically, interrogation of rare variation has been limited to resequencing a select group of candidate genes or structural differences. These structural differences, such as genomic losses and gains, are inferred from genotyping and tiling arrays. As technologies for exome, transcriptome, and whole genome sequencing become affordable (e.g. ABI SOLiD and Illumina Genome Analyzer), researchers can capture the entire realm of sequence variation in a sample population and test rare variant contributions to complex traits and diseases at a genome-wide scale.

Experimental Design

Like CDCV, testing the CDRV hypothesis begins by utilizing case/control design or quantitative design. Ideally, all cases and controls should be sampled from a homogenous population. Similar to testing the CDCV hypothesis, population structure is an issue for CDRV but for an entirely different reason. In CDCV, population structure is tested for and controlled. In CDRV, however, population structure is eliminated through removing outlier samples until a homogenous population remains. Outlier samples host variants that are private to the source population, but not present in the population being studied. Their removal is necessary because samples from different populations will inflate the number of rare variants detected.

The types of rare variants studied will determine the technology used to interrogate the samples. Rare structural variants have been interrogated using comparative genome hybridization (CGH) genotyping and tiling arrays [35], while resequencing can better interrogate rare SNPs and small insertions and deletions (indels) [36]. For studying a limited set of candidate genes, technologies like Sanger sequencing and Pyrosequencing are suitable; however, to study rare variants, genome-wide, cost-effective massively parallelized or "next gen(eration) sequencing" is required.

Several different methods for testing rare variant contributions to diseases have been proposed, each evaluating a different aspect of rare variants. A commonly used approach involves testing for an excess number of individuals carrying rare alleles versus

individuals not carrying rare alleles in cases compared to controls. Significance is assessed using Fishers Exact test for one rare variant per gene or the Mantel-Haenszel test for multiple rare variants at a given gene [37]. Li and Leal proposed that for any given gene, one method for testing association is to collapse all rare variants into a single variant and then to test this collapsed variant for association with the phenotype [38]. Another commonly used method utilizes Fisher's Exact Test to evaluate the ratio of rare missense mutations to silent mutations in cases and controls [39]. This method is used since rare protein-altering (nonsynonymous) mutations are more likely to be deleterious than silent (synonymous) mutations and be more prevalent in disease-associated genes in the affected cohort. A common extension of both methods is to use mutation severity predictors (e.g. MAPP [40], SIFT [41] and Polyphen [42]) to classify the rare missense mutations as protein damaging or benign, and test for excess of rare damaging variants within a gene.

Like GWAS, similar multiple testing corrections are applied and *P* values are used to determine significance. While summary statistics describing the impact of the variant association like odds ratio and relative risk are used, it is unclear what these measures indicate and whether they can accurately predict the risk of disease.

Examples

Ever since the realization of gene-scale sequencing for hundreds of samples, researchers have focused on resequencing candidate genes implicated in human disease

or intermediate phenotypes (e.g. high density lipoprotein cholesterol 'HDL-C' levels, blood pressure) to uncover their genetic factors. Studies include Cohen *et al*.'s investigation of rare allelic contributions to HDL-C levels in Canadians and participants of the Dallas Heart Study. An excess of nonsynonymous mutations private to the low HDL-C group was found in three candidate genes, adenosine triophosphate binding cassette transporter A1 '*ABC1*', apolipoprotein A1 '*APOA1*', and lecithin cholesterol acyltransferase '*LCAT*', when compared to private nonsynonymous mutations found in the high HDL-C or nonsynonymous variants found common to both low HDL-C and high HDL-C groups [39]. In a related study, Cohen *et al.* link low absorption of cholesterol to rare nonsynonymous variants in the candidate gene Niemann-Pick Type C1 Like 1 (*NPC1L1*) [43].

A Framingham Heart Study report showed carriers of rare variants (MAF < 0.001) in the genes *SLC12A3*, *SLC12A1*, and *KCNJ1* had significantly lower blood pressure than non-carriers [44]. Contrasting carrier and non-carrier siblings further supported this result, demonstrating that the rare variant carriers had lower blood pressures than their non-carrier siblings. The authors also report that most the rare variants were predicted to have severe effects using SIFT [41] and PolyPhen [42], which predict amino acid substitution effects based on phylogenetic conservation and physical properties of the amino acids. Both the excess of rare variants in individuals with low blood pressure and the predicted severity of those variants suggest these mutations are deleterious and strongly affect blood pressure.

As analysis of whole genomes becomes tractable in both the sequencing technology and bioinformatics arenas, rare variant detection and analysis will expand to genome-wide scans rather than candidate gene resequencing. For example, a recent whole-genome sequencing of a single quartet, with both offspring affected with Miller's syndrome and primary ciliary dyskinesia, implicated a handful of SNPs and genes likely to cause the diseases based on observed inheritance [45]. An earlier study of Miller's syndrome demonstrated resequencing could be used to find causal variants in rare, Mendelian disorders. Ng *et al* [46] used exome resequencing of two siblings and two unrelated individuals affected with Miller's syndrome. Using just the two siblings, only nine genes showed mutational patterns that matched the recessive disease model, adding just one more affected individual reduced the number of genes to a single gene, *DHODH*, as associated with Miller's syndrome. As these whole genome studies expand to multiple families, candidate regions will become further refined without the limitation of only resequencing candidate genes.

Limitations

The major technological limitation for genome-wide testing of the CDRV hypothesis is resequencing. Even with affordable sequencing technologies, the limitation lies in not in the physical sequencing, but in the bioinformatics work for accurately assembling short sequence reads and annotating sequence variation [47]. Efforts are being made to improve methods for distinguishing true mutations from those that arise

from experimental and alignment error. Even if experimental error is eliminated, the ability to detect rare variants remains limited, but improves with increased sample size. Li and Leal report that in a random sample of 100 individuals, 18.1% of variants with a population frequency of 0.001 will be detected genome-wide. However, when the sample size is increased to 1000 individuals, 86.5% of variants with frequency 0.0001 will be detected. When focusing on detecting all variants within a gene for diseased vs. control cohorts, there is a positive correlation in probability of detecting all rare variants and relative risk [48].

Methods for testing rare CDRV are also limited and one of the current areas of research. Historically, the genomic unit for detecting excess of rare variants has been defined as a gene or groups of candidate genes. As full genome sequencing becomes more popular, defining the physical unit to test CDRV will become more involved. For example, gene definitions could be extended to include up and down stream regions, tests could be completed in sliding window fashion, or genes could be grouped by gene families, pathways, and/or interacting genes. Existing methods for testing CDRV need refinement, expanding on current methods to include expectations of disease mutation evolution and population genetics.

Natural Selection

Overview

Natural selection affects genomic variation, from altering the distribution of allele frequencies to shaping haplotype structure in a population. The premise of natural selection is alleles that enhance survival and reproduction will increase in frequency [49]. The major types of natural selection include positive, negative or deleterious, and balancing selection. Mutations conferring increased fitness will increase in frequency or undergo positive selection. Negative selection occurs when mutations that decrease reproductive fitness are removed from the population. Finally, balancing selection, sometimes known as heterozygote advantage, occurs when allelic heterogeneity has increased reproductive fitness. Several tests of neutral evolution have been developed based on neutral theory and population genetic expectations.

Tests of Natural Selection

Model-based and qualitative tests have been developed to detect different types of selection which each leave different signatures in the genome. Tests of selection include site frequency based tests (e.g. Tajima's *D*), divergence based tests (e.g. McDonald Kreitman), or haplotype sharing tests (e.g. Long Range Haplotype).

In resequenced genomic regions, the distribution of allele frequencies is informative to the selective events that have occurred. Two examples of allele frequency based tests of natural selection are evaluating departures of a population's site frequency

spectrum (SFS) from neutral expectations and Tajima's *D*. The site frequency spectrum provides information about the history of a population, including selection. However, demography and selection can have similar effects on the SFS: population growth can look like negative selection, and bottlenecks can look like balancing selection. SFS-based inferences of selection must also account for demography, typically by estimating demographic parameters from an SFS derived from non-functional sites and estimating selection parameters (γ) from functional sites [50,51]. Like the SFS, demography and selection both affect Tajima's *D*. Tajima's *D* [52] is the contrast of two population mutation rate ($4N\mu$) estimators: θ_W based on the number of segregating sites detected and π based on the number of pair-wise differences. Under neutral theory of evolution, the two mutation rate estimators are equal and Tajima's *D* is zero. A positive Tajima's *D* indicates balancing selection or population bottlenecks, while negative Tajima's *D* can arise from positive selection or population growth [53]. Since demographic forces affect the whole genome and selection is locus specific, contrasting gene specific Tajima's *D* to genome-wide estimates is a method of isolating selection effects from demographic effects.

Differences in accumulation of synonymous and nonsynonymous mutations between species can be used to detect positive or balancing selection. McDonald and Kreitman [54] describe a test (MK test) of neutral evolution where the counts of synonymous and nonsynonymous fixed differences (D_s and D_n, respectively) between two species and the counts of synonymous and nonsynonymous polymorphisms (P_s and

P_n, respectively) make up a two by two contingency table. Under neutrality, the expected value of the expected value of the ratio P_s to D_s equals P_n to D_n and departures from neutrality are detected using a χ^2 test with 1 degree of freedom. Departures from neutrality can be assigned as diversifying positive selection when $D_n/D_s > P_n/P_s$ and as balancing selection when $D_n/D_s < P_n/P_s$ [55]. The MK test has been expanded to estimate the fraction of nonsynonymous divergent sites driven by positive selection (α) [56], the proportion of nonsynonymous polymorphisms that are deleterious [57], and the selection coefficient (γ) [58,59]. Unlike Tajima's *D*, the MK test is robust to demography, as demography affects both nonsynonymous and synonymous sites.

Qualitative based tests of positive selection evaluate haplotype sharing for each allele of a segregating variant. Selective sweeps leave a unique signature in the genome, as a specific allele rises in frequency due to positive selection, the haplotype the allele resides on will also increase in frequency, leaving a region in the genome with elevated LD surrounding the selected allele. This effect is known as the hitchhiking effect [60]. Partial selective sweeps, or selective sweeps that have not reached fixation, can be detected using the Long Range Haplotype (LRH) test. Haplotype homozygosity of an allele is measured using the extended haplotype homozygosity (EHH) statistic, defined as: given a core SNP allele and distance *x*, the probability two randomly chosen chromosomes carrying the core SNP allele are identical by state for the entire interval between the core region to the distance *x* [61]. Distance can be measured either as physical (bases) or genetic (cM) distance, however, using genetic distance is preferred to

control for variation in recombination rates. The LRH test tests for haplotypes with high frequency (>0.1) and high EHH relative to the genome-wide distributions and/or coalescent simulations are putative targets of positive selection. Extensions of this test include the iHS test [62] in which evaluates the log ratio of integrated EHH for each allele of the core SNP, and the XP-EHH [63] which contrasts haplotype homozygousity between populations to identify selective sweeps that have reached fixation in one population but not the other.

At the genome-wide level, EHH-based tests have an advantage over the MK test and Tajima's *D* since EHH can be used with ascertained data. Now that genome-wide resequencing is feasible, this advantage is no longer key. EHH-based tests also allow detection of very recent selective sweeps (less than 400 generations ago) with more power than Tajima's *D* or the MK test [61]. Tajima's *D* can be used to detect selection that has occurred 40,000 to 280,000 generations ago [64], and the MK test can be used to detect older selective events that have occurred since speciation.

Examples

Glucose-6-phosphate dehydrogenase (*G6PD*) is an X-linked gene that encodes an enzyme key to withstanding oxidant stress. There are three major alleles of *G6PD* (B, A, A-) with enzyme activities ranging from 100% to 12%. *G6PD* allele frequencies vary by geography [65]. At the global perspective, there is evidence of balancing selection through reports of positive Tajima's *D* and the observation all the alleles are frequent (>

10%), including alleles associated with reduced activity [66]. However, the LRH test showed a significant signal of positive selection for a *G6PD* haplotype found in the African populations and not the European-American or Asian populations [61]. The *G6PD* haplotype that showed extended haplotype homozygosity is one with reduced enzymatic activity and causes acute hemolytic anemia. In malaria endemic regions, the same reduced activity *G6PD* A- alleles confer a reduced risk (50%) for severe malaria [67], likely due to the anemic red blood cells being toxic to the parasite. This suggests in malaria endemic regions, carrying deficient *G6PD* alleles is more advantageous than carrying the normal *G6PD* allele. Meanwhile, in non-malaria endemic regions, the reduced activity *G6PD* alleles have been removed from the populations due to its reduction in fitness.

Complex Disease Model Predictions

The CDCV hypothesis postulates causal variants are selectively neutral due to the complex nature of the disease, however this assumption is often incorrect. For example, complex diseases like autism and schizophrenia have reduced reproductive fitness and thus, are not selectively neutral phenotypes[68,69]. Causal variants may be subject to deleterious selection and slowly removed from the population, suggesting the CDRV model better describes causal variants. At the other extreme, variants associated with advantageous phenotypes (e.g. parasite drug resistance) may be subject to positive selection and reach fixation in study populations or have allele frequencies that vary by

environment (e.g. *G6PD*). For these reasons, it is key to investigate both allelic associations and evidence of different types of selection when testing genetic contributions to complex diseases.

References

1. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, et al. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261: 921-923.

2. Lander ES (1996) The new genomics: global views of biology. Science 274: 536-539.

3. Tomlinson IP, Bodmer WF (1995) The HLA system and the analysis of multifactorial genetic disease. Trends Genet 11: 493-498.

4. Consortium IH (2003) The International HapMap Project. Nature 426: 789-796.

5. Consortium IH, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

6. Chee M, Yang R, Hubbell E, Berno A, Huang XC, et al. (1996) Accessing genetic information with high-density DNA arrays. Science 274: 610-614.

7. Jeffares DC, Pain A, Berry A, Cox AV, Stalker J, et al. (2007) Genome variation and evolution of the malaria parasite Plasmodium falciparum. Nat Genet 39: 120-125.

8. Mu J, Awadalla P, Duan J, McGee KM, Keebler J, et al. (2007) Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. Nat Genet 39: 126-130.

9. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

10. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

11. Consortium WTCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.

12. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74: 106-120.

13. (2005) nature04240-s5. 1-1.

14. Armitage P (1955) Tests for Linear Trends in Proportions and Frequencies. Biometrics 11: 375-386.

15. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100: 9440-9445.

16. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, et al. (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci USA 105: 4340-4345.

17. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870-874.

18. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087-1093.

19. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, et al. (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 41: 579-584.

20. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, et al. (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459: 528-533.

21. Weiss LA, Arking DE, Consortium GDPoJHtA, Daly MJ, Chakravarti A (2009) A genome-wide linkage and association scan reveals novel loci for autism. Nature 461: 802-808.

22. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, et al. (2009) Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460: 753-757.

23. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, et al. (2009) Common variants conferring risk of schizophrenia. Nature 460: 744-747.

24. Consortium IS, Purcell SM, Wray NR, Stone JL, Visscher PM, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748-752.

25. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317: 944-947.

26. Jallow M, Teo Y, Small K, Rockett K, Deloukas P, et al. (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet.

27. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39: 857-864.

28. Aranzana MaJ, Kim S, Zhao K, Bakker E, Horton M, et al. (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1: e60.

29. Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, et al. (2005) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Genet Genomics 274: 515-527.

30. Spencer CCA, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5: e1000477.

31. Li M, Li C, Guan W (2008) Evaluation of coverage variation of SNP chips for genome-wide association studies. Eur J Hum Genet 16: 635-643.

32. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124-137.

33. Xu B, Woodroffe A, Rodriguez-Murillo L, Roos JL, van Rensburg EJ, et al. (2009) Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. Proc Natl Acad Sci U S A 106: 16746-16751.

34. Xu B, Roos JL, Levy S, Van Rensburg EJ, Gogos JA, et al. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 40: 880-885.

35. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7: 85-97.

36. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, et al. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19: 1622-1629.

37. Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, et al. (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proc Natl Acad Sci USA 101: 15992-15997.

38. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311-321.

39. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305: 869-872.

40. Stone EA, Sidow A (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 15: 978-986.

41. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11: 863-874.

42. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894-3900.

43. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, et al. (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A 103: 1810-1815.

44. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40: 592-599.

45. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636-639.

46. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42: 30-35.

47. Mcpherson JD (2009) Next-generation gap. Nat Meth 6: S2-5.

48. Li B, Leal SM (2009) Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genet 5: e1000481.

49. Hartl DL, Clark AG (1997) Principles of population genetics. Sunderland, MA: Sinauer Associates. xiii, 542 p. p.

50. Williamson S, Perry SM, Bustamante CD, Orive ME, Stearns MN, et al. (2005) A statistical characterization of consistent patterns of human immunodeficiency

virus evolution within infected patients. Molecular Biology and Evolution 22: 456-468.

51. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4: e1000083.

52. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

53. Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. Mol Biol Evol 22: 63-73.

54. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.

55. Parsch J, Zhang Z, Baines JF (2009) The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in Drosophila. Mol Biol Evol 26: 691-698.

56. Smith NG, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415: 1022-1024.

57. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227-1234.

58. Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, et al. (2002) The cost of inbreeding in Arabidopsis. Nature 416: 531-534.

59. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.

60. Barton NH (2000) Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci 355: 1553-1562.

61. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

62. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. Plos Biol 4: e72.

63. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

64. Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413-429.

65. Ruwende C, Hill A (1998) Glucose-6-phosphate dehydrogenase deficiency and malaria. J Mol Med 76: 581-588.

66. Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, et al. (2002) Evidence for balancing selection from nucleotide sequence analyses of human G6PD. Am J Hum Genet 71: 1112-1128.

67. Ruwende C, Khoo SC, Snow RW, Yates SN, Kwiatkowski D, et al. (1995) Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376: 246-249.

68. Bassett AS, Bury A, Hodgkinson KA, Honer WG (1996) Reproductive fitness in familial schizophrenia. Schizophr Res 21: 151-160.

69. Lord C, Cook EH, Leventhal BL, Amaral DG (2000) Autism spectrum disorders. Neuron 28: 355-363.

Chapter 2 Global Genome-wide Variation is Shaped by Selection and Drug Resistance in the Agent of Malaria, *Plasmodium falciparum*

Rachel A. Myers[1,2], Kate M. McGee[3], Jon Keebler[1], Martine Zilversmit[1], Gilean A. T. McVean[4], Jianbing Mu[5], Junhui Duan[5], Xin-zhuan Su[5], and Philip Awadalla[1]

[1]Ste. Justine Research Centre, Department of Pediatrics, University of Montreal, Montreal, CA, H3T 1C5. [2]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695-7614, USA. [3]Laboratory of Experimental Immunology, National Cancer Institute-Frederick, Frederick, MD 21702, USA, [4]Deparment of Statistics, University of Oxford, Oxford OX1 3TG, UK. [5]Department of Biology, Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 12735 Twinbrook Parkway, Rockville, MD 20850, USA.

Submitted to *PLoS Genetics*

Abstract

Plasmodium falciparum, the etiological agent of the most severe form of human malaria, kills more people than all inherited diseases combined. Little is known, however, about the distribution and diversity of mutations in the *P. falciparum* genome, either among populations or in specific regions of the genome. Here we show the extent to which immunity and/or drug associated selective pressures shapes these parasite genomes. Analyses of a resequencing and microsatellite survey of 99 parasite genomes from 3 continents reveals that the genome-wide rate of mutation for *P. falciparum* is low relative to other eukaryotes, however, a number of mutations and genes, particularly those associated with drug resistance, are contributing to local adaptation among continents. For example, almost all amino acid changes in genes involved in host-pathogen interactions and electron transport are adaptive. After controlling for population structure, we identified a number of new, as well as previously reported, genes associated with resistance to three anti-malarial drugs. We find differences in associations between populations indicative of population specific adaptation, and often multiple drugs associated with the same polymorphism. We validate these association methods and demonstrate the power of our methods in the context of *P. falciparum* drug phenotypes and population structure. We note that a number of loci associated with drug resistance show signatures of positive selection. Model-based analyses of selection reveal that these associated genes, including *pfcrt*, have some of the highest levels of diversity in the

genome, suggesting that these genes were subject to a form of balancing selection previous to the introduction of drugs and drug selective sweeps.

Author Summary

Many investigations of drug resistance and selection in malaria parasites have adopted candidate loci approaches with limited geographical sampling. Here we use a global sample of parasites with genome-wide variation to make inferences of selection and drug resistance association in *P. falciparum*, a causative agent of malaria. By examining population structure, we evaluate the relationship of genotype and geography, make global and regional inferences of selection, and observe how response to selective pressures varies by continent using a wide range of population genetic tools. We observe a collection of haplotypes that have undergone continent specific selective sweeps. Additionally, the population structure results were used as corrections in a drug resistance association study of cultured parasites to eliminate spurious associations caused by population structure. The power of the association methods was evaluated in depth with simulated structured populations and experimental phenotypes. We also used two methods that treat population structure as either a fixed or random effect to further assess the reliability of these population-based association methods. In this portion of the study, we detect associations in a previously reported drug resistance associated gene, the putative chloroquine resistance transporter (*pfcrt*), as well as new genes associated with *in vitro* drug response.

Introduction

The deadly relationship between humans and *Plasmodium falciparum* has been a driving evolutionary force shaping genome composition in both species. Among populations, this antagonistic evolutionary process is predicted to increase rates of local adaptation of the parasite to different host populations or environments [1]. Global parasite populations are exposed to varying levels of immune system or drug pressures and therefore may use a number of deterministic mechanisms to adapt to these new environments. Natural or artificial selection leaves signatures in genomes that may contribute to population specific adaptations and is useful for identifying new candidate targets for vaccines [2,3]. Allelic variation at alternative loci may be responsible for differences in susceptibility to different host environments. Identifying the extent to which distinct populations of parasites use differing means to evade immunity is critical to identifying new potential targets for vaccines or drugs, and in understanding how these populations may respond to drugs or vaccines.

Our knowledge of population diversity in *P. falciparum* varies widely depending on either the genomic regions, markers or population studied [4,5,6,7]. For example, a number of previous studies of diversity and population structure in *P. falciparum* have focused specifically on sequence data of either known antigen genes that are likely to be under immune selection [for example 8] or non-coding regions [9]. Many genome-wide surveys have also used non-coding molecular markers, such as microsatellites [4] that are

unlikely to be the target of selection. Genotyping single nucleotide polymorphisms (SNPs) [for example 10,11] or probing single feature polymorphisms (SFPs) [3] are useful in capturing a genome-wide picture of variation. Ultimately, re-sequencing surveys of genomic variation is critical for capturing the fine-scale information necessary for making inferences about selection and demographic processes [11,12,13].

Determining the influence of selection on the evolution of the *P. falciparum* genome requires global approaches to analyzing patterns of coding variation. Currently, it is not known to what extent recent selection events, such as those associated with drug or immune evasion, are geographically restricted. For example, *P. falciparum* may evolve differently among populations due either to, varying levels of drug exposure or variable transmission rates in different regions. Selection due to the evolution of drug resistance in parasites is of particular importance to medical science, and is a worldwide problem; reported failure rates for the historically effective treatment chloroquine in South America are 80%, 50% in East and Central Africa, and 40% in South-East Asia (World Malaria Report 2005).

To investigate global and regional selection and population variability, we resequenced 93 genes found on chromosome 3, from 99 global isolates in four major regions (*n*=35, Africa; *n*=29, Asia; *n*=23, S. America; and *n*=11, Papua New Guinea). A subset of these SNPs were previously analyzed [10].We also resequenced 60 of the 93 loci on chromosome 3 for an out-group, *P. reichenowi* . This is the first re-sequencing survey of this large number of samples and genes for *P. falciparum*. We use the

resequencing results in conjunction with the previously published resequencing survey of transporter genes [14,15], a genome-wide microsatellite survey [16], and the genome-wide resequencing survey of five isolates [13] in this analysis.

Our analyses address three related issues about the evolutionary mechanisms shaping genome-wide variation. First we infer how different forms of selection shape variation genome-wide among different populations and identify potentially new candidate genes/SNPs subject to immune or drug pressures. Second, to address how different populations were adapting to different environments, we compare signatures of selection at individual genes among the three continents. We apply new population structure correction methods to the chromosome 3 and transporter re-sequencing surveys and ask if there are genotypes that correlate with drug resistance data for three anti-malarial drugs- atovaquone-proguanil, chloroquine, and quinine. Additionally we investigate microsatellite associations with drug resistance and the performance of these association methods. Finally, we asked whether functional variants, including those we find in association with drug resistance are differentiated among populations more so than random expectation and subject to selection using likelihood based methods. Here we show that these methods are robust even when used with relatively small samples.

Results and Discussion

Population Mutation Rates and Inferences of Selection

A recent genome-wide SNP survey revealed the overall population mutation rate at coding regions for *P. falciparum* is small ($2N\mu(2\text{-}s) = 0.0005$ per nucleotide, where N is the effective population size, μ is the per base mutation rate, and s is the fraction of the population produced by selfing [17]) relative to other organisms [13]. The high AT base composition (~80%) of *P. falciparum* has clearly played a significant role in limiting overall level of silent polymorphisms [18]. However, through surveys of polymorphism over large genomic regions from large numbers of parasites, we can draw some general and specific conclusions about the stochastic and deterministic forces shaping variation in *P. falciparum.*

Sequence data from a collection of 99 isolates from four major geographical regions revealed 368 single nucleotide polymorphisms (over 44,296 bases). Table 2-1 shows the breakdown of synonymous and nonsynonymous mutations per population. We observed an excess of nonsynonymous polymorphism, globally and in Africa (Table 2-1), when using the McDonald-Kreitman (MK) test [19] which included polymorphism data within *P. falciparum* and interspecific differences between *P. falciparum* and its most closely related species, *P. reichenowi*. Also, when genes were grouped by gene ontology (Generic GO Term Mapper, http://go.princeton.edu/cgi-bin/GOTermMapper), some groups showed an excess of nonsynonymous polymorphism consistent with balancing selection. The effect was more pronounced when we removed polymorphisms

segregating at < 5%. Other GO groups showed an excess of nonsynonymous substitutions between species consistent with adaptive evolution (Figure 2-1).

We inferred the genome-wide mutation rate and the deleterious mutation rate in coding regions, and asked what proportion of mutations is subject to positive selection [20,21]. All methods used to make these inferences make an assumption that silent mutations are neutral. The method of Bierne and Eyre-Walker [22] used the frequencies of silent and replacement polymorphisms within species, and similar counts between species (*P. falciparum* and *P. reichenowi*), to infer the proportion of amino acid substitutions (α) that were adaptive. For *P. falciparum* chromosome 3 data, this quantity was not significantly different from zero (not shown), similar to calculations made for humans [23]. However, using polymorphisms discovered genome-wide by Mu et al. [13] and Jeffares et al. [12], which were well-validated and were grouped according to gene ontology, specific functional categories showed a significant signature of positive selection (Table 2-2). These categories included genes associated with host-pathogen interactions and electron transport. Genes coding for transporter proteins were of particular interest because these proteins are known to be associated with drug resistance in a number of species including *P. falciparum*.

We mapped candidate regions along chromosome 3 that were associated with local adaptation in each of the populations surveyed. We first performed a number of standard tests for departures from neutrality. The Tajima's *D* [24] statistic assessed whether the frequency spectrum of polymorphism at each gene departs from neutrality.

Simulations were conducted of the coalescent to assess significance using demographic parameters inferred by Joy et al. [6]. Specifically, *p* values for the Tajima's *D* statistics for loci of parasites from Africa and South America were estimated based on 10,000 coalescent simulations of stationary population sizes conditioned on the number of segregating sites. Simulations of a population growth model were performed for comparisons to PNG and Asia Tajima's *D* statistics [6].

The distribution of Tajima's *D* values on the 3rd chromosome per gene varied spatially among populations (Table 2-3). Relative to other populations, Africa had a greater mean negative Tajima's *D* across genes (Table 2-4), suggesting a higher proportion of rare or singleton SNPs, possibly due to population expansion. PNG had more heterogenous distribution, a number of genes showed the signature of either recent population-specific selective sweeps or balancing selection (Table 2-3). For example, a number of contiguous coding regions on chromosome 3 among African parasites (regions included loci *PFC0175w* to *PFC0185w* and *PFC250c* to *PFC310c)* had significantly negative Tajima's *D* values, as would be expected for a local selective sweep in the region. Similarly, among Asian samples, a region at the 5' end of the chromosome consisting of 3 loci *(PFC0050c*, *PFC0060c*, and *PFC0075c*) also had significantly negative Tajima's *D* values. One locus was significantly positive in both Asia and PNG (*PFC0505w*).

Finally, selective sweeps that have not yet reached global fixation may drive new adaptive mutations up to intermediate frequency. In these cases, the favored allele

increases in frequency very fast, and as a result is associated with an unusually long haplotype of low diversity. We have previously showed that LD and recombination rates vary among populations [10]. Therefore, a haplotype that is unusually long will standout as a region subject to recent selection. To identify haplotypes that were unusually long, we calculated the integrated Haplotye Homozyougisty (iHH) for each SNP along chromosome 3 [25]. Simulations were performed to identify outliers. We computed these statistics for all populations combined as well as separate. Haplotypes near the sub-telomeric regions stood out as having an excess iHH relative to the rest of the chromosome where a number of potential antigen encoding genes reside [26]. The only other potential signature of a partial selective sweep was in Africa in a region spanning 197313 to 318121 bp that included loci *PFC0185w* through to *PFC0310c*. Interestingly, loci *PFC0185w, PFC0195w*, *PFC0245c, PFC0295c*, and *PFC0310c* also showed nominal associations with chloroquine and quinine resistance among African parasites (see below).

Geographic selection

We also asked whether functional and non-functional classes of sites evolve differently among populations by estimating rates of genetic drift for these different classes. Under selective neutrality, genetic differentiation between populations is determined by genetic drift; however, natural selection acting upon specific loci may homogenize differentiation (e.g. balancing selection) or lead to an excess of

differentiation (e.g. geographically restricted directional selection relative to neutrally evolving loci throughout the genome) [27,28,29,30]. To measure population differentiation among sites for all four populations, we used a Bayesian method [31,32] that fits a statistical model of population structure related to F_{ST}, the coefficient of population differentiation. The model defined the difference of each population from a hypothetical average, or ancestral, population by a variance parameter c_j for each population j [31] (Table 2-5). This parameter was analogous to F_{ST} values, but instead of being limited to a single estimate of differentiation for a pair or group of populations, we have an estimate of the variance parameter for each population.

Synonymous, nonsynonymous, and transporter gene SNPs exhibited significantly different patterns of differentiation among populations. In Africa and in America (South and Central), nonsynonymous sites were significantly more differentiated than synonymous sites (Rank-Wilcoxon $p<0.001$, $p=0.003$, respectively). A number of genes for which there was strong differentiation between populations also showed significant association with drug-resistance in one or more populations [14], and overall we could reject the hypothesis that the two are uncorrelated. This result was perhaps surprising, as one might expect drug resistance mutations to have reached global fixation. However, geographic variation in drug usage, and fitness-costs associated with resistance, will lead to a selection-migration balance that allows for fixation of advantageous variants in geographic regions of high drug-usage, and polymorphism in regions of lower usage [33]. We suggest that peaks of differentiation may therefore reflect loci involved in resistance

to anti-malarial drugs, and we tested the assertion by examining genotype - drug response associations.

Population Structure and Associations with Drug Resistance

Previous association studies conducted in *P. falciparum* have used geographic locations as indicators of population structure and correct for the structure by conducting within continent or sampling location association tests [14]. This approach is sensitive to spurious associations due to recent migrants, admixed isolates, mislabeling, and other instances where geography is an inaccurate indicator of the population structure. We improved on these population structure corrections by using a new but widely used [for example 34,35,36,37,38,39] principal component analysis (PCA) approach. PCA was both a test for population structure and also a correction method, permitting isolates to have continuous mixtures of ancestry that are defined genetically, rather than having discrete populations defined by collection location. The power and Type 1 error of the PCA implementation EIGENSOFT [40,41] and a simple sequence repeat (SSR) association test were evaluated using simulated structured populations and experimental phenotypes (chloroquine IC_{50}). We then tested the presence of population structure using the chromosome 3 SNPs along with the previously analyzed transporter genes [14] with EIGENSOFT. Additionally, the SNP data combined with variation at 348 SSR loci [16] spaced ~75 kb apart throughout the genome was analyzed using STRUCTURE [42,43] to also assess population structure. Finally, we used the PCA results for population structure

correction in the SNP data, and structure results for the SSR data, to detect statistical associations with *in vitro* drug resistance, measured as the concentration required for 50% inhibition of parasites (IC_{50}).

Power and Type 1 Error of Association Methods for P. falciparum Drug Response

The power to detect associations using EIGENSOFT has been reported and showed improvement over other association tests in structured populations, however these reports are only suggestive of the power and error in our study. We assessed the power and validity of EIGENSOFT to detect associations using the experimental phenotypes and simulated structured populations. Using Hudson's coalescent simulator ms [44], we simulated an island model with four populations, unequal migration, unequal population size and recombination ($4Nr = 50$). The PCA of the simulations showed several significant axes of variation indicating population structure in the datasets. To assign phenotypes, the simulated sites were screened for SNPs with frequency 0.366 – 0.42, corresponding to the frequency of chloroquine susceptible isolates in the sample. The resulting SNP was then used to assign phenotypic values from the chloroquine susceptible and resistant IC_{50} distributions to each sample, resulting in a completely linked SNP – phenotype dataset. One hundred replicates were analyzed using EIGENSOFT and *p* values were obtained by permutation tests. The power for this method experiment-wide was 0.62 and the Type 1 error was $< 8 \times 10^{-4}$. At the nominal level (p value ≤ 0.05) the power was .95 and the Type 1 error was 0.03. The power study

revealed the method had sufficient power to detect associations and a conservative Type 1 error.

SSR polymorphism data traditionally has not been used in the association framework and has the drawback of too much allelic diversity (relative to the number of samples), yet because SSRs were already available [16], we used them to also test for associations. We evaluated the performance of the SSR association test by examining the power and Type 1 error. The power to detect associations and Type 1 error of the SSR association method was estimated by simulating the largest sub-population, Africa, using the demographic parameters reported by Joy et al. [6]. Simcoal2 [45], a coalescent genetic diversity simulator, was used to generate samples (40) of completely linked SSRs and SNPs. A single SNP was removed and used to assign resistant or susceptible chloroquine IC_{50} values to each sample, resulting in completely linked SSR – phenotype samples. The simulated samples were then tested for associations. This process was repeated 1000 times. At the Bonferroni significance level ($\alpha = 0.05$), the power to detect a true association given the distribution of chloroquine phenotypes in Africa was 0.004, and at the nominal significance level (p value ≥ 0.05) the power was 0.17. The Type 1 errors experiment-wide and nominally were 0.001 and 0.045 respectively. The low power was expected due to small sample size and high allelic diversity of the SSRs, however the low Type I error demonstrated the statistical credibility of the SSR associations detected.

Finally, we used a linear mixed model association method, which used an identity by state matrix to control for population structure. Efficient Mixed Model Association

(EMMA) [46] differed from EIGENSOFT by modeling population structure as a random effect, meaning the observed population structure in the samples is a subset of the naturally existing population structure. Conversely, EIGENSOFT modeled population structure as a fixed effect meaning the observed structure is a complete sampling of the naturally existing population structure. The resulting population structure correction used in EMMA was larger due the larger variance estimator of the random effect and yielded a more conservative test of genotype-phenotype association.

Population Structure in Global Samples of P. falciparum

The PCA of the SNP data showed five significant axes of variation; the top 2 axes of variation accounted for 16% and 7.6% of the genetic variation and separated Africa and AsiaPNG, and then America from Africa (Figure 2-2 A). This population partitioning agreed with previous empirical [4,16,47] and STRUCTURE results [10]. The 3rd, 4th, and 5th axes (explained 4.0%, 3.7 % and 3.6% of the genetic variation) appeared to stratify the American population, separating samples from Brazil and Peru, with consistent variance in the remaining populations. Similarly, in the STRUCTURE analysis for the SNPs and 348 SSRs (Figure 2-2 B) the best-supported model was K = 3 populations consistent with the top 2 PCA results and the previous studies. We found a handful of isolates appearing to be an admixture of two or more groups which was also consistent with previous studies [10]. For example, we found that some Asian isolates appeared to be a mix of the "African" cluster and a 3rd cluster (Figure 2-2 C). We observed that for a large

proportion of Asian samples, genome-wide "mixing" or recombination was evident among African (in red) and Asian (in yellow) samples. Also clear, was the assignment of PNG samples to Africa. In this case, the PCA population structure correction provided an advantage over using continental groupings for association studies; the admixed samples having corrections according to all axes of ancestry, rather than a single population assignment.

Drug Response Association

Drug resistance, a continuous trait, was measured as the concentration required for 50% inhibition of the pathogens (IC_{50}) for atovaquone-proguanil (Table 2-6), chloroquine, and quinine. The latter two are described in Mu et al [14]. The SNP data was analyzed using an implementation of PCA (EIGENSOFT [40,41]) for detecting and controlling for population structure. The PCA revealed a clustering of chloroquine susceptible (IC_{50} 13.6-76.1 nM [14]) isolates (Figure 2-4) suggesting the population structure may reflect differences in drug tolerance. To avoid population structure corrections that were related to phenotype, the significant axes of variation were evaluated for correlations with the drug phenotypes. Each population class (Global, Africa, Asia, and America) was evaluated individually for population structure and associations. The five significant axes of variation discussed in the population structure analysis were used for corrections in the Global population; two significant axes of variation (p value = 1.5×10^{-8} and 0.0019, respectively) were used for population

structure correction in the Asian population, while the African and American populations had no principal component corrections due to phenotypic correlations. We detected 8 SNPs in two genes (Table 2-7), and 6 different SSR loci (Table 2-8), with significant drug resistance associations after Bonferroni multiple testing corrections ($\alpha = 0.05$). SNP and SSR drug resistance associations that were population specific may have evolved before or after population differentiation, perhaps in response to geographic differences in drug treatments over the past few decades.

Two genes having SNPs in association with drug phenotypes using both EIGENSOFT (Table 2-7) and EMMA were *MAL7P1.27*, or the putative chloroquine transporter (*pfcrt*), and a novel association with *PFC0850c*, a hypothetical protein. Polymorphisms in *pfcrt* have a well-documented role in chloroquine resistance [14,48,49,50] and served as positive controls for the PCA association method. *PFC0850c* was associated with quinine resistance at the global level and had nominal associations with both chloroquine and quinine in the AsiaPNG population. *PFC0850c* is orthologous to an endonuclease/exonuclease/phosphatase of the rodent malaria parasite *P. yoelii* and other *Plasmodium* species. *PFC0850c* has not been previously reported to be involved in drug resistance. The previous analysis of variance of the transporter genes revealed several significant associations [14]; only one gene was detected at the experiment-wide significance level in this study. This suggested the PCA and the EMMA correction methods used here were more conservative than the previous study.

A population specific SSR association (Table 2-6) was found at marker C13M30 (*p* value = 0.034) on chromosome 7 (599.7-.9 kb, 40.3 cM), overlapping *MAL7P1.50* encoding the erythrocyte membrane protein 1 (*pfemp1*), a cytoadherence protein expressed on the surface of infected red blood cells. Two novel associations included a chromosome 7 marker, 7A11 (332.87 kb, 23.1 cM), within the gene *PF07_0024,* (putatitive inositol phophatase) associated with chloroquine response (*p* value < 0.019). Also, two novel markers were identified on chromosome 4 to be in association with atovaquone-proguanil resistance in Africa.

The Bonferroni multiple testing correction tends to be conservative, therefore we also reported SNPs and SSRs with nominal *p* value < 0.05 in at least 2 of the 3 populations, or at the global level (Tables 2-9, 2-10, and 2-11 respectively). We observed chloroquine resistance associated with multiple *pfcrt* SNPs in both the African and American populations, reflecting differences in chloroquine resistance segregation between populations. The ABC transporter also showed a similar pattern of nominal associations with chloroquine in Asia and America. SSR marker C4M39 on chromosome 4 (666.2 kb, 31.6 cM) had associations with quinine and atovaquone-proguanil in at least 2 populations. We also saw a collection of SNPs, including those in *pfcrt* and *PFA0590w* (putative ABC transporter), had nominal associations with multiple drugs (chloroquine and quinine). Sites that were associated with different drugs acting in opposing directions (different alleles favorable for different drug resistances) may act to maintain variation among populations, which may be a form of balancing selection.

Selection at Drug Associated Loci

In the case of non-neutral phenotypes, association results may be considered jointly with selection. Selective sweeps can remove polymorphism, thus decrease the ability to detect associations. In contrast, balancing selection, which maintains polymorphisms within or among populations, resulting in alleles with higher frequencies, increases power to detect associations with phenotypes. Genes that have SNPs in association with drug phenotypes were tested for departure from neutrality using MLHKA [51], a maximum likelihood multi-locus extension of the HKA [52] test. Two models were tested; the alternative (selection) model that included two genes with SNP associations and *pfmdr1* as targets of selection and the null (neutral) model where all genes were selectively neutral. The maximum likelihoods of both models were determined and likelihoods were tested for statistical improvement of the alternative over the null. The test was carried out in each population separately using the same gene set. We observed a significant improvement for the selection model over the neutral model in all populations (Global, p value = 1.5e-5; Africa, p value = 0.0012; Asia - with PNG, p value = 5.4e-6; America, p value = 0.00037). The parameter k, a scaling factor which estimates how much diversity has been decreased or increased from neutral expectations in response to selection, was 11.5 globally, 8.4 in Africa, 9.9 AsiaPNG, and 11.2 in America for *pfcrt*, suggesting variation was high at these loci before a partial drug sweep. The high diversity at *pfcrt* suggested this locus was subject to balancing selection before

P. falciparum exposure to drug treatment, and that the use of drugs caused a selective sweep in some populations. Even if different mutations or haplotypes confer resistance, that variation had to be present for selection to act upon it. Similarly, hypothetical protein *PFC0850c* had *k* values ranging from 6.9 to 20.8 suggesting variation was maintained. Conversely, previously identified *pfmdr1* had *k* values that are all < 1, suggesting diversity has been decreased, which is consistent with the reported selective sweep.

Linkage disequilibrium (LD) block (regions where D' > 0.80) averaged range from 11.2 kb in Africa to 53 kb in PNG [10]. Therefore non-neutrally evolving loci within a 40 kb window of associated SSR loci were also of interest. Two such loci were *PFD0720w* close to SSR marker C4M39 and *PF14_0325* close to C14M115. *PFD0720w* appeared to have a deficit of fixed nonsynonymous sites (inferred from the MK test), suggesting selective constraint. The *PF14_0325* gene had a surplus of fixed differences suggesting positive selection.

Selective sweeps have previously been reported at *pfcrt* and *dhfr* (dihydrofolate reductase), with varying strengths, at different geographical locations [16,53]. Nash et al. [53] genotyped the *pfcrt* and *dhfr* loci in 130 parasites and showed losses of heterozygosity in strains resistant to chloroquine, consistent with loci subject to selective sweeps in a Southeast Asian population. Wootton et al. [16] evaluated genome-wide allelic diversity in the SSR dataset, noting high allele sharing along chromosome 7 and suggested that positive selection has occurred at *pfcrt* as separate events in the populations. Additionally, they reported LD extends > 200 kb from loci subject to

selection in chloroquine resistance parasites. Similarly, we observed chromosome 7 SSR associations ~ 100 and ~ 130 kb from *pfcrt*. As well, we saw SSR associations ~90 kb from *dhfr*, comparable to the 98 kb hitchhiking range in the Laos population reported by Nash et al., and an MK signal of non-neutrality at *PFD0720w*, 20 kb from *dhfr*. This suggested the SSR associations were reflections of drug induced selective sweeps and linkage disequilibrium rather than functional associations.

Conclusions

We have shown that *P. falciparum* has been subject to multiple forms of selection. Widespread anti-malarial drug-treatment has had a strong impact on the *P. falciparum* genome as noted by high differentiation of transporters SNPs and SNPs associated with drugs [26,54,55,56]. We showed that with even limited sample size, we could detect both known and novel associations. The sample sizes for our association were no doubt small relative to other studies in humans due to the laborious nature of testing for drug associations in the laboratory. However, our laboratory IC_{50} measures were fairly reliable compared to tests in the field [57], where starting and ending estimates of parasite numbers could be more accurately characterized. Finally, another novel finding in this study was the high degree of variation observed at previously characterized loci associated with partial sweeps. It raised questions about the function of these genes, apart from their association with drug resistance.

Finally, from our analysis of just one chromosome, we have identified a number of genes potentially subject to alternative forms of recent and ancient selection and a novel drug resistance association. The genome was fairly conserved relative to other genomes. No doubt, there were a number of constraints due to the complex life history of this particular pathogen. However, analysis of this one chromosome suggested that potentially 5% of loci genome-wide were subject to some form of adaptive selection. Although mutation rates appeared to be low in malaria, estimates were consistent with similar calculations made for other eukaryotes with short generation times. Regardless, a small proportion of amino-acid substitutions allowed the parasite to adapt to the ever-changing host environments.

Acknowledgements

We thank Tim Anderson for comments on an earlier version of this manuscript and M. Ferdig for the work on atovaquone-proguanil IC_{50} at the NIAID. This work was supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health. R.A.M. is supported by NIEHS training grant (T32 ES007329) in Bioinformatics and P.A. is supported by grants from the National Academies Keck Foundation Initiative (#Geno01), and the Human Frontiers in Science Program (# RGP0054/2006-C) and the NIH.

Methods

SNP and SSR Population Structure

The chromosome 3 sequence polymorphism was collected as described in Mu et al [13]. The transporter gene SNP set and the chloroquine and quinine IC_{50} for the parasites were described previously [14]. The chromosome 3 SNP data sets formed the SNP dataset (genotyped: $N_{Africa} = 36$, $N_{AsiaPNG} = 57$, $N_{America} = 23$). To ensure accuracy in allele frequency calculations, normalization, and the subsequent PCA, the haploid SNP data was recoded to diploid homozygotes (0, 2) data and then analyzed using the principle component analysis (PCA) based association method EIGENSOFT [41]. The SSR and SNP data were merged and analyzed using the admixture model implemented in STRUCTURE v. 2.2 [42] with ploidity = 1, K ranged from 2 to 5 and chains of 500000 iterations and 100000 burn-in iterations.

Ancestry of each locus was estimated from STRUCTURE output as follows:

$$\underbrace{\max}_{1..K}\left(p(k \mid n_{ij} = a) = \frac{p(a \mid k)\, p(k \mid n_i)}{\sum_1^K p(a \mid k)\, p(k \mid n_i)} \right)$$

where k is the ancestral population (1..K), n_{ij} is the genotype of sample i at locus j, $p(\, a \mid k\,)$ is the frequency of allele a at locus i in ancestral population k, and $p(\, k \mid n_i\,)$ is the ik^{th} element of the Q matrix, or the proportion of sample i's genome coming from population k.

SNP Association

The SNP data and atovaquone-proguanil, chloroquine and quinine IC_{50} datasets (phenotyped: $N_{Africa} = 24$, $N_{AsiaPNG} = 48$, $N_{America} = 16$) were analyzed for association analysis using EIGENSOFT, treating the phenotype as a continuous variable and correcting for population structure in both the genotype and phenotype. The 5 significant axes of variation from the PCA SNP structure analysis were used for structure corrections in EIGENSOFT for the association analysis of the Global population. Additionally, 2 axes of variation were used in the Asian population, and none in the remaining populations. The genotypes and phenotypes were adjusted by a factor of the location of the genotype or phenotype along the axis and the ancestry regression coefficient of the axis for all axes of variation used which is mathematically equivalent to using the axis of variation (eigenvector) matrix as a covariate in a multi-linear regression. The test statistic is $\chi^2 = (N - K - 1) * r^2$, where N is the sample size, K is number of axis of variation used in the adjustments, and r is the correlation of adjusted genotypes and adjusted phenotypes. The same recoding of genotypes (haploid to diploid) used in the PCA population structure analysis was used here. The recoding, or linear scaling, did not change the correlation (r^2) between genotypes and phenotypes; hence the test statistics and p values were unchanged.

Finally, phenotypes were permuted 250 000 times and (nominal) significance was obtained by counting the number of permuted test statistics equal to or greater than

the experimental test statistic. The Bonferroni multiple testing correction was then applied.

SSR Association

Each SSR marker *i* was tested using: $\chi_i^2 = \sum_{alleles} \frac{(P_a - n_a * \mu)^2}{n_a * \mu}$, where n_a is the number of isolates with allele *a*, μ is the phenotype mean of all samples, and P_a is the sum of the n_a phenotypic values with allele *a*. The analysis was done within each population; Africa, AsiaPNG, and America. Significance was obtained using the same permutation scheme as the SNP data. Sample sizes were small for the SSR dataset (68 samples phenotyped, $N_{Africa} = 34$, $N_{AsiaPNG} = 18$, $N_{America} = 16$) and subsequently reduced power of detection, particularly since allelic diversity was high, resulting in high degrees of freedom (up to 33 df) for each locus. We examined the effect of binning alleles into 2, 3, 4, 5, and 6 alleles total for each site to reduce degrees of freedom and increase power. The binned data was analyzed using a χ^2 test for each marker. Phenotypes were permuted 10000 times and (nominal) significance was assessed. The Bonferroni multiple testing correction was then applied. Percent of loci with a *p* value below 0.05, .001, and .005 for binning schemes and drugs were obtained, and although binning appeared to increase the number of loci detected, no one binning strategy improved detection over all phenotypes.

SNP Power Analysis

Ninety-nine samples of 201 sites were simulated from a four island model with unequal migration rates, unequal population sizes and recombination, generated using Hudson's ms [44]. The experimental chloroquine IC_{50} values were divided into two distributions- resistant and susceptible, defined by a natural break in the over all distribution, and resulted in a frequency of 0.39 for chloroquine susceptibility. The simulated sites where screened for segregation in at least 2 of the 4 populations and a minor allele frequency range of 0.366 - 0.425. The SNP meeting the frequency and segregating conditions was used to assign phenotypes, sampling with replacement from the susceptible and resistant IC_{50} distributions. The resulting datasets were analyzed using EIGENSOFT with corrections on 5 axes of variation. *P* values were obtained by permuting the phenotypes 250000 times and counting the number of times a test statistic was as or more extreme as the original results. The process was repeated 100 times to determine power, counting the number of the 100 replicates in which the linked SNP was significant. To determine Type 1 error, a phenotype permutation was taken as the result and compared to the rest of the permutations to obtain *p* values in each of the 100 simulations. Type 1 error was report as the number of times an association was detected at the unlinked SNP out of the 100 simulations. Both power and Type 1 error were determined with and without the Bonferroni multiple testing correction.

SSR Power Analysis

A sample size of 40 completely linked SSR and 10 SNPs were simulated using simcoal2 [45], and growth and population size parameters reported by Joy et al [6]. The SNPs were filtered for minor allele frequency > 0.225 and then used to assign phenotypes. To mimic the African population, the chloroquine IC_{50} values of the African isolates were divided into two distributions, resistant and susceptible, using the same break as in the SNP power analysis. The phenotypes were assigned to the simulated samples based on the filtered SNP, sampled with replacement from the two different distributions. The resulting data set was tested for associations and *p* values of the associations were obtained from 10000 permutations of the phenotypes. The process was repeated for 1000 simulations and power was determined as the number of times an association between the SSR and phenotype was detected. The same simulation process was conducted for an unlinked SSR and phenotype to determine Type 1 error. Both power and Type 1 error were determined with and without the Bonferoni multiple testing correction.

Maximum Likelihood Hudson Kreitman and Aguade (MLHKA) test

Polymorphism data and divergence data from Jeffares et al [12] were combined for the 2 genes with SNP associations, *pfmdr1*, and 10 randomly chosen genes from chromosome 3 and evaluated for signals of selection using MLHKA [51,52], a maximum likelihood multiple locus version of the HKA test. The null model consisted of all 13

genes as neutral loci and the alternative model was the 3 associated genes as selected loci. To evaluate model improvement, 2(*ln*(alt) – *ln*(null)) was assumed to be distributed χ^2 with 3 degrees of freedom. This test was repeated for the African, Asian, American, and PNG populations.

Proportion of Positively Selected Amino Acids

Inferring the proportion of adaptive mutations genome-wide or at particular groups of genes could be estimated by a simple extension of the McDonald–Kreitman test [19,20,21]. The assumption was made that synonymous substitutions and polymorphism are neutral. To estimate the average proportion of amino-acid substitutions which were driven by adaptive evolution we needed to combine data across genes (see rationale and method in Smith and Eyre-Walker [20]). The proportion of positively inferred amino acids, α, was approximately equal to $1-(D_s/D_n)(P_n/(P_s+1))$ where P_n and P_s were the number of polymorphic amino acid or silent sites within species. To estimate α we used the likelihood approach outlined by Bierne and Eyre-Walker [22]. Estimates were made genome-wide and for separate GO function categories but only for primate malaria as substantial population data does not exist for rodent malaria.

Calculation of M and U.

Sequences were aligned by ClustalX [58]. For each pair of genes, we calculated the rates of synonymous and nonsynonymous substitution. We used methods as described

in Eyre-Walker and Keightley (2000). The genomic amino acid (*M*) and deleterious (*U*) mutation rates were calculated as

$$M = Z(\overline{K_{ts}N_{ts}} + \overline{K_{tv}N_{tv}}) \quad U = M - Z\overline{K}_n/3$$

where averages were unweighted averages across genes and *Z* is a constant that converts the per site estimate to a per genome per generation estimate: e.g., for human/chimpanzee,

$$Z = \frac{2(\text{genomes}) \times 80{,}000(\text{genes}) \times 1500(\text{bp for a gene}) \times 25(\text{years for a generation}}{12 \times 10\text{^}6(\text{years of divergence})}$$

.

We used the Goldman and Yang [59] approach implemented in PAML to calculate synonymous and nonsynonymous substitution rates. We used the codon frequency model 3, which is a free parameter model and calculated substitution rates using base composition and codon frequency from the data. $M = Z(\underline{D_s})$ and $U = M - Z(D_n)/3$. The synonymous substitution rate was correlated to the level of synonymous codon bias in some species. This may be due to selection on synonymous codon use or a downward bias in the methods used to correct for multiple hits in sequences with high codon bias. Four-fold and two fold substitution rates were correlated to codon bias. We calculated L which was the departure from expected codon usage assuming no codon bias calculated as a weighted 2 statistic [60], summed over amino acids, assuming an empirical genomic G and C base content of 21% averaged across species. To correct our mutation rate estimates, we regressed the estimates of *M* and *U* obtained from each gene against the

average L value taking the y intercepts as our corrected M and U values; these were the predicted values of M or U for no synonymous codon bias.

Estimating Genetic Drift

Inferences of ancestral populations and measures of differentiation: The model took the form of a Beta-binomial

$$x_{ij} \sim Bin(n_{ij},\alpha_{ij});\ \alpha_{ij} \sim Beta\left(\frac{\pi_i(1-c_j)}{c_j},\frac{(1-\pi_i)(1-c_j)}{c_j}\right)$$
$$i = 1,...,L;\ \ j = 1,..P$$

where P is the number of populations, L is the number of loci, n_{ij} is the number of chromosomes typed at the i^{th} SNP in the j^{th} population and x_{ij} is the number of copies of the chosen SNP variant at locus i in population j. We estimated the model parameters in a Bayesian framework.

References

1. Ebert D (1994) Virulence and Local Adaptation of a Horizontally Transmitted Parasite. Science 265: 1084-1086.

2. Kidgell C, Winzeler EA (2006) Using the genome to dissect the molecular basis of drug resistance. Future Microbiol 1: 185-199.

3. Kidgell C, Volkman SK, Daily J, Borevitz JO, Plouffe D, et al. (2006) A systematic map of genetic variation in Plasmodium falciparum. PLoS Pathog 2: e57.

4. Anderson TJ, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, et al. (2000) Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. Mol Biol Evol 17: 1467-1482.

5. Hughes AL, Verra F (2001) Very large long-term effective population size in the virulent human malaria parasite Plasmodium falciparum. Proc Biol Sci 268: 1855-1860.

6. Joy DA, Feng X, Mu J, Furuya T, Chotivanich K, et al. (2003) Early origin and recent expansion of Plasmodium falciparum. Science 300: 318-321.

7. Mu J, Duan J, Makova KD, Joy DA, Huynh CQ, et al. (2002) Chromosome-wide SNPs reveal an ancient origin for Plasmodium falciparum. Nature 418: 323-326.

8. Escalante AA, Lal AA, Ayala FJ (1998) Genetic polymorphism and natural selection in the malaria parasite Plasmodium falciparum. Genetics 149: 189-202.

9. Volkman SK, Barry AE, Lyons EJ, Nielsen KM, Thomas SM, et al. (2001) Recent origin of Plasmodium falciparum from a single progenitor. Science 293: 482-484.

10. Mu J, Awadalla P, Duan J, McGee KM, Joy DA, et al. (2005) Recombination hotspots and population structure in Plasmodium falciparum. PLoS Biol 3: e335.

11. Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, et al. (2007) A genome-wide map of diversity in Plasmodium falciparum. Nat Genet 39: 113-119.

12. Jeffares DC, Pain A, Berry A, Cox AV, Stalker J, et al. (2007) Genome variation and evolution of the malaria parasite Plasmodium falciparum. Nat Genet 39: 120-125.

13. Mu J, Awadalla P, Duan J, McGee KM, Keebler J, et al. (2007) Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. Nat Genet 39: 126-130.

14. Mu J, Ferdig MT, Feng X, Joy DA, Duan J, et al. (2003) Multiple transporters associated with malaria parasite responses to chloroquine and quinine. Mol Microbiol 49: 977-989.

15. Trimnell AR, Kraemer SM, Mukherjee S, Phippard DJ, Janes JH, et al. (2006) Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. Mol Biochem Parasitol 148: 169-180.

16. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, et al. (2002) Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature 418: 320-323.

17. Nordborg M, Donnelly P (1997) The coalescent process with selfing. Genetics 146: 1185-1195.

18. Neafsey DE, Hartl DL, Berriman M (2005) Evolution of noncoding and silent coding sites in the Plasmodium falciparum and Plasmodium reichenowi genomes. Mol Biol Evol 22: 1621-1626.

19. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.

20. Smith NG, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415: 1022-1024.

21. Fay JC, Wyckoff GJ, Wu CI (2002) Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415: 1024-1026.

22. Bierne N, Eyre-Walker A (2004) The genomic rate of adaptive amino acid substitution in Drosophila. Mol Biol Evol 21: 1350-1360.

23. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.

24. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

25. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.

26. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498-511.

27. Lewontin RC, Krakauer J (1975) Letters to the editors: Testing the heterogeneity of F values. Genetics 80: 397-398.

28. Cavalli-Sforza LL (1966) Population structure and human evolution. Proc R Soc Lond B Biol Sci 164: 362-379.

29. Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. Heredity 86: 641-647.

30. Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, et al. (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci U S A 88: 839-843.

31. Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, et al. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64: 695-715.

32. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512-517.

33. Morgan AD, Gandon S, Buckling A (2005) The effect of migration on local adaptation in a coevolving host-parasite system. Nature 437: 253-256.

34. Idaghdour Y, Storey J, Jadallah S, Gibson G (2008) A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. PLoS Genetics 4: e1000052.

35. Kibriya MG, Jasmine F, Argos M, Andrulis IL, John EM, et al. (2008) A pilot genome-wide association study of early-onset breast cancer. Breast Cancer Res Treat.

36. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870-874.

37. Han J, Kraft P, Nan H, Guo Q, Chen C, et al. (2008) A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genetics 4: e1000074.

38. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, et al. (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci USA 105: 4340-4345.

39. Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, et al. (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature 448: 591-594.

40. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190.

41. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

42. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

43. Falush D, Stephens M, Pritchard J (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

44. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

45. Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20: 2485-2487.

46. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.

47. Conway DJ, Roper C, Oduola AM, Arnot DE, Kremsner PG, et al. (1999) High recombination rate in natural populations of Plasmodium falciparum. Proc Natl Acad Sci USA 96: 4506-4511.

48. Sidhu AB, Verdier-Pinard D, Fidock DA (2002) Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations. Science 298: 210-213.

49. Su X, Kirkman LA, Fujioka H, Wellems TE (1997) Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant P. falciparum in Southeast Asia and Africa. Cell 91: 593-603.

50. Wellems TE, Walker-Jonah A, Panton LJ (1991) Genetic mapping of the chloroquine-resistance locus on Plasmodium falciparum chromosome 7. Proc Natl Acad Sci USA 88: 3382-3386.

51. Wright SI, Charlesworth B (2004) The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168: 1071-1076.

52. Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

53. Nash D, Nair S, Mayxay M, Newton PN, Guthmann JP, et al. (2005) Selection strength and hitchhiking around two anti-malarial resistance genes. Proc Biol Sci 272: 1153-1161.

54. Gardner MJ, Shallom SJ, Carlton JM, Salzberg SL, Nene V, et al. (2002) Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. Nature 419: 531-534.

55. Hall N, Pain A, Berriman M, Churcher C, Harris B, et al. (2002) Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13. Nature 419: 527-531.

56. Hyman RW, Fung E, Conway A, Kurdi O, Mao J, et al. (2002) Sequence of Plasmodium falciparum chromosome 12. Nature 419: 534-537.

57. Anderson TJ, Nair S, Qin H, Singlam S, Brockman A, et al. (2005) Are transporter genes other than the chloroquine resistance locus (pfcrt) and multidrug resistance gene (pfmdr) associated with antimalarial drug resistance? Antimicrob Agents Chemother 49: 2180-2188.

58. Ma L, G B, Np B, R C, Pa M, et al. (2007) ClustalW and ClustalX version 2.0. Bioinformatics.

59. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725-736.

60. Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5: 704-716.

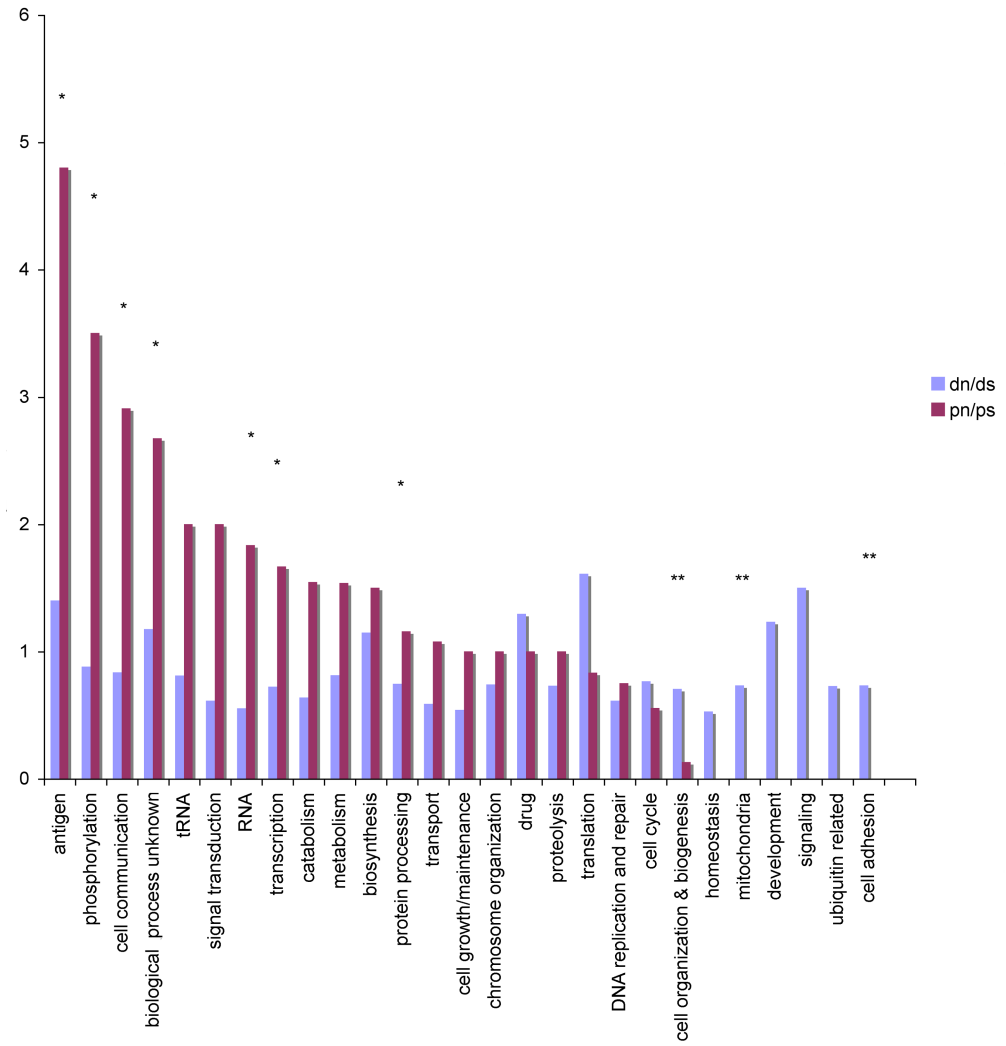
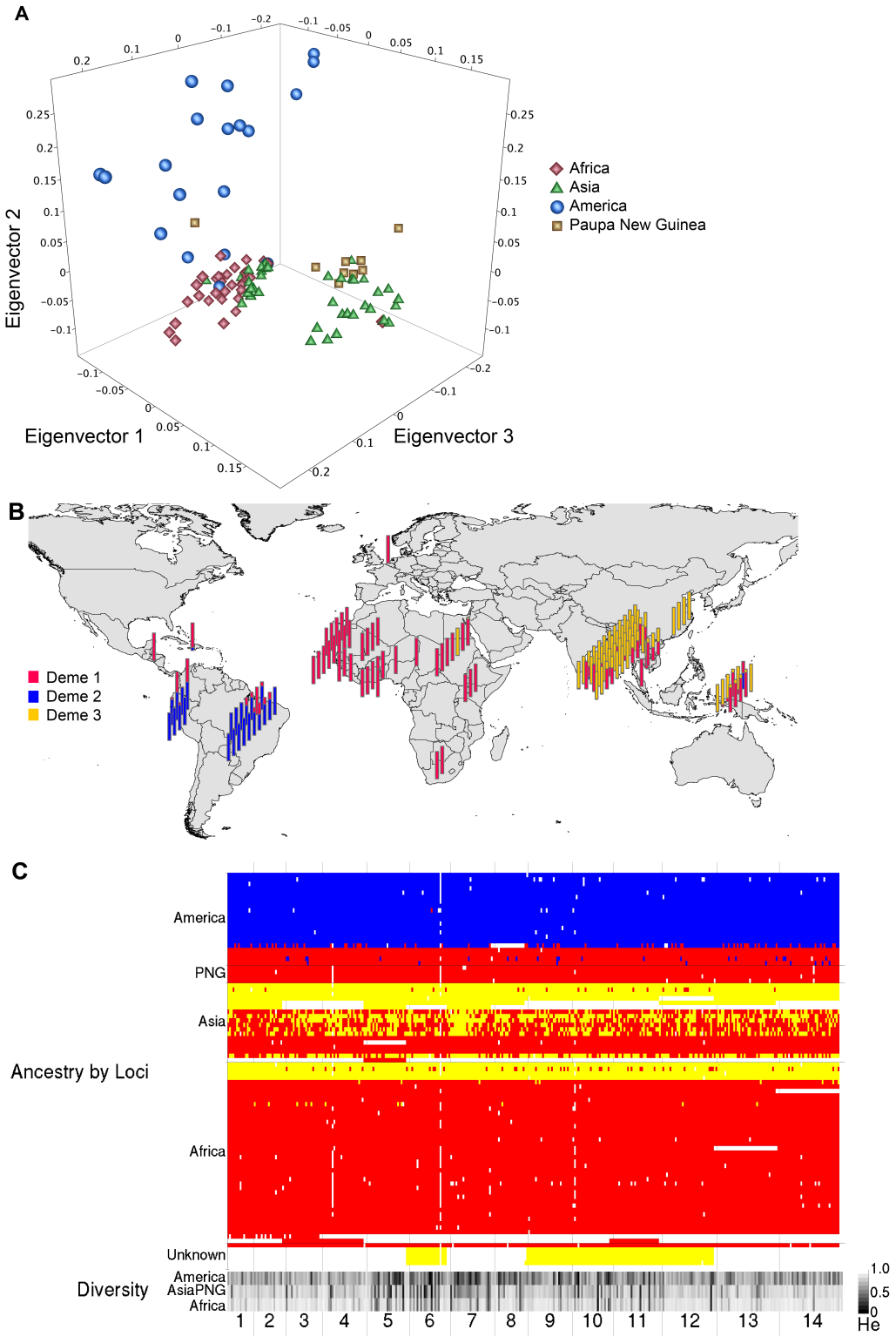


Figure 2-1. The ratios of nonsynonymous and synonymous polymorphisms (*pn/ps*) and divergence (*dn/ds*) for genes grouped by Gene Ontology.
Indicated are groups that are significant (* denotes p value < 0.05, ** denotes p value < 0.01) differentiated at nonsynonymous sites between species using a Fisher's Exact Test (usually indicative of positive selection) and those groups that show a significant excess of nonsynonymous polymorphisms (indicative of and excess of either deleterious - nonsynonymous alleles at low frequency- or balancing selection – nonsynonymous alleles at high frequency).

Figure 2-2. Global Population Structure in *P. falciparum*. **A**) EIGENSOFT inferences of population structure. The top 3 axes of variation from the principle component analysis of random chromosome 3 SNPs. **B**) Population structure for SNP and SSR data combined (N_{Africa} = 40, N_{Asia} = 55, $N_{America}$ = 29, and N_{PNG} = 10) using the admixture model implemented in Structure v.2.2. **C**) The top is ancestry of each SSR locus in each isolate inferred from STRUCTURE v2.2 output. The grey shadings at the bottom is expected heterozygosity (*He*) for the SSR data plotted by loci across the genome.


A
Eigenvector 2
Eigenvector 1
Eigenvector 3
Africa
Asia
America
Paupa New Guinea
B
Deme 1
Deme 2
Deme 3
C
Ancestry by Loci
America
PNG
Asia
Africa
Unknown
Diversity
America
AsiaPNG
Africa
1 2 3 4 5 6 7 8 9 10 11 12 13 14
1.0
0.5
0
He

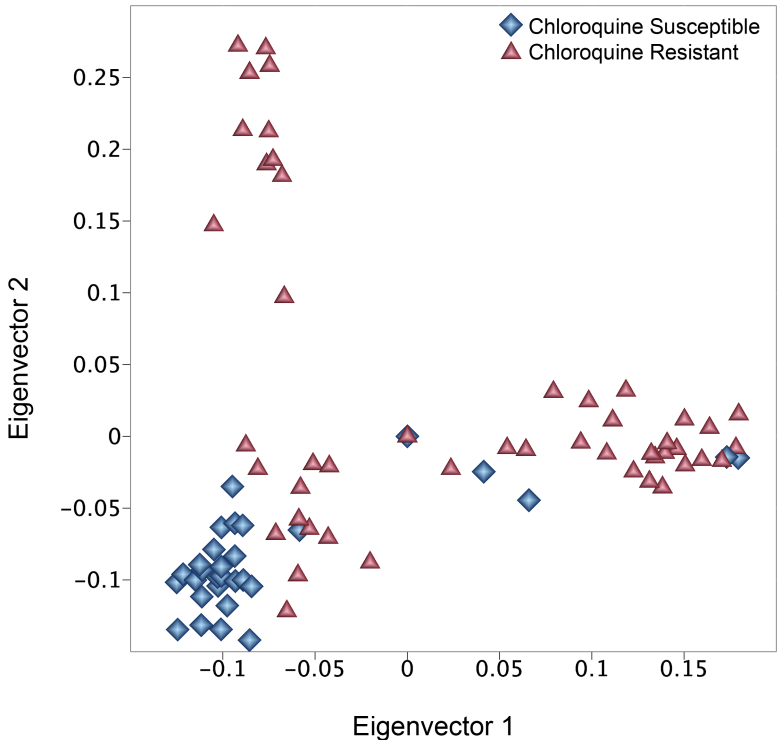


Figure 2-3. EIGENSOFT inferences of population structure.
Indicated are individual isolates exhibiting chloroquine resistance (IC_{50} from 170.3 – 951.6 nM) or susceptibility (IC_{50} from 13.6-76.1 nM)

Table 2-1. McDonald-Kreitman Tests for Departures from Neutrality for the Length of Chromosome 3.

	Dn	***Ds***	***Pn***	***Ps***	***p* Value**
All samples (*n*=99)	279	189	92	40	0.042
Africa (*n*=36)	279	189	65	26	0.035
Asia (*n*=29)	284	193	47	21	0.145
PNG (*n*=11)	284	196	36	14	0.095
South America (*n*=23)	284	193	46	24	0.361

Table 2-2. Frequency of Positively Selected Amino Acids per Gene Ontology Group. Proportions inferred according to Smith and Eyre-Walker (2002).

GO term	**# of genes in GO group**	**Proportion of a.a**
cell cycle	16	0.25
cytoskeleton organization and biogenesis	10	0.34
secondary metabolism	2	0.43
cytoplasm organization and biogenesis	20	0.43
organelle organization and biogenesis	16	0.45
cell organization and biogenesis	24	0.49
electron transport	5	0.62
host-pathogen interaction	5	1.0

Table 2-3. Tajima's *D* by Gene and Population.

	Africa		Asia		America		PNG	
Gene	Tajima's D	p	Tajima's D	p	Tajima's D	p	Tajima's D	p
PFC0050c	-0.1977	0.6860	-1.5325	0.0418	-0.3922	0.6858	-0.7781	0.5814
PFC0060c	-1.2529	0.2578	0.0000	NA	-0.6620	0.4464	-1.1284	0.9372
PFC0070c	0.0968	0.7982	-1.6865	0.0244	-1.6496	0.0444	0.5308	0.5836
PFC0075c	0.0000	NA	-1.7328	0.0038	-1.1608	0.2860	-1.1284	0.9980
PFC0080c	-1.7520	< 0.0001	-0.4303	0.6320	-1.1939	0.1378	0.6714	0.7638
PFC0090w	-0.9900	0.3610	-0.3355	0.7560	-1.3033	0.1674	-1.7910	0.0012
PFC0095c	0.2683	0.5296	0.6129	0.4572	1.4793	0.1410	-1.4295	0.5552
PFC0100c	0.4951	0.4234	-0.7529	0.3306	1.0838	0.2114	0.0000	NA
PFC0125w	0.6948	0.4998	0.1609	0.7992	0.2700	0.6922	-0.6484	0.1774
PFC0135c	0.0000	NA	-1.1493	0.0642	0.0000	NA	0.0000	NA
PFC0165w	-0.6680	0.3472	-0.2880	0.7346	0.5996	0.4900	-1.1139	0.6616
PFC0170c	-0.7279	0.3172	-1.5092	0.0286	0.2752	0.5864	-1.1284	0.9702
PFC0175w	-1.7233	0.0008	0.0000	NA	-1.1608	0.2848	0.0000	NA
PFC0180c	-1.6270	0.0258	1.5339	0.1148	1.4329	0.1052	-0.7781	0.7644
PFC0185w	-1.3024	0.0268	1.4190	0.1338	-1.6789	0.0522	-1.7116	0.6480
PFC0195w	-0.2318	0.5464	-1.5092	0.0300	0.2426	0.6220	-1.1284	0.5614
PFC0215c	-0.8750	0.4058	-1.0805	0.3034	-0.6620	0.4640	-0.7781	0.7766
PFC0220w	-1.1331	0.2566	1.1680	0.2196	0.1859	0.4850	0.0362	0.5984
PFC0230c	-1.5612	0.0694	-1.1505	0.1858	1.2194	0.2454	0.1013	0.4922
PFC0235w	-1.7233	0.1384	0.0000	NA	0.0000	NA	0.0000	NA
PFC0245c	-0.5442	0.5946	-0.7529	0.3316	1.5826	0.0428	-1.1284	0.0672
PFC0250c	-2.0069	< 0.0001	-0.6908	0.5640	0.0000	NA	0.0000	NA
PFC0255c	-1.5612	0.0030	0.0000	NA	0.0799	0.7554	0.0000	NA
PFC0260w	-1.1331	0.1772	-1.1493	0.0702	0.0000	NA	-1.4295	0.3680
PFC0265c	-1.4953	0.0244	-1.1493	0.0728	0.0000	NA	0.3619	0.0704
PFC0270w	0.0000	NA	-1.1493	0.0632	0.0000	NA	0.0000	NA
PFC0285c	-1.1331	0.1434	-1.5092	0.0266	1.0838	0.2162	0.0000	NA
PFC0295c	-1.4953	0.0368	1.4425	0.1450	0.0000	NA	-0.1000	0.8630
PFC0310c	-1.2664	0.0268	-1.1493	0.0690	0.2752	0.5844	-1.1284	0.7914
PFC0325c	-1.0860	0.2142	0.0000	NA	-1.1608	0.2828	0.0000	NA
PFC0330w	-1.1331	0.0676	0.0000	NA	-1.1608	0.2904	-1.1284	0.0496
PFC0335c	-1.1331	0.3686	0.0000	NA	-1.1608	0.2874	0.0000	NA
PFC0340w	-0.7279	0.4426	0.0000	NA	-0.5905	0.6180	0.0000	NA
PFC0345w	0.4807	0.5706	1.1534	0.2630	-1.1608	0.2920	-1.4295	0.0654
PFC0355c	-0.8750	0.2850	-0.3870	0.4828	0.5681	0.4656	-0.2895	0.2334
PFC0370w	-1.8850	< 0.0001	0.0000	NA	-1.1608	0.3016	0.0000	NA
PFC0380w	-0.7014	0.8336	-1.1493	0.0676	-0.6620	0.4500	-1.1284	0.0272
PFC0395w	-1.1331	0.5318	0.0000	NA	0.0000	NA	-1.1284	0.1160
PFC0420w	-1.4953	0.0762	0.0000	NA	0.1859	0.4840	0.0000	NA
PFC0425w	-0.6718	0.4686	-0.3870	0.4846	-0.3674	0.7430	-1.1284	0.0950
PFC0435w	0.0000	NA	-0.3870	0.4778	-1.1608	0.2978	-1.1284	0.3598
PFC0440c	-0.4210	0.6380	-0.7885	0.4210	0.0000	NA	-1.1284	0.6740
PFC0441c	-1.0860	0.1940	1.5949	0.0896	0.0000	NA	1.4427	0.0118
PFC0460w	-1.1946	0.1854	-1.0036	0.3354	-0.1442	0.9086	-0.9316	0.0410
PFC0465c	0.3528	0.3318	-1.1493	0.0746	1.5327	0.0798	-1.1284	0.0376
PFC0485w	-0.9176	0.2078	-1.5092	0.0300	0.1859	0.4672	-1.1284	0.3938
PFC0486c	0.0000	NA	0.0000	NA	-1.1608	0.2888	0.0000	NA
PFC0505c	0.6019	0.6470	1.9755	0.0442	-0.9469	0.3504	1.4379	0.0046
PFC0510w	-0.9006	0.3356	-1.7328	0.0122	-0.6620	0.4390	0.0000	NA
PFC0515c	-1.4953	0.0574	0.1124	0.7516	0.0000	NA	-1.1284	0.0538
PFC0530w	-1.1331	0.2946	-0.7529	0.3224	-1.1608	0.2852	0.0000	NA
PFC0545c	0.0000	NA	-1.1493	0.0790	-0.6620	0.4566	0.0000	NA
PFC0550w	-0.5128	0.4832	0.0000	NA	0.0000	NA	0.0000	NA

Table 2-3. Continued

PFC0560c	-1.4953	0.1582	0.0000	NA	-1.1608	0.2964	0.0000	NA
PFC0580c	0.1486	0.8396	0.0924	0.7590	0.8344	0.2878	0.0000	NA
PFC0600w	-0.8133	0.4262	0.0000	NA	-0.2131	0.6008	0.0000	NA
PFC0615w	-1.4953	0.2292	0.5277	0.3922	0.5351	0.3680	0.0000	NA
PFC0625w	-0.8133	0.2650	0.0000	NA	0.5351	0.3714	-1.1284	0.0212
PFC0630w	-1.4089	0.0058	-1.1493	0.0724	0.0000	NA	0.0000	NA
PFC0640w	0.0700	0.9518	-0.2880	0.7622	-0.8639	0.4156	-1.4295	0.0104
PFC0650w	-1.4953	0.0072	-0.0472	0.9910	0.0000	NA	-1.1284	0.3834
PFC0701w	-0.9006	0.3132	0.0000	NA	-1.1608	0.2840	0.0000	NA
PFC0705c	-1.0732	0.2592	2.0142	0.0362	-0.8639	0.4330	-1.4295	0.0042
PFC0730w	0.0000	NA	0.0000	NA	-1.1608	0.2890	0.0000	NA
PFC0735w	-1.1331	0.1672	0.0000	NA	0.0000	NA	0.0000	NA
PFC0745c	0.4951	0.4188	-0.2480	0.7734	1.2833	0.1598	0.6714	0.7230
PFC0755c	0.0000	NA	0.0000	NA	0.0000	NA	-0.1267	0.6614
PFC0760c	-1.4089	0.0226	-1.1493	0.0760	-0.6401	0.5594	-1.1284	0.0130
PFC0765c	-1.2843	0.0708	-1.1493	0.0718	-0.6361	0.5814	0.0000	NA
PFC0770c	0.0000	NA	0.0000	NA	-1.1608	0.2732	0.0000	NA
PFC0775w	0.0000	NA	0.0000	NA	0.0000	NA	-1.1284	0.6754
PFC0790w	0.0000	NA	0.0000	NA	-0.2131	0.6256	0.0000	NA
PFC0805w	1.1929	0.1286	-0.3870	0.4786	1.4329	0.1094	0.0000	NA
PFC0830w	-0.3937	0.3978	-0.0012	0.9800	-0.6153	0.6122	0.0362	0.3412
PFC0831w	-1.1331	0.2164	-1.1493	0.0894	0.0000	NA	0.6714	0.7762
PFC0835c	0.0000	NA	0.0000	NA	-0.6620	0.4476	0.0000	NA
PFC0840w	-1.2843	0.0752	-0.3870	0.4850	0.8344	0.2814	-1.1284	0.2302
PFC0850c	-1.5612	0.0162	1.5938	0.1114	-1.1608	0.2904	-0.1000	0.2680
PFC0870w	-1.1331	0.3824	0.0000	NA	1.5327	0.0772	0.0000	NA
PFC0880c	-0.6768	0.5092	1.1680	0.2270	-0.2781	0.7162	0.6714	0.0126
PFC0900w	0.0000	NA	0.0000	NA	-1.1608	0.2928	0.0000	NA
PFC0905c	0.3464	0.8462	-0.3870	0.5010	-1.1608	0.2938	-1.1284	0.8058
PFC0910w	1.6290	0.0756	-0.3281	0.7244	-0.6620	0.4528	-0.1000	0.4216
PFC0915w	-0.5128	0.5898	0.0000	NA	0.0000	NA	0.0000	NA
PFC0930c	0.3077	0.8566	-0.9887	0.3636	-1.2600	0.1558	-1.1284	0.9588
PFC0935c	0.0299	0.6142	-0.7885	0.3926	-0.6620	0.4474	1.4427	0.0900
PFC0940c	1.1018	0.1834	-0.7078	0.5006	0.4731	0.5786	2.0449	0.0044
PFC0945w	-1.1331	0.5856	-0.0516	0.6158	-1.1608	0.2850	0.0000	NA
PFC0955w	-0.4213	0.7224	-0.0516	0.5970	0.0000	NA	0.0000	NA
PFC0970w	-0.2318	0.6204	-1.0087	0.3202	1.5826	0.0244	0.0000	NA
PFC0995c	-0.2318	0.6036	0.1924	0.6914	-0.2131	0.6352	-0.1000	0.1394
PFC1000w	-0.7279	0.8500	0.1124	0.7236	-1.1608	0.2800	0.0000	NA
PFC1011c	-0.8133	0.5466	0.0000	NA	-1.1608	0.2662	0.0000	NA
PFC1030w	-1.0860	0.1888	0.5277	0.4104	0.0000	NA	-0.1000	0.9606
PFC1035w	-1.1331	0.3618	0.0000	NA	0.0000	NA	0.0000	NA
PFC1045c	-1.2664	0.1570	-0.3870	0.4844	0.5030	0.5516	0.0000	NA
PFC1055w	-0.6965	0.5328	-0.1085	0.9056	0.1124	0.7658	1.4427	0.0062
PFC1060c	-1.3691	0.2616	-1.8573	0.0132	0.0000	NA	0.2228	0.4838

Table 2-4. Tajima's *D* Across Loci.
Mean (μ),and the Coefficient of Variation (C_v) for Tajima's *D*.

Pop.	μ	$\lvert C_v \rvert$
Africa	-0.80	-1.1
America	-0.25	-4.0
Asia	-0.42	-2.4
PNG	-0.15	-5.7

Table 2-5. Estimates of Genetic Drift.
Mean (and standard deviations) of genetic drift (cj) per population for all synonymous and nonsynonymous sites surveyed throughout chromosome 3 of P. falciparum. The larger values indicate greater departures from the assumed ancestral admixed population. n = sample size.

Population	***n***	**Synonymous**	**Nonsynonymous**
Africa	36	0.269 (0.052)	0.370 (0.040)
America	23	0.217 (0.048)	0.448 (0.042)
Asia	29	0.419 (0.056)	0.370 (0.025)
PNG	11	0.535 (0.057)	0.500 (0.048)

Table 2-6. IC_{50} of ATO

Isolate	ATO IC50	Isolate	ATO IC50	Isolate	ATO IC50
3D7	1.572991	418	0.097255	THY16	0.102545
123/5	1.089802	425	0.07836	THA19	0.110213
128/4	0.612145	Hu425	0.080435	PC49	0.682294
102/1	0.498696	433	0.141312	PC09	4.432078
124/8	0.225995	456	0.299375	PC15	0.934087
REN	0.497985	FAB6	0.709475	PC17	0.805298
106	2.476265	Fab9	0.757012	PC26	0.904983
M2	0.701456	713	0.267092	7G8	0.797188
P13	0.212846	SL/D6	0.738421	DIV17	8.076699
M5	0.568626	LF4/1	0.677747	DIV30	0.959382
S35	0.409527	DD2	1	PAD	1.15635
M24	0.572479	Indo	3.200849	ICS	2.871632
K39	1.277634	Camp	0.118033	ECP	0.575541
KMVII	1.541071	FCB	1.767855	DIV14	3.107774
9013	0.196026	MR80	0.387707	ECU	0.157172
9016	0.110681	V1/S	2.757307	JAV	0.246775
9020	0.344246	JCK	1.397756	HB3	0.815011
9021	0.119735	D5	0.449352	Haiti	0.871613
M97	0.92325	P31	0.180714	D10	1.873546
434	0.047237	TM284	3.561734	PNG2	0.181418
449	0.083717	T2/c6	0.572768	PNG3	0.101868
601	0.107081	TM191c	0.429208	PNG4	0.155626
M190	0.079998	C2A	2.361409	PNG13	0.29489
224	0.260456	MT/S-1	0.284781		

Table 2-7. Significant SNP Associations.
CQ – chloroquine, QN – quinine, Bonferroni corrected p value of 0.05 or less

Gene	SNP	Drug	Pop	*p* Value	Gene Description
MAL7P1_27 (pfcrt)	74	CQ	Africa	< 0.00098	chloroquine resistance transporter, putative
MAL7P1_27 (pfcrt)	75	CQ	Africa	< 0.00098	chloroquine resistance transporter
MAL7P1_27 (pfcrt)	76	CQ	Africa	< 0.00098	chloroquine resistance transporter
MAL7P1_27 (pfcrt)	220	CQ	Africa	< 0.00098	chloroquine resistance transporter
MAL7P1_27 (pfcrt)	271	CQ	Africa	< 0.00098	chloroquine resistance transporter
MAL7P1_27 (pfcrt)	326	CQ	Africa	< 0.00098	chloroquine resistance transporter
MAL7P1_27 (pfcrt)	371	CQ	Africa	< 0.00098	chloroquine resistance transporter
PFC0850c	850	QN	Global	0.0299	Hypothetical protein

Table 2-8. Significant SSR Associations.
CQ – chloroquine, QN – quinine, ATO – atovaquone proguanil, *p* value are Bonferroni corrected. MK *p* values are for concatenated genes within a 40kb window of the associated marker (Jeffares et al. 2007)

Chr	Marker	Drug	Population	*p* Value	Min MK *p* Value	Genes	Descriptions
7	7A11	CQ	Africa	<0.0342	1	*PF07_0024*	Inositol phosphatase, putative
4	c4m48	ATO	Africa	<0.0342	1	*PFD0495c*	hypothetical protein
4	c4m39	ATO	Africa	<0.0342	0.01915	*PFD0705c*	hypothetical protein
7	C13M30	ATO	Africa	0.03434	1	*MAL7P1.50*	erythrocyte membrane protein 1
5	m6	QN	AsiaPNG	0.0343			
11	B5M124	ATO	AsiaPNG	0.0343	0.1964	*PF11_0450*	hypothetical protein

Table 2-9. SNPs with Nominally Significant Associations in All Three Populations.

Gene	Snp	Drug	Africa	Asia	America	Gene Description	MK *p* value
MAL7P1.27	76	CQ	< 4.000e-6	>0.05	0.02394	Chloroquine resistance transporter, putative	0.5304
MAL7P1.27	220	CQ	< 4.000e-6	>0.05	0.02394	Chloroquine resistance transporter, putative	0.5304
PFA0590w	1318	CQ	0.006117	>0.05	0.003092	ABC transporter, putative	1

Table 2-10. SSR with Nominally Significant Associations in Two or More Populations.
*****Minimum p value McDonald Kreitman test.

Marker	**Drug**	**Africa**	**Asia**	**America**	**Gene**	**Description**	**MIN MK *p* value***
B7M97/B	ATO	0.0417	0.0008	0.9706	*PFA0245w*	hypothetical protein	0.5773
c3m64	ATO	0.0254	0.0122	0.5388	*PFC0475c*	hypothetical protein, conserved	0.183
c4m30/Y	ATO	0.0216	0.0248	0.439			
c4m39-1	ATO	< 0.0001	0.007	0.6057	*PFD0705c*	hypothetical protein	0.01915
c4m39-1	QN	0.0054	0.0031	0.0608	*PFD0705c*	hypothetical protein	0.01915
c4m39-2	QN	0.4042	0.0204	0.0119	*PFD0705c*	hypothetical protein	0.01915
B5M123/Y	CQ	0.0312	0.0414	0.274	*PFF1400w*	hypothetical protein	1
B5M24/G	CQ	0.7703	0.0091	0.0334			
7A11	CQ	< 0.0001	0.0065	0.2003	*PF07_0024*	hypothetical protein	1
9B12/Y	CQ	0.0062	0.0014	0.2022	*PF07_0022*	hypothetical protein	1
B5M100/Y	ATO	0.0105	0.1506	0.0115	*MAL7P1.224*	hypothetical protein, conserved in P. falciparum	1
B5M47/Y	CQ	0.0397	0.0014	0.0169	*PF07_0021*	hypothetical protein	1
B5M77	CQ	0.0372	0.0316	0.2351	*PF07_0018*	hypothetical protein	1
B5M97/G	CQ	0.0149	0.0227	0.1368	*PF07_0018*	hypothetical protein	1
BM7/B	CQ	0.3124	0.0085	0.0292	*MAL7P1.56*	erythrocyte membrane protein 1 (PfEMP1)	1
PE14F/G	CQ	0.8003	0.05	0.0334	*MAL7P1.204*	hypothetical protein	0.6772
PS590/G	CQ	0.0108	0.0003	0.3812	*MAL7P1.21*	origin recognition complex subunit, putative	1
C9M11/B	QN	0.0347	0.0269	0.7315			
C9M26	ATO	0.0184	0.0426	0.2993	*PFI1710w*	cytoadherence-linked protein	0.3

Table 2-10 Continued

C9M33	ATO	0.0018	0.5895	0.0246	*PFI1650w*	DNA excision-repair helicase, putative	1
B7M78/B	ATO	0.0464	0.0109	0.1986	*PF10_0091*	hypothetical protein	1
TA117	CQ	0.6933	0.0207	0.0342	*PF11_0464*	hypothetical protein	0.5294
TA31/G	ATO	0.0459	0.1263	0.0212	*PFI1_0185*	hypothetical protein	0.3154
XB8M18/Y	CQ	0.4148	0.0148	0.036			
C12M62	ATO	0.0362	0.0184	0.9254			
C12M89	QN	0.0415	0.7326	0.0274	*PFL1755w*	hypothetical protein	1
C13M5	CQ	0.3394	0.0067	0.0043	*MAL13P1.296*	hypothetical protein	1
C14m58	CQ	0.177	0.0417	0.0009	*PF14_0233*	hypothetical protein	1
C14M87	ATO	0.03	0.0157	0.641	*PF14_0589*	valine - tRNA ligase, putative	0.1667

Table 2-11. SNPs with Nominally Significant Associations in the Global Population.

Gene	SNP	Drug	P value
MAL7P1.27	76	CQ	0.018992
PF08_0078	440	QN	0.008468
PF14_0133	1740	QN	0.02436
PFC0090w	90	CQ	0.015772
PFC0090w	90	CQ	0.015772
PFC0090w	90	CQ	0.018344
PFC0125w		CQ	0.03248
PFC0125w		CQ	0.02858
PFC0125w		CQ	0.004484
PFC0165w	165-3	ATO	0.014732
PFC0295c	153	QN	0.011728
PFC0295c	295-1	QN	0.005832
PFC0355c	355-2	QN	0.0426
PFC0505c	505-NEW	ATO	0.00588
PFC0505c	505-3	QN	0.033496
PFC0580c	580-2	QN	0.021464
PFC0615w	615	CQ	0.013036
PFC0615w		CQ	0.0161
PFC0650w	650	CQ	0.013992
PFC0745c	745	ATO	0.021632
PFC0831w	831	CQ	0.007152
PFC0850c	850-2	QN	0.000124
PFC0850c	850-2	ATO	0.019864
PFC0850c	850	CQ	0.013676
PFC0940c	940	QN	0.014244
PFC0955w	955	CQ	0.01794
PFE1150w	1034	QN	0.004668
PFE1150w	1034	ATO	0.04204
PFE1150w	1042	QN	0.017488

Chapter 3 Genome-wide Positive Selection, Recombination Hotspots, and Loci Associated with *Plasmodium falciparum* Resistance to Antimalarial Drugs

Jianbing Mu[1], Rachel A. Myers[2,3], Hongying Jiang[1], Shengfa Liu[1,4], Stacy Ricklefs[5], Michael Waisberg[6], Kesinee Chotivanich[7], Polrat Wilairata[7], Srivicha Krudsood[7], Nicholas J. White[8], Rachanee Udomsangpetch[9], Liwang Cui[10], May Ho[11], Fengzheng Ou[12], Haibo Li[12], Jiangping Song[12], Guoqiao Li[12], Xinhua Wang[13], Suon Seila[14], Sreng Sokunthea[14], Duong Socheat[14], Daniel E. Sturdevant[5] , Stephen F. Porcella[5], Rick M. Fairhurst[1], Thomas E. Wellems[1], Philip Awadalla[2] & Xin-zhuan Su[1]

[1]Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA; [2]Department of Pediatrics, University of Montreal, Faculty of Medicine, Ste. Justine Research Centre, Montreal, Quebec, Canada H3T 1C5; [3]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27606, USA; [4]School of Life Sciences, Xiamen University, Xiamen, Fujian, The People's Republic of China; [5]Genomics Unit, Research Technologies Section, RTB, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT 59840, USA; [6]Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA; [7]Department of Clinical Tropical Medicine, Faculty of Tropical Medicine,

Mahidol University, Bangkok 10400 Thailand; [8]Wellcome-Trust-Mahidol University-Oxford Tropical Medicine Research Programme, Mahidol University, 420/6 Rajvithi Road, Bangkok 10400, Thailand; [9]Pathobiology Department, Faculty of Science, Mahidol University, 420/6 Rajvithi Road, Bangkok 10400, Thailand; [10]Department of Entomology, The Pennsylvania State University, 501 ASI Building, University Park, PA 16802; [11]Department of Microbiology and Infectious Disease, University of Calgary, Calgary, Alberta, T2N1N4 Canada; [12] Research Center for Qinghao, Guangzhou University of Chinese Medicine, Guangzhou, People's Republic of China; [13]Guangzhou University of Chinese Traditional Medicine, Guangzhou, People's Republic of China; [14]National Centre for Parasitology, Entomology and Malaria Control, Phnom Penh, Cambodia.

Published in Nature Genetics, 42, 268-271 (31 January 2010).

Abstract

Antimalarial drugs impose strong pressure on *Plasmodium falciparum* parasites and leave signatures of selection in the parasite genome [1,2]. Search for signals of selection may lead to genes encoding drug or immune targets [3]. The lack of high-throughput genotyping methods, inadequate knowledge of parasite population history, and time-consuming adaptations of parasites to *in vitro* culture have hampered genome-wide association studies (GWAS) of parasite traits. Here we report genotyping of DNA from 189 culture-adapted *P. falciparum* parasites using a custom-built array with thousands of single nucleotide polymorphisms (SNPs). Population structure, variation in recombination rate, and loci under recent positive selection were detected. Parasite IC_{50} values to seven antimalarial drugs were obtained and used in GWAS to identify genes associated with drug responses. The SNP array and genome-wide parameters provide valuable tools and information for new advances in *P. falciparum* genetics.

Introduction

Drug resistance in *P. falciparum* parasites has evolved and spread rapidly, leading to the loss of chloroquine (CQ) and sulfadoxine-pyrimethamine (SP) as first-line treatments in most endemic areas. Resistance to all antimalarial drug classes has been reported, including recently the artemisinin (ART) derivatives [4-7]. Mutations in the *P. falciparum* CQ resistance transporter gene (*pfcrt*) and the genes encoding dihydrofolate reductase (*pfdhfr)* and dihydropteroate synthase (*pfdhps)* have been shown to confer

resistance to CQ and SP, respectively. Additionally, copy number and/or point mutations at the gene encoding a homolog of human P-glycoprotein (*pfmdr1*) on chromosome 5 have been associated with parasite response to mefloquine (MQ), quinine (QN), ART, and other antimalarial drugs, although other unknown genes may have roles in the responses [8]. *P. falciparum* resistance to antimalarial drugs has occurred only since widespread deployment of the drugs (i.e. within the past 60 years), and there may have not been enough time for recombination to break down completely linkages between causal alleles and nearby genetic markers. Indeed, by scanning for regions of high LD, the chromosome segment carrying the *pfcrt* locus was correctly identified using 342 genome-wide microsatellite (MS) markers and 92 parasite isolates collected from different parts of the world [1]. Here we report the first genome-wide *P. falciparum* maps of population recombination events, signatures of recent positive selection, and GWAS of multiple drug resistant phenotypes and SNP genotypes obtained using a custom-built SNP typing microarray.

Results

We collected and adapted 189 independent *P. falciparum* isolates into *in vitro* culture, including 146 from the Asia (Thailand and Cambodia), 26 from Africa, 14 from America, and 3 from Papua New Guinea. We developed a custom 3K oligo probe array based on the molecular inversion probe (MIP) technology (Affymetrix Inc, Santa Clara, CA) [9] to interrogate 3354 SNPs we identified previously [3]. The MIP array provides a

simple and reliable method to genotype the 23 megabase (mb) *P. falciparum* genome with a coverage averaging ~one SNP per 7 kilobase (kb). Among the 3257 (97.1%) SNPs called, 2763 (82.4%) had call rate >90%, and only seven were different from those the sequenced 3D7 genome sequence (0.2%). One thousand, eight hundreds, and eighty-nine (58.3%) SNPs had a minor allele frequency (MAF) greater then 2% among all the parasites; 1216 (37.3%) SNPs had MAF >2% in the Asian population; 1637 (50.3%) SNPs had MAF >2% in the African population; and 813 (24.9%) had MAF >2% in the American populations.

We tested for genetic heterogeneity that may be associated with geography. STRUCTURE analyses [10] showed that the parasites could be clustered into continental populations, with a group of Cambodian parasites separated from the majority of the those of Thailand and Cambodia (Figure 3-1a). Similarly, principal component analysis (PCA) using EIGENSOFT[11] identified significant axes of variations partitioning the parasites into clusters of Asia, Africa, America (Figure 3-1b) as well as distinct groups of parasites from Thai-Cambodian regions (Figure 3-1c). The clustering of Cambodian parasites, which were collected from sites within a radius of ~50 km, into different groups suggests either a recent population admixture or possibly the presence of SNPs that could distinguish parasites with different phenotypes. These population clusterings were corroborated with Wright's *Fst* values (Africa *vs.* Asia 0.054; Africa *vs.* America 0.136; Asia *vs* America 0.028, and between the two Cambodian populations 0.254). The

large *Fst* value for the Cambodian populations was due to fixation of ~75% SNPs (765/1024) in the outlier Cambodian population.

Using genome-wide SNPs, we generated population recombination maps for all 14 chromosomes. Interestingly, the five largest chromosomes (9-14) had relatively fewer recombination events than the smaller chromosomes (Figure 3-2). Similar to those observed on chromosome 3 [12], many recombination hot- or cold-spots appeared to be 'conserved' among populations. There were several loci with extremely high levels of recombination activity, including a locus at one end of chromosome 1 and a segment on chromosome 7 containing *pfcrt* (from 400 to 800 kb) that had a mosaic recombination pattern. The chromosome 7 recombination hotspots flanked a central 100 kb segment (containing *pfcrt*) with a reduced recombination activity, suggesting a recent selective sweep. In contrast, balancing selection on the nearby *var* and other genes may favor higher rates of allelic exchange.

We mapped chromosomal loci potentially under selection using relative extended haplotype homozygosity (REHH) [13,14], integrated haplotype score (iHS) [15], and cross population extended haplotype homozygosity (XP-EHH) [14,16]. We generated genome-wide maps of selection for parasite populations from Asia, Africa, and America, separately, and detected many loci that were under significant positive selection (Figure 3-3). Examples of recent positive selection from REHH included the locus on chromosome 7 containing *pfcrt*, a locus on chromosome 11 containing *pfama-1*, and a locus on chromosome 13 containing an ABC transporter (PF13_0271) (Figure 3-3a). The

pfcrt gene is under CQ selection [1]; *pfama-1* is a target of host immune response [17]; and the gene encoding the ABC transporter on the chromosome 13 was predicted to transport iron into mitochondrion at PlasmoDB. Other signals evident in Figure3-3a were likely to represent regions containing genes for either antigens and/or putative transporters that may be under immune or drug selection pressures.

Similarly, iHS detected strong selection signals at the *pfcrt and pfama-1* loci (Figure 3-3b), consistent with REHH results. Additional interesting iHS signals included PFA0655w (encoding a member of SURFIN) [18] on chromosome 1 and a putative metabolite/drug transporter (PF14-0260) on chromosome 14. If we used an iHS score of 2.3 as a significant cutoff value (approximately top 1% of theoretical iHS distribution), we identified many SNPs that were also identified using REHH (data not shown). We also performed XP-EHH to detect selective sweeps that drive some alleles to fixation in one population but remains polymorphic in others. Indeed, many extended haplotypes were detected between populations (Figure 3-3c). Again, the *pfcrt* locus had highly significant *P*-values, particularly in comparison of African *v.s.* American (AF/AM) and African *v.s* Asian (AF/AS) populations. Another gene with very significant XP-EHH *P*-value was PFE1445c on chromosome 5 that encoded a *Plasmodium* conserved protein (Figure 3-3). There were also several large extended haplotypes (519126-922368 bp on chromosome 7, 831749-925515 bp on chromosome 8, 319075-495408 bp on chromosome 9) between African and American populations. A total of 11 genes under significant selection were detected by all the three methods (Table 3-1), although the

signatures at the chromosome 7 locus may be due to selective sweeps and hitchhiking [1]. Other genes such as PFC0940c and PFE1445c were highly polymorphic with predicted transmembrane domains and were likely conserved antigen genes in *Plasmodium*.

To detect genes associated with drug responses, we measured half maximum inhibitory concentrations (IC_{50}) of CQ, QN, MQ, SP, dihydroartemisinin (DHA), amodiaquine (AMQ), and piperaquine (PQ) from 185 culture-adapted parasites using a SYBR green method [19] (Figure 3-4a) and conducted GWAS. Except for CQ and SP that had bimodal distributions of IC_{50} values, the distributions of IC_{50} values for the other five drugs were more unimodal. All the parasites were sensitive to PQ and DHA. The range of IC_{50} values for PQ was small (5 folds) while the IC_{50} range for SP was large (~56,000 folds). The IC_{50} ranges for the other drugs were 10-fold or higher (16 fold for DHA; 17 fold for AMQ; 26 fold for QN; 34 fold for MQ; 70 fold for CQ). Parasites from the Thai-Cambodian population had similar distributions of IC_{50} values to those of the worldwide population, except that there were only 2 and 6 (out of 143) parasites that were sensitive to CQ and SP, respectively (Figure 3-4b). We also compared the IC_{50} values of the parasites from the two genetically distinct Cambodian populations and found that the average IC_{50} for the all drugs were not significantly different (unpaired *t*-test; data not shown).

Delayed parasite clearance following artesunate treatment or artemisinin combination therapy (ACT) has been reported from patients at the Thai-Cambodian border [6,7]. The Cambodian parasites did have a significantly higher mean IC_{50} value to

DHA (5.2±1.5 nM) compared to the parasites from Thailand (2.0±1.0 nM) and America (2.5±1.1 nM) (*t*-tests, $P<0.001$), but not from Africa (3.1±2.4 nM) ($P=0.09$). As reported previously [20,21], multivariate analyses showed a strong positive correlation ($R^2= 0.78$) between IC_{50} values of MQ and DHA, some positive correlations between CQ IC_{50} values and those of SP ($R^2= 0.47$), AMQ ($R^2= 0.52$), and QN ($R^2= 0.52$), and slight negative relationships between DHA/AMQ, MQ/AMQ, CQ/MQ, and CQ/DHA among all the parasites (Figure 3-4c) and those from the Thai-Cambodian population (Figure 3-4d). The strong positive correlation between the responses to DHA and MQ suggests either co-selection by the drugs and/or a common resistance mechanism. This association may also be partly explained by *pfmdr1* amplification.

We performed GWAS on individual populations using PLINK [22] and EIGENSOFT (Figure 3-5 and Table 3-2). Quantile-Quantile plots suggested effective correction of potential population structure (Figure 3-6). Although several genes were associated with responses to CQ, QN, DHA and MQ, only MAL7P1.27-9 (*pfcrt*), PFA0665w-18 (*pfsurfin*) and PFE1150w-4 (*pfmdr1*) had a minor allele frequency higher than 15%. All of the three genes were also under positive selection. The association of *pfcrt* with CQ response is well established [23]. Likewise, the association of *pfmdr1* with QN is consistent with the linkage of QN response to polymorphisms in the gene [24] and with altered QN IC_{50} values in parasites engineered to have wild type *pfdmr1* allele replaced with a mutant allele [25]. Association of PFA0665w (SURFIN) with responses to antimalarial drugs has not been reported previously. SURFIN was reported to be co-

transported with PfEMP1 and RIFIN to the infected erythrocyte surface [18] and could be part of a protein complex involved in binding or transport chemical compounds. There were also two *Plasmodium* conserved genes (PF11_0079 and PFC0460w) with significant *P*-values from both EIGENSOFT and PLINK. However, the associations of some of these candidate genes could be due to linkage to genes nearby that might be the real actors. The functions of these candidate genes in the associated loci and their contributions to antimalarial drug resistance require further studies.

Conclusions

Our *in vitro* assays suggest that *P. falciparum* strains from different continents remain sensitive to DHA and PQ, although parasites from Cambodia are generally more resistant to the drugs. Many genes under recent positive selection were identified, some of which could be drug or immune targets. The candidate genes associated with responses to the antimalarial drugs require further verification due to small parasite sample size and low minor allele frequencies. Gene copy number variation has been reported to contribute to parasite drug response [26,27] and need to be investigated too. The high throughput MIP array, estimates of genome-wide recombination events and recent positive selection maps provided important tools and information for GWAS to identify genes controlling various malaria traits.

Methods

Parasite collection

All the parasites used in this study were culture-adapted clonal lines collected from 23 different countries. Some of the Asian parasites and all the parasites from Africa, America, and PNG were described previously [28,29]. Thirty-four parasites from Cambodia were collected in a clinical study approved by the IRBs of the National Institute Allergy and Infectious Diseases, USA; the Ministry of Health of the Kingdom of Cambodia; and the Guangzhou University of Chinese Medicine, Guangzhou, People's Republic of China, with informed consent obtained from all subjects. The identity and clonality of the parasites were verified using multiple microsatellites before drug assays.

DNA extraction and SNP genotyping using MIP array

Parasite culture and genomic DNA extraction were as described [30]. Genomic DNA isolated from *Plasmodium falciparum* grown in culture was genotyped using the custom designed 3K Malaria Panel (Affymetrix Inc, Santa Clara, CA). Samples were prepared with the Malaria 3K Panel following the GeneChip® Scanner 3000 Targeted Genotyping System Protocol and hybridized to Universal 3K Tag arrays (Affymetrix Inc, Santa Clara, CA). The only modification in the assay protocol was to normalize samples to a starting concentration of 65ng/μL that equates to a total gDNA input of 871ng. Following hybridization and scanning, genotypes were assigned using the GeneChip® Targeted Genotyping Analysis Software (Affymetrix Inc.) with the following changes to

the default clustering parameters: MinHetToHalfRatio=0.5, and MinAssayCallRate=90. Genotypes were scored and stored in Excel sheets for further analyses.

Drug assays and IC_{50} calculation

Drug assays were performed as described previously [30,31]. To ensure high quality of phenotypic data, we repeated all drug assays at least 3 times independently using the same drug stock solutions. CQ, QN, MQ and DHA were purchased from Sigma-Aldrich (St. Louis, USA); SP was obtained from Roche (Indianapolis, USA); AMQ was bought from LGC Promochem (UK), and PQ was obtained from Guangzhou University of Traditional Chinese Medicine, China. The same stock solution for each drug (10mM in ethanol, except SP in dimethylsulfoxide) was used in all drug assays. The 3D7 parasite was included in all drug assays as a control for plate-to-plate variation.

Structure, Fst and principal component analysis

We applied PCA, a Bayesian clustering approach, as implemented in the program EIGENSOFT [11] and STRUCTURE (v2.2) [10], respectively, and *Fst* to investigate potential population structure. We used Wright's population differentiation estimator *Fst* to ensure ploidy independence. To run the STRUCTURE program, we applied the same conditions described previously [12]. Briefly, ten runs of 50,000 burn-ins and 100,000 iterations were performed for K=1 to 10 using the admixture model. For PCA, we used the LD correction and calculated the top 10 eigenvectors or principal components (PCs) from the

genotypes of the African, Asian, and American populations. We identified and removed isolates that were greater than 6 standard deviations from the PC mean along any of the top 5 PCs and repeated the PCA calculation and outlier detection for 10 iterations.

Estimate of recombination events

Nonparametric estimates of the number of recombination event (Rh) were calculated using the Myers and Griffiths method as described previously [12]. The 14 chromosomes were analyzed individually for African, Asian, American and the Cambodian populations.

Detection of recent positive selection

We used long-range haplotype (LRH) and integrated haplotype score (iHS) to detect loci under recent natural selection in parasite genome [13,15]. For LRH analysis, we compare the REHH extending 100kb in both directions from a core SNP. For iHS**,** extended haplotype homozygosity (EHH) was calculated with a window size of 10 SNPs in each direction from the core SNP, and then EHH was integrated using physical distance resulting in the integrated EHH (iHH) for each allele at the core SNP. The log ratio of the major allele iHH to minor allele iHH was taken and, conditioning on minor allele frequency, standardized to have mean = 0 and variance = 1, resulting in the iHS score for the core SNP. Theoretical cutoffs for the 1% of signals genome-wide was considered as strong signals to indicate candidate selection regions. Isolates from three

different geographic locations were tested separately. For XP-EHH analysis, we calculated EHH and the log ratio iHH for the pair-wise tests of the African, Asian and American populations as described [16]. The log ratios were standardized to have mean 0, variance 1, and assigned *P* values assuming a normal distribution. SNPs with *P*-values less than 0.05 where considered strong signals.

Genome-wide association analysis

The individual populations were analyzed for association to the seven antimalarial drugs using EigenstratQTL in the EIGENSOFT program, utilizing PCA to control for population structure within the populations. Population structure was corrected using three, one, and zero significant PCs in the PCA for the Asian, African, and American populations, respectively. The correction is a function of sample position and the regression of genotypes at PC position for that sample, which adjusted genotypes and phenotypes and effectively eliminated population structure within each individual population. The correction for the genotype of sample i at SNP j is: $g_{ij,adjusted} = g_{ij} - y_i a_j$ $y_i = \frac{\sum_j a_j g_{ij}}{\sum_j a_j^2}$

Where a_j is the ancestry/position of individual j in the PC. Test statistic is (N – K)* correlation (corrected genotypes, corrected phenotypes)^2, where N = number of isolates (N = 133), and K = number of PCs used for correction (K = 3). The correlation between corrected genotypes and corrected phenotypes were obtained with the top 3 PC's as fixed

effects. Nominal *P*-values were determined using the Chi-sq distribution, df =1. Bonferroni *P*-values were determined as 1-(1-nominal *P*-value)^number of successful tests.

Association analysis was also performed using software PLINK [22]. Because PLINK does not have PCA correction within its test, population outliers from PCA analysis (those outside the circle in Fig 1c) were removed before association analysis. A linear regression was fitted to test for each SNP for its association with *in vitro* IC_{50} values of the seven antimalarial drugs. Significant SNPs ($P<0.05$) were determined after Bonferroni correction. Quantile-Quantile plots for both methods were obtained by contrasting uncorrected and corrected (if applicable) experimental *P value* distributions to the expected uniform 0 to 1 distribution.

Web Resources

PlasmoDB, http://plasmodb.org/plasmo/;

PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/;

EIGENSOFT, http://helix.nih.gov/Applications/eigensoft.html;

Acknowledgements

This work was supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, and funds from

the Canadian Institute of Health Research #11284, National Academies Keck Genome Initiative, and the Human Frontiers in Science Program #RGP54/2006 for P.A. KC and NW are supported by the Wellcome Trust. S.L. was supported by the 973 National Basic Research Program of China, #2007CB513103. We also thank Dr. Jetsumon Sattabongkot for help in parasite shipping; Ms. Josephine Dunn and Mr. Louie Zhang for assistance in parasite culture; and NIAID intramural editor Brenda Rae Marshall for assistance.

Author Contributions

J.M. parasite culture and DNA extraction, drug assay, data analysis, writing; R.A.M statistical design, data analysis and writing; H.J. parasite collection, culture, and drug assay; S.L. parasite culture and drug assay; S.R., D.E.S. and S.P. array development and genotyping; M.W. drug assay software; K.C., P.W., S.K., N.J.W., R.U., L.C., M.H., F.O., H.L., J.P., X.W., G.L, S. Seila, S. Sokunthea, D. S., field studies and parasite collection; R.M.F. and T.E.W., field work and writing; P.A. statistical analysis and design, and writing; X-z. S. project design, data analysis, and writing.

References

1 Wootton, J. C. *et al.* Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. *Nature* 418, 320-323 (2002).

2 Roper, C. *et al.* Intercontinental spread of pyrimethamine-resistant malaria. *Science* 305, 1124 (2004).

3 Mu, J. *et al.* Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. *Nat. Genet.* 39, 126-130 (2007).

4 Zalis, M. G., Pang, L., Silveira, M. S., Milhous, W. K. & Wirth, D. F. Characterization of Plasmodium falciparum isolated from the Amazon region of Brazil: evidence for quinine resistance. *Am. J. Trop. Med. Hyg.* 58, 630-637 (1998).

5 Baird, J. K. Effectiveness of antimalarial drugs. *N. Engl. J. Med.* 352, 1565-1577 (2005).

6 Wongsrichanalai, C. & Meshnick, S. R. Declining artesunate-mefloquine efficacy against falciparum malaria on the Cambodia-Thailand border. *Emerg. Infect. Dis.* 14, 716-719 (2008).

7 Noedl, H. *et al.* Evidence of artemisinin-resistant malaria in western Cambodia. *N. Engl. J. Med.* 359, 2619-2620 (2008).

8 Hayton, K. & Su, X.-z. Drug resistance and genetic mapping in Plasmodium falciparum. *Curr. Genet.* (2008).

9 Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* 21, 673-678 (2003).

10 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* 155, 945-959 (2000).

11 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904-909 (2006).

12 Mu, J. *et al.* Recombination hotspots and population structure in Plasmodium falciparum. *PLoS Biol* 3, e335 (2005).

13 Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832-837 (2002).

14 Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913-918 (2007).

15 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol* 4, e72 (2006).

16 Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19, 826-837 (2009).

17 Escalante, A. A., Lal, A. A. & Ayala, F. J. Genetic polymorphism and natural selection in the malaria parasite Plasmodium falciparum. *Genetics* 149, 189-202 (1998).

18 Winter, G. *et al.* SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes. *J. Exp. Med.* 201, 1853-1863 (2005).

19 Liu, S., Mu, J., Jiang, H. & Su, X.-z. Effects of Plasmodium falciparum mixed infections on in vitro antimalarial drug tests and genotyping. *Am. J. Trop. Med. Hyg.* 79, 178-184 (2008).

20 Brockman, A. *et al.* Plasmodium falciparum antimalarial drug susceptibility on the north-western border of Thailand during five years of extensive use of artesunate-mefloquine. *Trans. R. Soc. Trop. Med. Hyg.* 94, 537-544 (2000).

21 Basco, L. K. & Le Bras, J. In vitro activity of artemisinin derivatives against African isolates and clones of Plasmodium falciparum. *Am. J. Trop. Med. Hyg.* 49, 301-307 (1993).

22 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559-575 (2007).

23 Fidock, D. A. *et al.* Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* 6, 861-871 (2000).

24 Ferdig, M. T. *et al.* Dissecting the loci of low-level quinine resistance in malaria parasites. *Mol. Microbiol.* 52, 985-997 (2004).

25 Sidhu, A. B., Valderramos, S. G. & Fidock, D. A. pfmdr1 mutations contribute to quinine resistance and enhance mefloquine and artemisinin sensitivity in Plasmodium falciparum. *Mol. Microbiol.* 57, 913-926 (2005).

26 Cowman, A. F., Galatis, D. & Thompson, J. K. Selection for mefloquine resistance in Plasmodium falciparum is linked to amplification of the pfmdr1 gene and cross-resistance to halofantrine and quinine. *Proc. Natl. Acad. Sci. U. S. A.* 91, 1143-1147 (1994).

27 Dharia, N. V. *et al.* Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in Plasmodium falciparum. *Genome Biol* 10, R21 (2009).

28 Su, X.-z., Kirkman, L. A., Fujioka, H. & Wellems, T. E. Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant P. falciparum in Southeast Asia and Africa. *Cell* 91, 593-603 (1997).

29 Mu, J. *et al.* Multiple transporters associated with malaria parasite responses to chloroquine and quinine. *Mol. Microbiol.* 49, 977-989 (2003).

30 Jiang, H. *et al.* Genome-wide compensatory changes accompany drug- selected mutations in the Plasmodium falciparum crt gene. *PLoS ONE* 3, e2484 (2008).

31 Raj, D. K. *et al.* Disruption of a plasmodium falciparum multidrug resistance-associated protein (PFMRP) alters its fitness and transport of antimalarial drugs and glutathione. *J. Biol. Chem.* (2008).

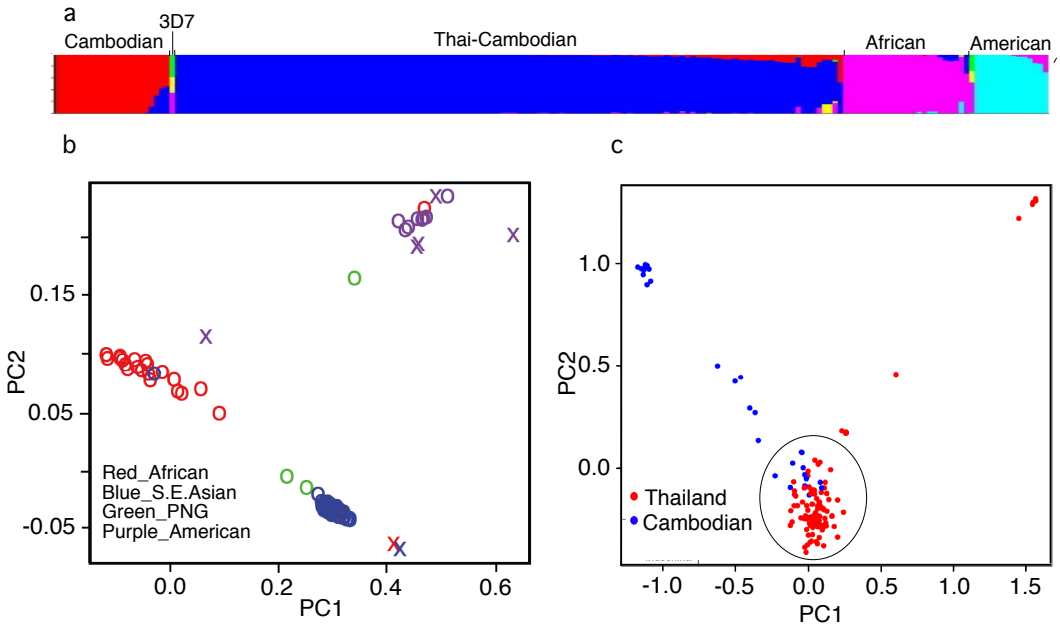
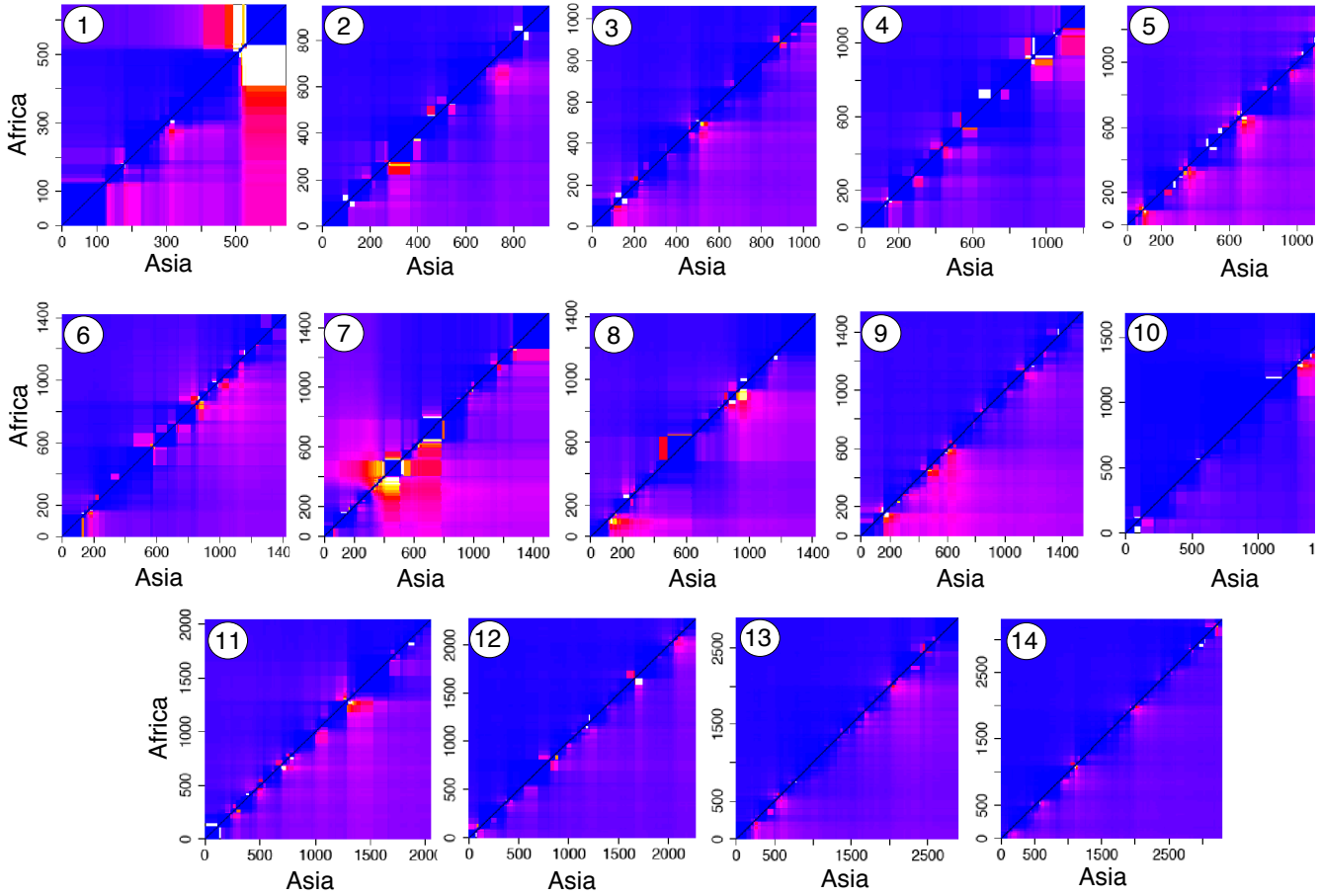


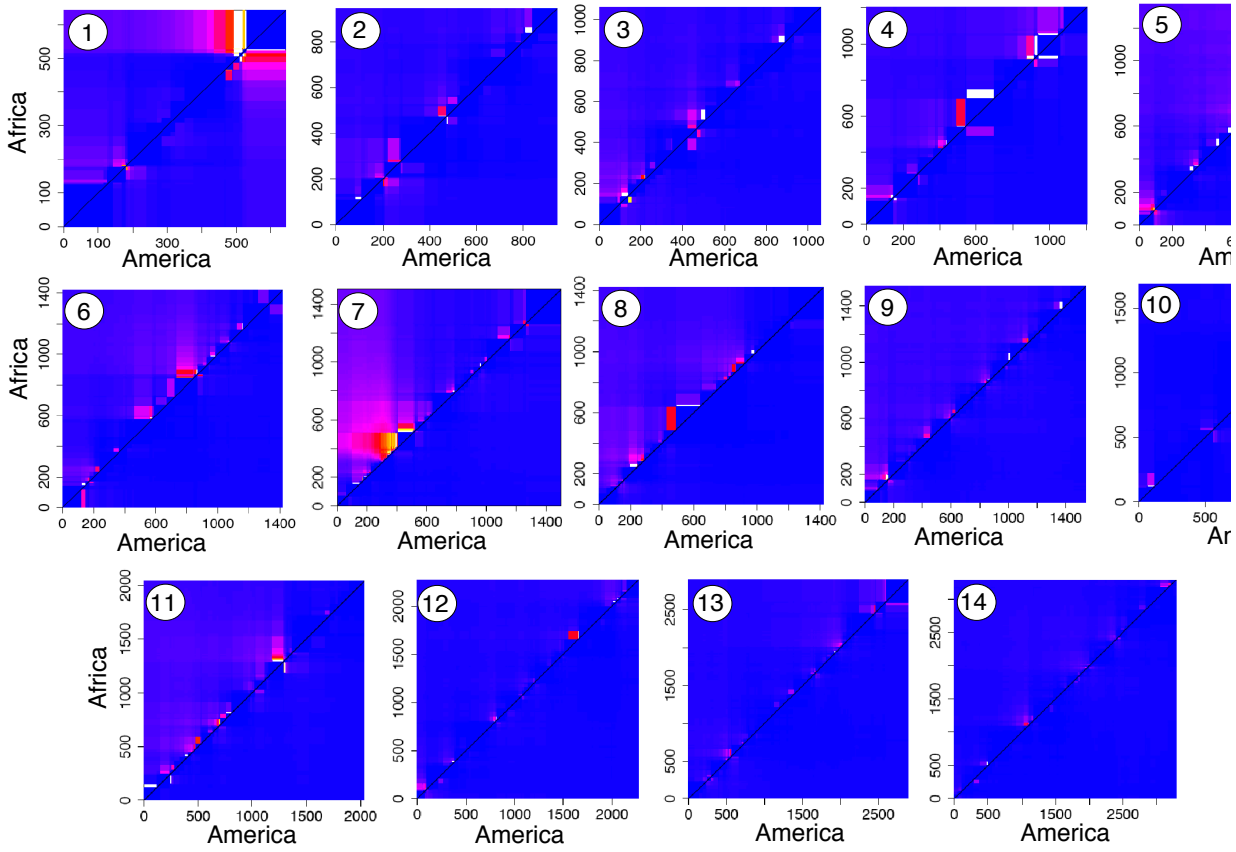
Figure 3-1. Population structure and principal component analysis (PCA) of *Plasmodium falciparum* parasite populations.
a) Population partitions using STRUCTURE (v2.2) [10]. The Cambodian group (red) consists of parasites CP195, CP201, CP216, CP285, CP286, CP291, CP313, CP268, CP325, CP305, CP307, CP256, CP238, and CP211 from Cambodia. b) PCA plot of all the parasites. Parasite continental origins are as color-coded and 'X' indicates outliers. PNG, Papua New Guinea. c) PCA plot of the Thai-Cambodian parasites showing outliers from the region.

Figure 3-2. Recombination events and hotspots on the 14 chromosomes of parasites.
Asian and Africa (a), Africa and America (b), and Cambodia (c). Recombination counts along the chromosomes were plotted and compared between parasites populations (a and b) or between chromosomes with the Cambodian population(c), with lighter color (white) reflecting more recombination counts and darker color for fewer recombination events. The counts depend on the numbers of isolates. Because there were more isolates from Asia, more recombination events were detected. Panels 1-14, as marked in (a) and (b), represent data from each of the 14 chromosomes. The numbers on both 'x' and 'y' are nucleotide positions on each chromosome.

(a)



(b)



(c)

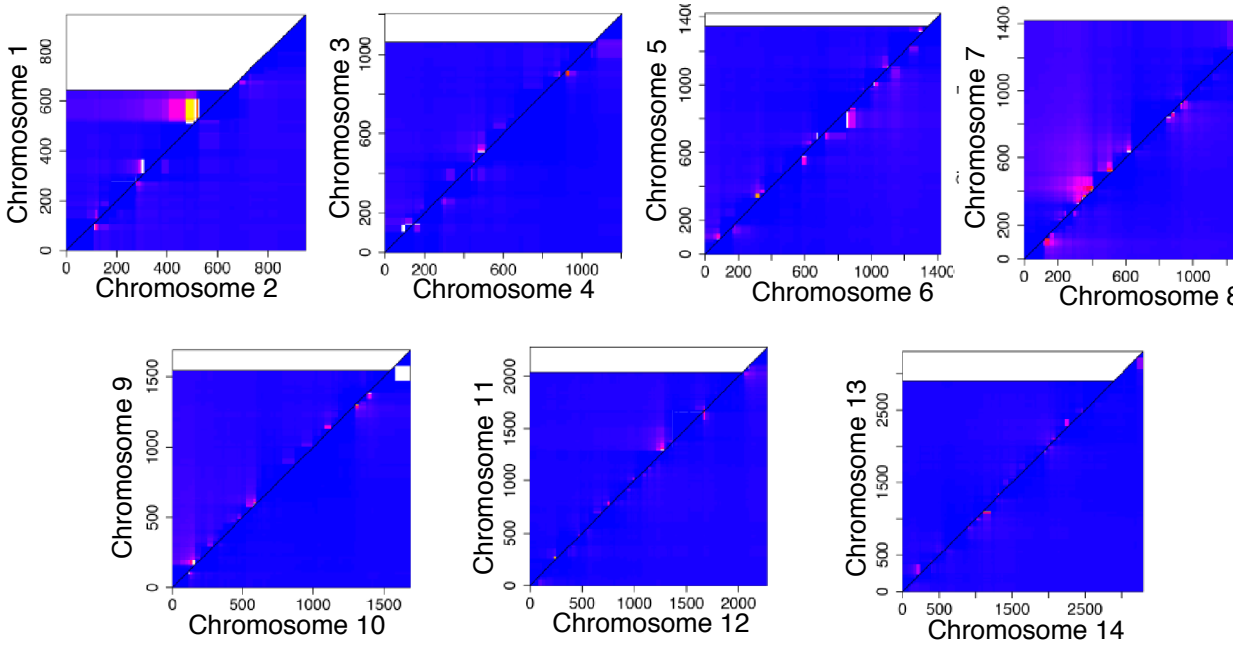
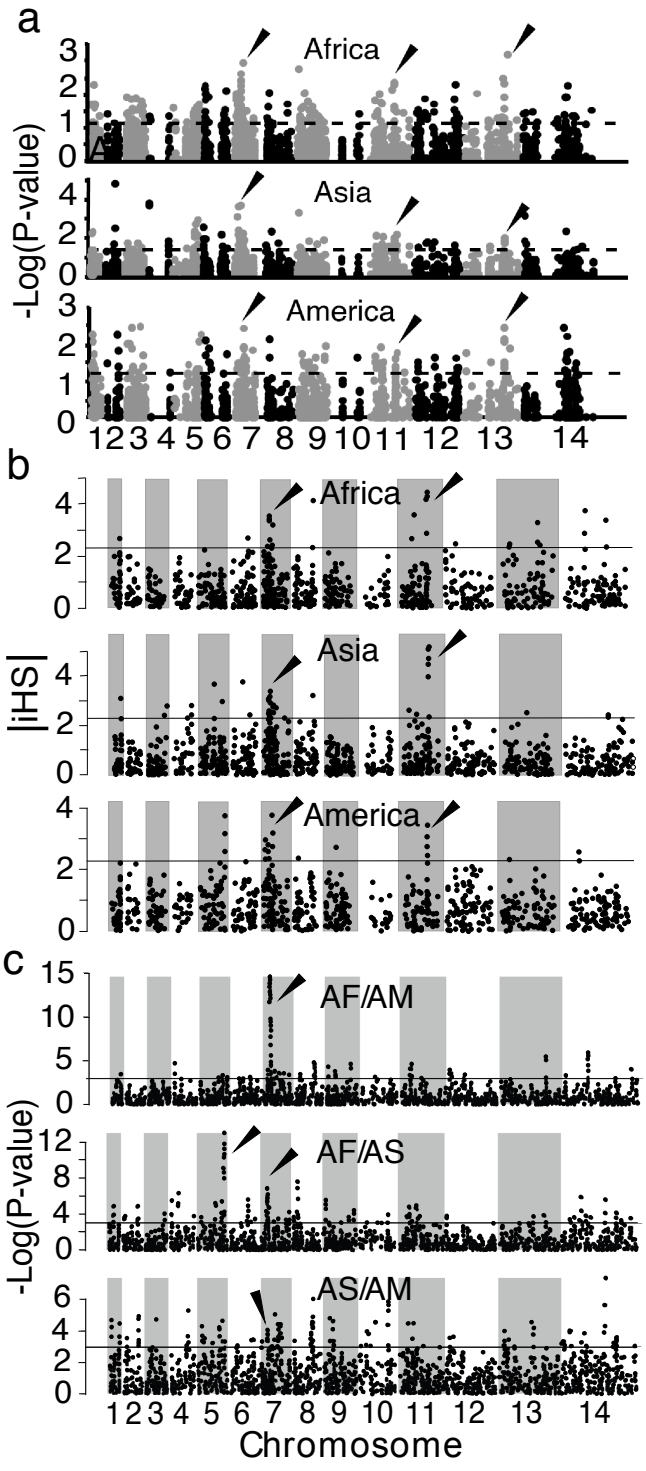

Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
Chromosome 13
Chromosome 14

Figure 3-3. Loci subject to positive selection in *Plasmodium falciparum* populations from Africa, Asia and America.

a) plots of –Log *P* values showing loci significantly under positive selection. Arrowheads point to loci containing the genes encoding chloroquine resistance transporter (*pfcrt*) on chromosome 7, the apical membrane antigen (*pfama-1*) on chromosome 11, and an ABC transporter on chromosome 13, respectively; Dots above the dash lines indicate significant. (b) plots of integrated haplotype scores (iHS) showing loci under selection. Arrowheads indicate the *pfcrt* and *pfama-1* loci on chromosome 7 and 11, respectively. SNPs with │iHS│ values≥2.3 were those above the horizontal line in each graph. Each dot represents an │iHS│ value from a window of 21 SNPs (a core SNP plus 10 SNPs on each side). (c) plots of –log*P* values from cross population extended haplotype homozygosity (XP-EHH) analyses. AF/AM, comparison of African and American populations; AF/AS, comparison of African and Asian populations; AS/AM, comparison of Asian and American populations. The horizontal lines indicate significant *P*-values (<0.05), and the arrowhead points to the *pfcrt* locus on chromosome 7 and PFE1445c locus on chromosome 5, respectively.


a
Africa
Asia
America
-Log(P-value)
1 2 3 4 5 6 7 8 9 10 11 12 13 14
b
Africa
Asia
America
|iHS|
c
AF/AM
AF/AS
AS/AM
-Log(P-value)
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Chromosome

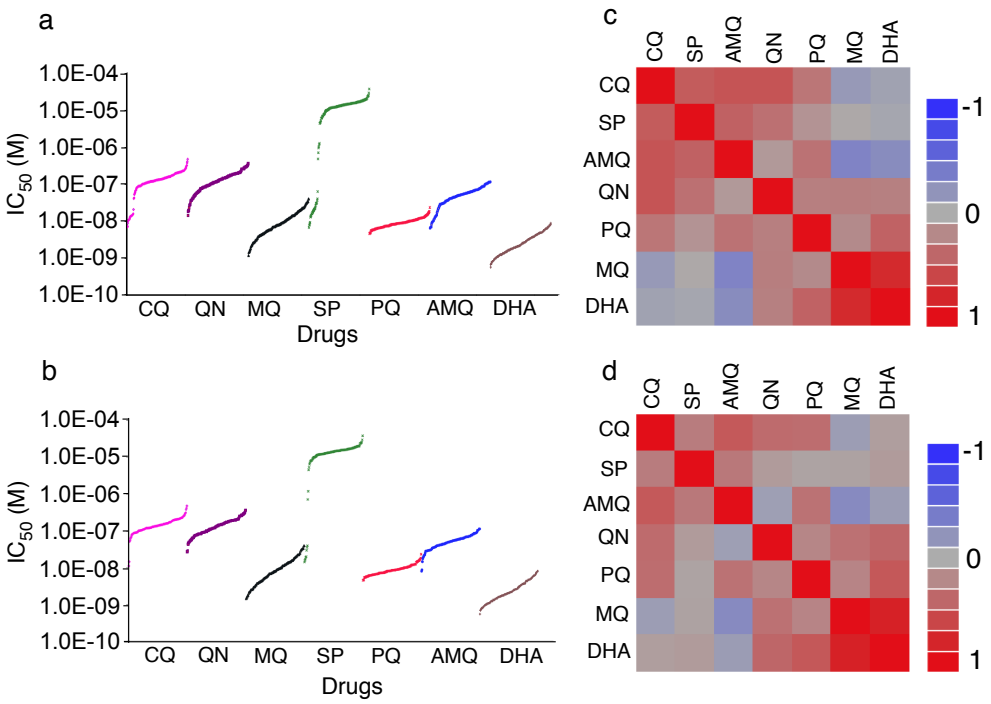


Figure 3-4. *In vitro* parasite responses (IC_{50}) to seven antimalarial drugs.
(a) IC_{50} values to seven different antimalarial drugs from 185 parasites were sorted from the lowest to the highest values. Note gaps in IC_{50} values in parasite responses to chloroquine (CQ) and sulfadoxine-pyrimethamine (SP), but continuous distributions for the other drugs. IC_{50} curves for each drug as marked in the figure; (b) similar plots as in (a) for parasites from Thai-Cambodian population after removing parasites with different genomic structure; (c), (d) multivariate analyses showing correlations between responses to seven different drugs for all the parasites (c) and Thai-Cambodian parasites (d). Dihydroartemisinin (DHA) and mefloquine (MQ) had strong positive correlation; CQ, amodiaquine (AMQ), piperaquine (PQ), quinine (QN), and SP also had positive correlation to some degree; whereas AMQ/DHA, AMQ/MQ had negative correlation.

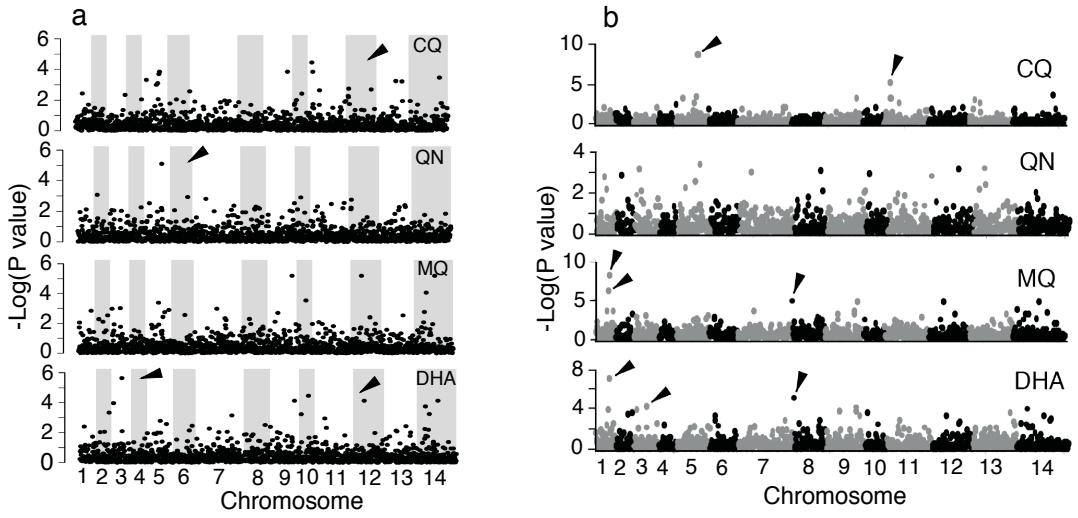


Figure 3-5. **Genome-wide scan for SNPs associated with responses to antimalarial drugs in the Asian population.**
Values of -Log *P* for four drugs were plotted against chromosomal positions. The arrowheads indicate SNPs with Bonferroni corrected $P<0.05$. (a) plots from EIGENSOFT; (b) plots from PLINK. CQ, chloroquine; QN, quinine; MQ, mefloquine; DHA, dihydroartemisinin.

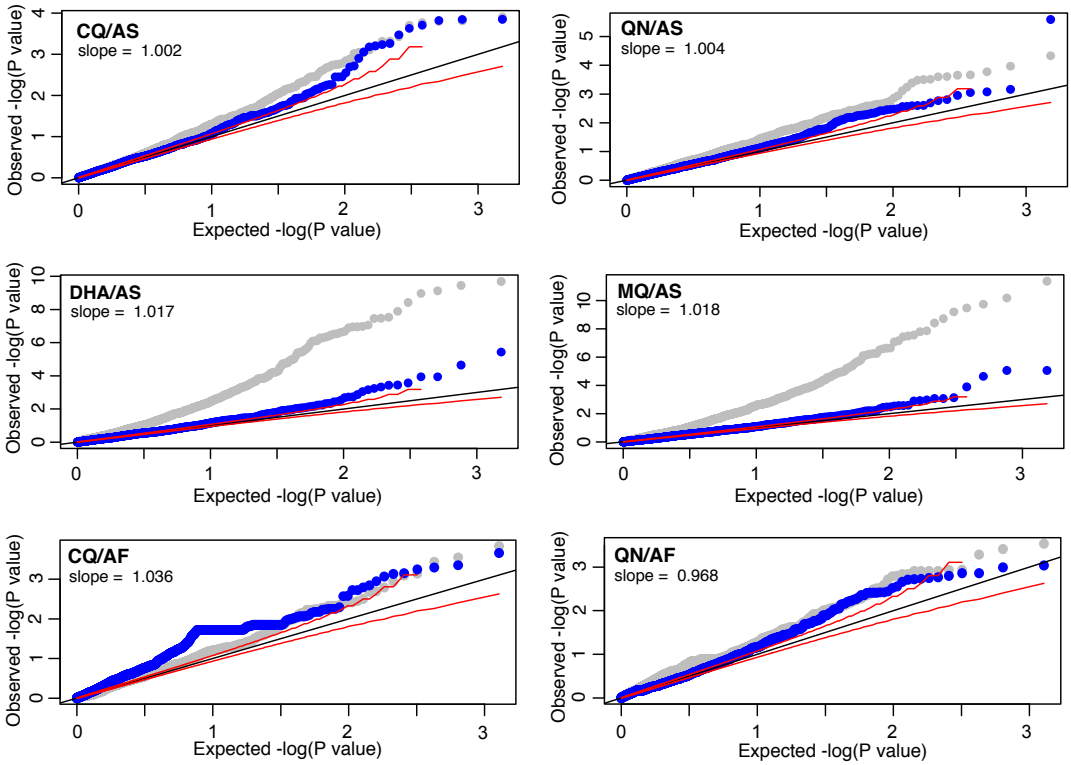


Figure 3-6. Quantile-Quantile plots of P-values before and after principal component analysis (PCA) correction for genome-wide scans.
Observed and PCA corrected -Log10P-values were plotted. Gray dots represent the observed P-values, and the blue dots represent the data obtained after PCA correction. The black line is the excepted line under the null distribution, and the red lines are 95% confident intervals. Data from four drugs, chloroquine (CQ/AS), quinine (QN/AS), mefloquine (MQ/AS), and dihydroartemisinin (DHA/AS) from the Asian (AS) population, and CQ and QN from the African (CQ/AF and QN/AF) population are presented. Large inflation of observed data across the entire distribution was seen in DHA/AS and MQ/AS, reflecting the existence of population structure in the data. After correction for population structure, the signals were much closer to the expected lines.

Table 3-1. Genes under significant selection detected by all the three haplotype tests. Genes, Gene ID from PlasmoDB (www.PlasmoDB.org); Chr, chromosome; Haplotype tests, indicate genes significantly under selection detected by the three methods. REHH, relative extended haplotype homozygosity; iHS, integrated haplotype score; XP-EHH, cross-population extended haplotype homozygosity; AF, AM, and AS in parenthesis indicate African, American, and Asian populations, respectively. Predicted functions, predicted gene functions from PlasmoDB; No. TM, number of predicted transmembrane domains; SNPs, total SNPs from all parasite isolates.

Gene	Chr	Haplotype tests	Predict function	No. TM	No. SNPs
PFC0940c	3	REHH (AF,AM,AS); iHS (AS); XP_EHH (AF/AS)	conserved Plasmodium protein, unknown function	2	23
PFE1445c	5	REHH (AM, AS); iHS (AM); XP-EHH (AF/AM, AF/AS)	conserved Plasmodium protein, unknown function	2	7
MAL7P1.27	7	REHH (AF,AS,AM); iHS (AF); XP-EHH (AF/AM,AF/AS)	chloroquine resistance transporter	10	25
PF07_0033	7	REHH (AF); iHS (AM); XP-EHH (AF/AS)	cg4 protein	0	2
PF07_0035	7	REHH (AF,AS); iHS (AS); XP-EHH (AF/AM,AF/AS)	cg1 protein	0	41
PF07_0036	7	REHH (AF); iHS (AM); XP-EHH (AF/AM,AF/AS)	cg6 protein	0	4
PF07_0037	7	REHH (AF,AS); iHS (AS); XP-EHH (AF/AM,AF/AS,AM/AS)	cg2 protein	0	72
PF07_0041	7	REHH (AF,AS); iHS (AS); XP-EHH (AF/AM,AF/AS)	conserved Plasmodium protein, unknown function	1	10
PF07_0042	7	REHH (AF,AS); iHS (AS,AM); XP-EHH (AF/AM,AF/AS)	conserved Plasmodium protein, unknown function	0	251
PF07_0068	7	REHH (AS); iHS (AS); XP-EHH (AF/AM,AF/AS)	cysteine desulfurase, putative	1	7
PF11_0188	11	REHH (AF,AM); iHS (AF,AS); XP-EHH (AM/AS)	heat shock protein 90, putative	0	10

Table 3-2. Genes/SNPs significantly associated with drug responses within Asian and African parasite populations*.
CQ, chloroquine, QN, quinine; MQ, mefloquine; DHA, dihydroartemisinin; SNP, single nucleotide polymorphism; NTs, the two nucleotides detected in the parasite populations; MAF (%), minor allele frequency; Chr, chromosome; Position, SNP position on chromosome; Unadj-P, unadjusted P-values; Bonf, Bonferonni adjusted P-value. *P*-values in bold are significant.* No gene/SNP was significant in the American population.

									EIGENSOFT		PLINK	
Drug	**Pop.**	**SNP**	**NTs**	**Annotation**	**MAF (%)**	**Chr**	**Position**	**Selection**	**Unadj-P**	**BONF-P**	**Unadj-P**	**BONF-P**
CQ	Asian	PFE1150w-5	G-T	PfMDR1	5.3	5	961620	iHS	0.00014	0.18	2.03×10^{-9}	**2.5×10^{-6}**
CQ	Asian	PFE1150w-3	A-T	PfMDR1	6.3	5	960984	iHS, XP-EHH	0.0002	0.25	2.03×10^{-9}	**2.5×10^{-6}**
CQ	Asian	PF11_0079-1	A-G	conserved Plasmodium protein	6	11	281940		0.00034	**0.048**	6.9×10^{-6}	**0.0085**
CQ	African	MAL7P1.27-9	A-G	PfCRT	20	7	310066	REHH, iHS, XP-EHH	0.0021	0.93	9.7×10^{-6}	**0.016**
QN	Asian	PFE1150w-4	A-G	PfMDR1	15.8	5	961008	iHS, XP-EHH	7.7×10^{-6}	**0.011**	0.00034	0.41
MQ	Asian	PFA0655w-27	G-T	SURFIN	6	1	518494	iHS, XP-EHH	0.0015	0.88	1.2×10^{-8}	**1.5×10^{-5}**
MQ	Asian	PFA0655w-18	C-G	SURFIN	18	1	517438	iHS, XP-EHH	0.039	1	1.1×10^{-6}	**0.0013**
MQ	Asian	MAL8P1.101-1	A-G	RNA binding protein	5.3	8	745328		0.98	1	1.7×10^{-5}	**0.021**
DHA	Asian	PFC0460w-4	A-G	conserved Plasmodium protein	7.5	3	466483		2.27×10^{-6}	**0.0032**	0.0033	1
DHA	Asian	PF10_0309-1	C-T	DEAD/DEAH box helicase	6.8	10	1272199		3.4×10^{-5}	**0.047**	9.6×10^{-8}	**0.00012**

Table 3-2. Continued

DHA	Asian	PFA0655w-27	G-T	SURFIN	6	1	518494	iHS, XP-EHH	0.018	1	7.5×10^{-6}	**0.009**
DHA	Asian	MAL8P1.101-1	A-G	RNA binding protein	5.3	8	745328		0.73	1	0.0037	1

Chapter 4 Direct Measure of the *de novo* Mutation Rate in Autism and Schizophrenia Cohorts

Philip Awadalla,[1,2,3,15]* Julie Gauthier,[3,15] Rachel A. Myers,[1,7,15] Ferran Casals,[1] Fadi F. Hamdan,[2,3] Alexander R. Griffing,[7] Mélanie Côté,[3] Edouard Henrion,[3] Dan Spiegelman,[3] Julien Tarabeux,[3] Amélie Piton,[3] Yan Yang,[3] Adam Boyko,[8] Carlos Bustamante,[8] Lan Xiong,[3] Judith L. Rapoport,[9] Anjené M. Addington,[9] J. Lynn E. DeLisi,[10] Marie-Odile Krebs,[11] Ridha Joober,[12] Bruno Millet,[11] Éric Fombonne,[13] Laurent Mottron,[4] Martine Zilversmit,[1] Jon Keebler,[1,7] Hussein Daoud,[3] Claude Marineau,[3] Marie-Hélène Roy-Gagnon,[2] Marie-Pierre Dubé,[5] Adam Eyre-Walker,[14] Pierre Drapeau,[6] Eric A. Stone,[7] Ronald G. Lafrenière,[3] and Guy A. Rouleau[1,2,3]**

[1]Department of Pediatrics; Université de Montréal; Montréal, Quebec H3T 1C5, Canada; [2]CHU Sainte-Justine Research Centre; Université de Montréal; Montréal, Quebec H3C 1G7, Canada; [3]Centre of Excellence in Neuromics of Université de Montréal, Centre Hospitalier de l'Université de Montréal, and Department of Medicine; Université de Montréal; Montréal, Quebec H2L 2W5, Canada; [4]Department of Psychiatry, Hôpital Rivière-des-Prairies; Université de Montréal; Montréal, Quebec H1E 1A4, Canada; [5]Centre de recherche Institut de Cardiologie de Montréal, Department of Pharmacology; Université de Montréal; Montréal, Quebec H1T 1C8, Canada; [6]Groupe de recherche sur le système nerveux central, Department of Pathology and Cell Biology; Université de Montréal; Montréal, Quebec H3C 3J7, Canada; [7]Bioinformatics Research Center; North Carolina State University; Raleigh, North Carolina 27606, USA;

[8]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA; [9]Child Psychiatry Branch; National Institute of Mental Health; Bethesda, Maryland 20892, USA; [10]Center for Advanced Brain Imaging; Nathan S Kline Institute; Orangeburg, New York 10962, NY, USA; [11]University Paris Descartes ; INSERM, Laboratoire de Physiopathologie des Maladies Psychiatriques, Centre de Psychiatrie et Neurosciences, U894 ; Sainte-Anne Hospital, Paris 75014, France ; [12]Department of Psychiatry; McGill University and Douglas Hospital; Montreal, Quebec H3A 1A1, Canada; [13]Department of Psychiatry; McGill University and Montreal Children's Hospital; Montreal, QC, H3Z 1P2, Canada; [14]Centre for the Study of Evolution, School of Life Sciences; University of Sussex; Brighton BN1 9QG, United Kingdom

[15]These authors contributed equally to this work

*Correspondence: philip.awadalla@umontreal.ca

**Correspondence: guy.rouleau@umontreal.ca

Published in The American Journal of Human Genetics, 87, 1-9 (10 September 2010).

Summary

The role of *de novo* mutations (DNMs) in common diseases remains largely unknown. Nonetheless, the rate of *de novo* deleterious mutations and the strength of selection against *de novo* mutations are critical to understanding the genetic architecture of a disease. Discovery of high impact DNMs requires substantial high-resolution interrogation of partial or complete genomes of families using re-sequencing. We hypothesized that deleterious DNMs may play a role in cases of autism spectrum disorders (ASD) and schizophrenia (SCZ), two etiologically heterogeneous disorders with significantly reduced reproductive fitness. We present a direct measure of the *de novo* mutation rate (μ) and selective constraints from DNMs, estimated from a deep resequencing data set generated from of a large cohort of ASD and SCZ cases (n=285) and a population of controls (n=285) with available parental DNA. A survey of ~ 430 Megabases (Mb) of DNA from 401 synapse-expressed genes across all cases and 25 Mb of DNA in controls found 28 candidate DNMs, 13 of which were cell line artifacts. Our calculated direct neutral mutation rate (1.36×10^{-8}) is similar to previous indirect estimates but we observed a significant excess of potentially deleterious DNMs in ASD and SCZ individuals. Our results emphasize the importance of DNMs as genetic mechanisms in ASD and SCZ, and the limitations of using DNA from archived cell lines to identify functional variants.

Introduction

The rate at which human genomes mutate is critical to understanding every aspect of medical, statistical, and evolutionary genomics. To date, human mutation rate estimates have been indirectly inferred either from a human-chimpanzee divergence approach [1], from the analysis of mutations causing human Mendelian diseases [2; 3], or more recently using next generation sequencing in one nuclear family [4]. These data suggest that there will be ~2 *de novo* mutations (DNMs) in the genome-wide coding regions per zygote, so that such mutations may contribute to some common diseases. Indeed, DNMs in individuals with complex disorders could explain genetic factors that are not detectable through genome-wide association studies. Deep resequencing of patients suffering from various diseases, and their parents, holds the promise of discovering DNMs that potentially could have a significant impact on disease prevalence and severity. Disease-causing mutations are more likely to involve selectively constrained positions where mutations are likely to be less tolerated or may have a substantial impact on fitness. If DNMs contribute significantly to a disorder, then there should be more functional and potentially deleterious mutations: 1) in cases versus control samples, and 2) in functionally constrained sites versus non-functional, and thus unconstrained (neutral), sites within the same cohort.

The disruption of gene function by rare deleterious penetrant mutations could represent an important cause of neurodevelopmental disorders such as schizophrenia (SCZ, MIM181500) and autism spectrum disorders (ASD, MIM209850). In fact,

deleterious DNMs may explain observations such as the high global incidences of ASD (~0.45%)[5] and SCZ (~0.4%) [6] despite extremely variable environmental factors and reduced reproductive fitness [7], and increased risk with increasing parental age[8; 9]. Indeed, recent studies report an excess of *de novo* copy number variants (CNVs) in ASD and SCZ compared to controls [10-12]. We hypothesized that sequencing of families with affected individuals will identify an excess of missense relative to silent *de novo* mutations and that these mutations are candidate causal mutations for ASD and SCZ. As part of the Synapse to Disease Project (S2D), we resequenced synaptic genes in ASD and SCZ cases, and a group of population controls. Such a resequencing project will capture DNMs at greater resolution, with the potential to unambiguously identify missense or frameshift mutations not detectable using linkage, association or CNV methods. To test our hypothesis we examined variants identified by resequencing 401 genes in a cohort of 285 ASD and SCZ individuals, and for a subset of 39 of these genes in 285 population control individuals. Our analyses demonstrate a neutral mutation rate similar to that already reported and an excess of *de novo* deleterious mutations associated with the disease cohorts.

Subjects and Methods

Diagnostic screening and selection of patients

The cohort of patients used for the sequencing of candidate genes for discovery of DNMs included 142 unrelated ASD patients (122 males and 20 females) previously

described [13], 65% of which had no family history of ASD or related neurological disorders. All patients were diagnosed using the Diagnostic and Statistical Manual of Mental Disorders criteria. Depending on the recruitment site, the Autism Diagnostic Interview-Revised or the Autism Diagnostic Observation Schedule was used. In addition, the Autism Screening Questionnaire (ASQ) was completed for all the subjects. We excluded patients with an estimated mental age <18 months, a diagnosis of Rett syndrome or Childhood Disintegrative Disorder and patients with evidence of any other psychiatric and neurological conditions including: birth anoxia, rubella during pregnancy, fragile-X syndrome, encephalitis, phenylketonuria, tuberous sclerosis, Tourette and West syndromes. The 143 SCZ subjects (95 males and 48 females) were collected from 5 different centers and included 28 cases of childhood onset schizophrenia (COS) and 115 sporadic or familial cases (with unaffected parents) of adult onset schizophrenia or schizoaffective disorder [14-17]. Sixty percent of the SCZ subjects had no family history of schizophrenia or other related neurological disorders. They were evaluated by experienced investigators using the Diagnostic Interview for Genetic Studies (DIGS 3.0) [18] or Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS), and multidimensional neurological, psychological, psychiatric, and pharmacological assessments. Family history for psychiatric disorders was also collected using the Family Interview for Genetic Studies (FIGS). All DIGS and FIGS were reviewed by two or more psychiatrists for a final consensus diagnosis based on DSM-IIIR or DSM-IV at each centre. Exclusion criteria included patients with psychotic symptoms mainly caused by

alcohol, drug abuse, or other clinical diagnosis including major cytogenetic abnormalities. We selected patients for which blood DNA was available so that DNMs could be validated and for which DNA was available from both parents to test for inheritance of the variants. The population control cohort (150 males and 135 females) consisted of unrelated individuals collected for the Quebec Newborn Twin Study (QNTS) [19] where DNA samples were available from both parents and both twins (either monozygotic or dizygotic); however only one sibling was chosen randomly for sequencing. All samples were collected through informed consent following approval of each of the studies by the respective institutional ethics review committees. Ethnic origins of the grand-parents were self-reported by the parents of probands and population control subjects. The ASD cohort is composed of French Canadians (85), other European descent (54) and non- European descent (3). The SCZ cohort is composed of European descent (136) and Asians (7). The control population is composed of French-Canadians (204), other European descent (55), non-European descent (18) and individuals of mixed origin (8).

Selection of candidate genes

The list of genes screened in the S2D project was generated from a repertoire of approximately 5000 potential synaptic genes compiled from several synaptic lists from different synapse databases, including the Genes-to-Cognition (G2C) database, which include an extensive list of postsynaptic genes [20], the Synapse database (SynDB) which

uses Synapse Ontology algorithms to mine potential synaptic genes [21], and from an extensive list of synaptic vesicle genes [22]. It also comprises genes identified through manual searches of PubMed that are either localized at the synapse or affect synapse-related functions (i.e. plasticity, axon or dendrite outgrowth, dendritic spine morphology, learning and memory).

The sequencing data reported here were generated from the screening of the coding regions and splice site junctions of 122 X-linked and 279 autosomal potentially synapse related genes in 142 ASD and 143 SCZ subjects using Sanger technology. The excess of X-linked genes is due to the S2D selection procedure which sought to include all potentially synaptic X-linked genes because of their importance in neurodevelopmental diseases[23] and because ASD is more common in males than females [24]. The genes on the autosomes were largely that encode glutamate receptors (including NMDA receptors, AMPA receptors, kainate receptors, metabotropic glutamate receptors), as well as genes that encode proteins that interact with them, in particular those complexed with the NMDAR, a majority of which were identified by large proteomic studies [25] and reported in the G2C database [20]. The autosomal gene list is composed of 203 genes from the glutamate receptor complex (23 known glutamate receptors and 180 of their synaptic interactors), and 73 genes implicated in synapse function and/or cognition, interaction with ASD genes, or due to their disruption in the context of small copy number variations (CNVs) or balanced translocations in patients with ASD, SCZ, or mental retardation (MR). An additional 3 genes were included for

which mutations reported to cause ASD, SCZ, or the related neurodevelopmental disease, non-syndromic mental retardation (NSMR) which is known to co-exist with ASD. Mutations which multiple reports have previously found associations with diseases that are not related to ASD or SCZ were eliminated. In addition, data generated from the resequencing of 39 of these 401 genes in 285 population control samples were used. The 39 genes were resequenced in controls after discovering a *de novo* or deleterious mutation in ASD or SCZ samples.

DNA preparation, sequencing and variant identification

Genomic DNA was extracted from peripheral blood lymphocytes and/or lymphoblastoid cell lines using Puregene extraction kits (Gentra System, USA). A panel of 7 microsatellite markers was used to confirm parentage for all samples [26]. To overcome the issue of limited DNA material, the gene screening was performed on DNA isolated from an Epstein-Barr Virus transformed lymphoblastoid cell line for most cases (n= 224). The rest being done on blood DNA (n=61). The ASD cell lines samples had been frozen or regrown a maximum of two times. For the SCZ cell lines, 59 DNA samples were acquired from Coriell and are available through a request to J. L. DeLisi. All unique variants (heterozygous in a single individual) detected during the screen were tested in the parents DNA. All potential *de novo* variants (not seen in the parents) were reconfirmed by reamplifying the fragment and resequencing the proband blood DNA and both his/her parents using reverse and forward primers in order to eliminate PCR or

sequencing artifact. All *de novo* identified originally from cell lines DNA were retested in the subject DNA extracted from blood to rule out variations that could have occurred during production or growth of the lymphoblastoid cell line. An identity DNA test was performed to eliminate any cell lines/blood inconsistencies caused by sample identification errors or non-paternity. Primers were designed using the Exon Primer program from the UCSC Genome Browser. PCR products were sequenced on one strand using Sanger technology done at the Genome Quebec Innovation Centre in Montreal, Canada on a 3730XL DNA Analyzer System. PolyPhred (v.5.04), PolyScan (v.3.0), and Mutation Surveyor (v.3.10, SoftGenetics Inc.) were used for variant detection.

Estimation of base-pairs screened

To estimate the number of base-pairs (bp) screened, we determined the amount of coding and non-coding (intronic, UTR) sequence screened for each gene based on our PCR amplicon designs. Briefly, genomic intervals were calculated based on Forward and Reverse PCR primer sequences for each amplicon. Because the DNA sequence overlapping each PCR primer is not surveyed, these genomic intervals did not include those corresponding to each PCR primer. Furthermore, because the first ~30 bp of sequence from each sequence read are of low quality, we trimmed 30 bp from each genomic interval (usually at the Forward primer end). Overlapping amplicons were merged to form a single genomic interval. Then each genomic interval was annotated as

to coding and non-coding sequence for each gene using tools available on the refGene table from the UCSC Human Genome Browser.

We defined functional and non-functional sites sequenced as those in which a nucleotide change would, or would not, lead to an altered protein sequence, respectively. We estimated that 71.2% of coding sites were functional, whereas non-functional sites were estimated at 28.8% of the coding and 100% of the intronic sites [27]. The number of CpG sites was estimated as 2.8% for functional sites and 1% of non-functional sites [28]. The corrections for increased mutation rates in CpG regions was 10*number of CpG sites, yielding an effective bases sequenced of 10*CpG sites + Non-CpG sites for both functional and non-functional sites.

Evaluation of false negative mutation calls

A subset of 71 of our ASD samples was also genotyped using Affymetrix 500K arrays. We tested our ability to detect heterozygous variants by comparing our heterozygous calls from the resequencing screen of these 71 samples with calls made for overlapping genotyped SNPs on the array. Of the 1649 heterozygous calls made in 90 autosomal SNPs on the Affymetrix 500K chip overlapping our screened amplicons, we failed to detect 41 of those heterozygous calls in our resequencing survey, suggesting that our false negative rate is ~2%.

Prediction of missense severity

The potential consequence of each missense variant was evaluated using the MAPP [29], PolyPhen [30], SIFT [31], and PANTHER [32] programs. Orthologous protein sequence alignments were obtained using tools available on the Galaxy Browser website for the generation of MAPP scores.

Statistical Analysis

Excess of unctional relative to non-functional DNMs in each category (Initial, CpG, non-CpG, and Effective Bases sequenced) was measured using 1) Binomial test (P[X ≥ # functional DNMs | *q,* functional bases sequenced], where *q* = neutral mutation rate) and 2) Fisher's Exact test (FET). Functional DNMs are defined as missense and nonsense DNMs, while non-functional DNMs include silent and intronic DNMs.

Results

Identification of de novo mutations

By resequencing the coding and splice junction regions of 401 genes in 142 ASD and 143 SCZ samples (and 19 of these genes in 285 QNTS control samples), we identified 6,184 DNA variants. Of these, 2,437 were unique (i.e. heterozygous in 1/285 unrelated individuals tested). Each of these 2,437 unique variants was resequenced in the proband and both parental samples to determine inheritance mode (transmitted versus *de novo*). A total of 15 unique variants were confirmed in the proband blood DNA sample

but not detected in either parent blood DNA sample (Table 4-1). These 15 validated DNMs are either germ-line derived mutations or arose as somatic mutations in blood tissue. A further 13 variants were not detected in parents DNA, nor were they detected in the proband blood DNA sample, and were assumed to be generated during lymphoblastoid cell line development (Table 4-2).

Of the 15 confirmed DNMs, 14 were detected in the ASD and SCZ cohorts (2 nonsenses, 5 missenses, 3 frameshifting indels, 2 silents and 2 intronic) and one was a missense DNMs found in the population control group (Table 4-1). Five of the 11 point mutations in cases were transitions, 4 being CpG mutations, and 6 were transversions. Eight of the 14 DNM detected among ASD and SCZ samples were disruptive to translation or protein structure and/or function, including three frameshifts and two nonsenses, and 3 missenses predicted to significantly disrupt protein structure using the computational prediction method MAPP[33]. Significant MAPP scores reflect a potentially damaging amino-acid change and predict deleterious consequences. Reassuringly, MAPP scores relate to allele frequencies as would be predicted by population genetics, with increasingly deleterious alleles tending toward lower frequencies (Figure 4-1). Three MAPP *p*-values of DNMs in SCZ and ASD samples are significantly low (Table 4-1); two were found in kinesin encoding genes (R349P in *KLC2* [MIM 611729]; R802C in *KIF5C* [MIM 604593]). One nonsense mutation (Y575X) was also found in a kinesin-encoding gene (*KIF17* [MIM 605037]). One nonsense and one splice site deletion were found within *SHANK3* [MIM 606230] and a frameshift mutation was found in each of

IL1RAPL1 [MIM 300206] and *NRXN1* [MIM 600565]. We have also confirmed the damaging predicted functional impact of the 5 *de novo* missenses using three other prediction programs (Panther, SIFT and PolyPhen) (data not shown). Only one DNM was discovered in an X-linked gene (*IL1RAPL1*) [12]. More detailed description of these genes and their potential role in ASD and SCZ will be presented elsewhere (manuscript in preparation).

Estimates of the Neutral Human Mutation Rate

To calculate human mutation rates, we estimated the total initial count of bases we resequenced in all cohorts as 458.8 Mb (Table 4-3 and Methods). This includes 230.6 Mb of protein-coding and 228.3 Mb of intronic sequence. In the ASD and SCZ cohorts, exonic material sequenced is 215,186,702 bases and intronic is 218,145,769 bases. The total number of replacement sites in the cases is 153,704,787 and the number of silent (synonymous and intronic) sites is 279,627,684. In the QNTS cohort, 15,422,960 bases were exonic and 10,122,419 bases were intronic. The number of replacement sites is 11,016,400 and the number of silent sites is 14,528,979.

We distinguished functional from non-functional sites based on the effect of a mutation on transcription or translation of the protein at a given position (see Methods). When addressing whether there was an excess of functional DNMs [34] relative to non-functional or silent base pairs, we calculated the "effective bp count" [35]. Because mutations are 10 times more likely to occur at CpG sites than at non-CpG sites, we

calculated the total number of CpG and non-CpG sites for both functional and non-functional sites [28], and calculated the effective bp count (non-CpG sites + 10 x CpG sites) to account for increased mutation rates in CpG sites (Table 4-3).

We calculated the neutral mutation rate by examining the number of DNMs found in non-functional sites in our ASD, SCZ and QNTS samples. In total we sequenced ~294 Mb of non-functional DNA in which we observed 4 DNMs (Table 4-1, in genes *GSN* [MIM 137350], *MAP2K1* [MIM 176872], *BSN* [MIM 604020] and *ATP2B4* [MIM 604020]). Since these mutations are unlikely to be pathogenic (referred to here as "neutral") they allowed us to directly estimate the rate of neutral point mutations. We estimated this to be 1.36×10^{-8} mutations per site per generation (95% Poisson Confidence Interval: 0.34×10^{-8}, 2.7×10^{-8}). These estimates of neutral mutation rates are similar to, and not significantly different from the estimate of 2.5×10^{-8} derived from phylogenetic analyses [1; 2] and the intergeneration estimate of 1.1×10^{-8} derived from next generation sequencing data [36].

Excess of functional DNMs in ASD and SCZ cohorts

If DNMs cause sporadic cases of ASD and SCZ, then DNMs will be more common in functional than in the non-functional sites in our disease cohorts. Based on an observed 4 DNMs in 294 Mb of neutral (silent and intronic) DNA, we expect 1.3 DNMs in the 96 Mb of non-synonymous DNA sites resequenced in the cases with no family history of disease (65% of ASD and 60% of SCZ cases). However, among trios without

family histories (Table 4-4), we observed 6 non-synonymous DNMs surveyed in individuals, representing a significant enrichment of non-synonymous DNMs (p=0.003 in one-tail binomial test; p=0.022 FET, Table 4-3). This excess remains significant even when we take into account that CpG dinucleotides mutate faster than other sites and are more common in exons than introns (p=0.008 in one-tail binomial test, p=0.032 FET; see Table 4-3 and Methods).

If DNMs cause disease, we also expect point mutations with larger effects to be more frequent than expected in the disease group. Among our 5 non-synonymous DNMs in trios with no-family history, 2 are nonsense mutations (ratio 1:2.5) similar to previous estimates [3] for DNMs causing Mendelian diseases (1:3.9) that are catalogued in the Human Gene Mutation Database (HGMD). Also the ratio of synonymous to missense DNMs in the ASD and SCZ cohort is similar to that observed for HGMD [3]. Under a neutral model, we would expect a ratio of 1 nonsense to 19.7 missenses [3] when only point mutations are considered. In HGMD the number of missense to nonsense DNMs was significantly higher than the neutral expectation. Using a binomial test, our observed number of missense to nonsense DNMs was also significantly higher than the neutral expectation (p=0.04), suggesting that some of the mutations are predisposed to be pathogenic. All of these observations suggest an excess of potentially disease predisposing DNMs in the SCZ and ASD cohort. Taken together, these lines of evidence suggest that mutations with functional effects are over-represented within the synapse genes sequenced in individuals showing sporadic ASD and SCZ.

Comparing DNMs and segregating variant ratios

Functional and non-functional segregating variants provide an expectation of the proportion of functional and non-functional DNMs. We compared the ratio of functional and non-functional DNMs to the ratios of the same classes of segregating variants in the ASD and SCZ cohorts (Table 4-5). Similar observations were found for the QNTS cohort. The comparison was significant when the functional and non-functional DNMs were compared to all segregating sites (FET, $p < 0.001$) and to Unique SNP classes in the ASD and SCZ cohort (FET, $p = 0.003$). Given that the ratio of functional to non-functional was two times higher for DNMs relative to segregating sites, this suggests an excess of deleterious DNMs. Furthermore, given that rare SNP classes are likely enriched for slightly-deleterious missense mutations [3], this significant comparison can be considered conservative. Under the expectation that most highly deleterious mutations will be selectively removed in one generation, the unique comparisons above provide insight into the proportion of deleterious mutations in humans that are selectively removed relative to segregating variation. Potentially disease-causing DNMs were more frequent than non-functional DNMs in our cohorts relative to expectations inferred from segregating mutations.

Discussion

In the present study we have attempted to directly estimate the mutation rate using a large set of resequencing data generated from a common disease based project. In addition we have tried to validate that DNMs are a possible genetic factor of ASD and SCZ. There are 3 main conclusions that can be drawn from our study. First, the source of biological material (blood DNA vs. cell line DNA) is crucial while doing experimental analyses using resequencing data seeking DNMs. All DNMs analyzed here were confirmed by resequencing, using standard Sanger technology, from DNA samples extracted from blood in the proband and in the parents. In so doing, we discovered that ~50% of our identified DNMs are the result of mutations that most likely occurred during the transformation and propagation of lymphoblastoid cell lines thus creating false-positive DNMs. This observation was also recently stressed in CNV analyses by The Wellcome Trust Case Control Consortium [37]. This biological artifact, if unnoticed, would have significantly biased our results and would have contributed to a doubling of mutation rates for all classes of sites. Interestingly, 8 of 13 cell line mutations were X-linked, suggesting that this chromosome is particularly susceptible to generation or accumulation of deleterious mutations after transformation of lymphoblasts with Epstein-Barr Virus. These mutations which are hemizygous in males, may also be positively selected because they contribute to higher fitness in cells carrying these mutations. An awareness of the high rate of mutation observed in cell lines, some of which are archived at the Coriell Institute, is critical to any large-scale whole genome sequencing project,

and potentially to those taking advantage of next-generation sequencing technologies, to capture rare and/or pathogenic mutations. Not only will the inherent error rate of the technologies be critical, but so too will the choice of samples and the way those samples are being maintained or cultured. Second, using a direct calculation and classical Sanger sequencing, we validated the reported estimates of the neutral mutation rate in humans. Third, our study confirms the critical role that large-sample, high-resolution nucleotide surveys play in detecting potentially disease-causing DNMs.

We acknowledge that there are weaknesses in our present study. The amount of resequencing in the controls is substantially lower than the resequencing in the cases. In fact, direct comparison in terms of sequences covered between cases and controls would be the optimal way to directly estimate the rate of mutation and detect significant differences in mutation rate between cases and controls. Nevertheless, our analyses show that the rate of potentially deleterious DNMs is significantly higher in functional compared to non-functional sites within the disease cohorts suggesting a role of functional DNMs in the etiology of ASD and SCZ. Given that our estimates of the neutral human mutation rate is consistent with a recent genome-wide estimate[36] and the accumulation of more direct observations of mutation, the confidence intervals of mutation rate estimates will begin to narrow. The rate of functional mutations in this survey is almost five times that of neutral or genome-wide rates, supporting our conclusions that we have likely detected mutations that are causal with respect to ASD and SCZ. By resequencing the genes where DNMs were discovered among QNTS

(random) participants, we were able to validate that these genes among a substantial number of random individuals do not carry functional DNMs with the exception of one locus (*SHANK3*, Table 4-1). At *SHANK3*, a nonsynonymous mutation of low predicted functional impact was discovered in the QNTS cohort. *SHANK3* has been previously implicated in ASD [38] and may be a rapidly evolving gene in humans, with substantial neurological phenotypic impact.

By demonstrating that functional DNMs are at higher relative frequencies than segregating polymorphisms (Table 4-5), we show that DNMs may have a substantial role in ASD and SCZ etiology. The power of the genomics approaches employed here are that DNMs are not subject to the same demographic processes that shape segregating site variation. As a result, it is not necessary to test or correct for population structure, nor are our analyses subject to population stratification or admixture issues associated with GWAS analyses. By using a simple genomics approach that compares different classes of sites, we have sufficient power to map candidate mutations that are more likely to contribute to these diseases.

From sequencing only 8% of genes expressed in the synapse, functional DNMs were found in 5% of individuals, having no family history, exhibiting a wide range of clinical phenotypes (see Methods and Table 4-4). Although we biased our sampling strategy towards likely candidate genes, our predictions appear to have been poor regarding the X-chromosome. It is therefore possible that by sequencing all 5,000 synapse-related genes we may uncover many of the mutations responsible for sporadic

cases of ASD and SCZ. Furthermore, since non-functional DNMs are predicted to be relatively rare in ASD and SCZ genes (1.36×10^{-8} non-synonymous DNMs per site) it may be easy to determine the likely causative mutations.

Acknowledgments

We would like to thank all the families and individuals who participated in this study. We are thankful for the efforts of the members of the Genome Quebec Innovation Centre Sequencing and Bioinformatic groups. This work was supported by Genome Canada and Génome Québec, and received co-funding from Université de Montréal for the Synapse to Disease (S2D) Project as well as funding from the Canadian Foundation for Innovation to both G.A.R and P.A and co-funding from the MDEIE of Quebec. G.A.R. holds the Canada Research Chair in Genetics of the Nervous System; P.A. holds a career award from the FRSQ and Genome Quebec.

Web Resources

The URLs for data presented herein are as follows:

Coriell Institute, http://www.coriell.org

Galaxy Browser, http://main.g2.bx.psu.edu/

UCSC Human Genome Broswer, http://genome.ucsc.edu/cgi-bin/hgGateway

Human Gene Mutation Database, http://www.hgmd.cf.ac.uk/ac/index.php

McGill University and Genome Quebec Innovation Centre,
http://www.genomequebecplatforms.com/mcgill/

NCBI, http://www.ncbi.nlm.nih.gov/omim

Panther, http://www.pantherdb.org/tools/csnpScoreForm.jsp

PolyPhen, http://genetics.bwh.harvard.edu/pph/

SIFT, http://blocks.fhcrc.org/sift/SIFT.html

Synapse-to-Disease Project, http://www.synapse2disease.ca

References

1. Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297-304.

2. Kondrashov, A.S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 21, 12-27.

3. Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80, 727-739.

4. Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328, 636-639.

5. Rutter, M. (2005). Incidence of autism spectrum disorders: changes over time and their meaning. Acta Paediatr 94, 2-15.

6. Saha, S., Chant, D., Welham, J., and McGrath, J. (2005). A systematic review of the prevalence of schizophrenia. PLoS Med 2, e141.

7. Bassett, A.S., Bury, A., Hodgkinson, K.A., and Honer, W.G. (1996). Reproductive fitness in familial schizophrenia. Schizophr Res 21, 151-160.

8. Croen, L.A., Najjar, D.V., Fireman, B., and Grether, J.K. (2007). Maternal and paternal age and risk of autism spectrum disorders. Arch Pediatr Adolesc Med 161, 334-340.

9. Malaspina, D., Brown, A., Goetz, D., Alia-Klein, N., Harkavy-Friedman, J., Harlap, S., and Fennig, S. (2002). Schizophrenia risk and paternal age: a potential role for de novo mutations in schizophrenia vulnerability genes. CNS Spectr 7, 26-29.

10. Xu, B., Roos, J.L., Levy, S., van Rensburg, E.J., Gogos, J.A., and Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 40, 880-885.

11. Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E., et al. (2008). Large recurrent microdeletions associated with schizophrenia. Nature 455, 232-236.

12. Piton, A., Michaud, J.L., Peng, H., Aradhya, S., Gauthier, J., Mottron, L., Champagne, N., Lafreniere, R.G., Hamdan, F.F., Joober, R., et al. (2008). Mutations in the calcium-related gene IL1RAPL1 are associated with autism. Hum Mol Genet 17, 3965-3974.

13. Gauthier, J., Bonnel, A., St-Onge, J., Karemera, L., Laurent, S., Mottron, L., Fombonne, E., Joober, R., and Rouleau, G.A. (2005). NLGN3/NLGN4 gene mutations are not responsible for autism in the Quebec population. Am J Med Genet B Neuropsychiatr Genet 132B, 74-75.

14. DeLisi, L.E., Shaw, S.H., Crow, T.J., Shields, G., Smith, A.B., Larach, V.W., Wellman, N., Loftus, J., Nanthakumar, B., Razi, K., et al. (2002). A genome-wide scan for linkage to chromosomal regions in 382 sibling pairs with schizophrenia or schizoaffective disorder. Am J Psychiatry 159, 803-812.

15. Gochman, P.A., Greenstein, D., Sporn, A., Gogtay, N., Nicolson, R., Keller, A., Lenane, M., Brookner, F., and Rapoport, J.L. (2004). Childhood onset schizophrenia: familial neurocognitive measures. Schizophr Res 71, 43-47.

16. Joober, R., Rouleau, G.A., Lal, S., Dixon, M., O'Driscoll, G., Palmour, R., Annable, L., Bloom, D., Lalonde, P., Labelle, A., et al. (2002). Neuropsychological impairments in neuroleptic-responder vs. -nonresponder schizophrenic patients and healthy volunteers. Schizophr Res 53, 229-238.

17. Mechri, A., Bourdel, M.C., Slama, H., Gourion, D., Gaha, L., and Krebs, M.O. (2009). Neurological soft signs in patients with schizophrenia and their unaffected siblings: frequency and correlates in two ethnic and socioeconomic distinct populations. Eur Arch Psychiatry Clin Neurosci 259, 218-226.

18. Gourion, D., Goldberger, C., Bourdel, M.C., Bayle, F.J., Millet, B., Olie, J.P., and Krebs, M.O. (2003). Neurological soft-signs and minor physical anomalies in schizophrenia: differential transmission within families. Schizophr Res 63, 181-187.

19. Lemelin, J.P., Boivin, M., Forget-Dubois, N., Dionne, G., Seguin, J.R., Brendgen, M., Vitaro, F., Tremblay, R.E., and Perusse, D. (2007). The genetic-environmental etiology of cognitive school readiness and later academic achievement in early childhood. Child Dev 78, 1855-1869.

20. Croning, M.D., Marshall, M.C., McLaren, P., Armstrong, J.D., and Grant, S.G. (2008). G2Cdb: the Genes to Cognition database. Nucleic Acids Res.

21. Zhang, W., Zhang, Y., Zheng, H., Zhang, C., Xiong, W., Olyarchuk, J.G., Walker, M., Xu, W., Zhao, M., Zhao, S., et al. (2007). SynDB: a Synapse protein DataBase based on synapse ontology. Nucleic Acids Res 35, D737-741.

22. Trinidad, J.C., Specht, C.G., Thalhammer, A., Schoepfer, R., and Burlingame, A.L. (2006). Comprehensive identification of phosphorylation sites in postsynaptic density preparations. Mol Cell Proteomics 5, 914-922.

23. Laumonnier, F., Cuthbert, P.C., and Grant, S.G. (2007). The role of neuronal complexes in human X-linked brain diseases. Am J Hum Genet 80, 205-220.

24. Skuse, D.H. (2000). Imprinting, the X-chromosome, and the male brain: explaining sex differences in the liability to autism. Pediatr Res 47, 9-16.

25. Collins, M.O., Yu, L., Coba, M.P., Husi, H., Campuzano, I., Blackstock, W.P., Choudhary, J.S., and Grant, S.G. (2005). Proteomic analysis of in vivo phosphorylated synaptic proteins. J Biol Chem 280, 5972-5982.

26. Gauthier, J., Champagne, N., Lafreniere, R.G., Xiong, L., Spiegelman, D., Brustein, E., Lapointe, M., Peng, H., Cote, M., Noreau, A., et al. (2010). De novo mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. Proc Natl Acad Sci U S A 107, 7863-7868.

27. Eyre-Walker, A., and Keightley, P.D. (1999). High genomic deleterious mutation rates in hominids. Nature 397, 344-347.

28. Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A 103, 1412-1417.

29. Stone, E.A., and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Research 15, 978-986.

30. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30, 3894-3900.

31. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31, 3812-3814.

32. Thomas, P.D., Kejariwal, A., Campbell, M.J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., et al. (2003). PANTHER: a browsable database of gene products organized by biological function, using

curated protein family and subfamily classification. Nucleic Acids Res 31, 334-341.

33. Stone, E.A., and Sidow, A. (2007). Constructing a meaningful evolutionary average at the phylogenetic center of mass. BMC Bioinformatics 8, 222.

34. Smith, N.G., and Eyre-Walker, A. (2001). Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol Biol Evol 18, 982-986.

35. Eyre-Walker, A. (1998). Problems with parsimony in sequences of biased base composition. J Mol Evol 47, 686-690.

36. Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science.

37. Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713-720.

38. Moessner, R., Marshall, C.R., Sutcliffe, J.S., Skaug, J., Pinto, D., Vincent, J., Zwaigenbaum, L., Fernandez, B., Roberts, W., Szatmari, P., et al. (2007). Contribution of SHANK3 mutations to autism spectrum disorder. Am J Hum Genet 81, 1289-1297.

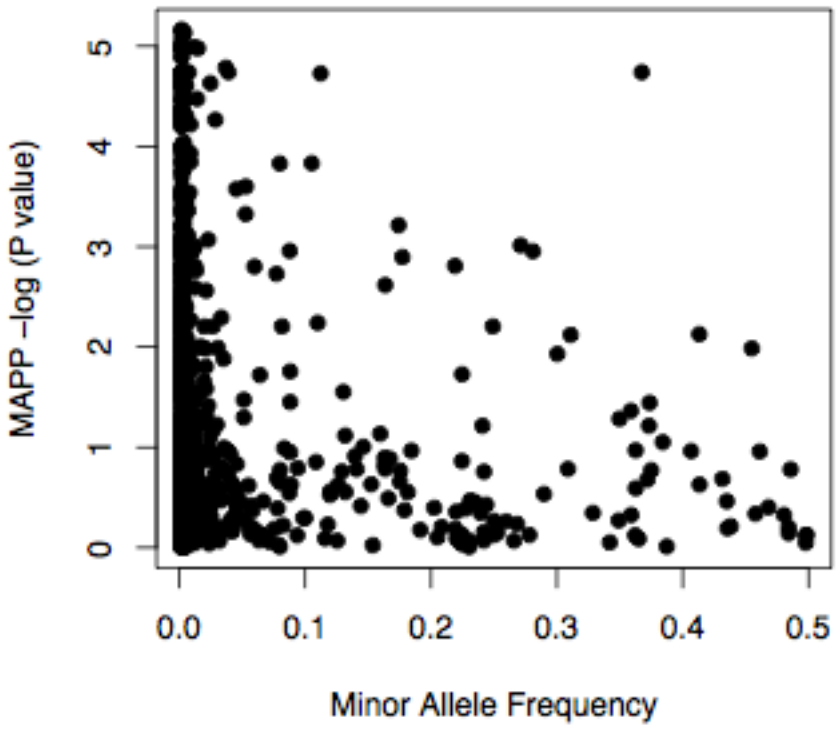


Figure 4-1. MAPP P values by Minor Allele Frequency.
Log p-values of disruption scores inferred using MAPP for missense mutations discovered in the ASD and SCZ cohort and the unaffected cohort plotted vs. minor allele frequency. We observed a clear significant ($P<0.001$) negative correlation as most significant predicted disruptive mutations using the comparative approach are also at the lowest frequency in the cohorts.

Table 4-1. *De novo* mutations discovered by re-sequencing.
Total of 458,877,850 nucleotides of DNA in ASD, SCZ and QTNS control individuals.

Gene	Sample	Diagnosis	Mutation Type	Mutation Location	Chr	Position	Nucleotide Change & Genomic Context	Amino acid /structural change	MAPP *p*-value
SHANK3	S00004	Autism Disorder	Indel	Coding	22	49500342	CGAGATTAGC (G/-)TAAGGGCCAC	Splice site delG	--
IL1RAPL1	S00015	Asperger Syndrome	Indel	Coding	X	29869731	CTTGGTGCTA (TACTCTT/-)GCTGCTTGTA	I367SfsX6	--
GSN	S00099	Asperger Syndrome	Intronic	Intronic	9	123104277	GTGAGGCTGG (C/G)CCTGCCCAGC	Within intron	--
KLC2	S00036	Autism Disorder	Missense	Coding	11	65788196	TACTATCGGC (G/C)GGCACTGGAG	R349P	0.001
KIF5C	S00044	Autism Disorder	Missense	Coding	2	149575030	GGACCGTAAG (C/T)GCTACCAGCA	R802C, R872C	0.001
FLJ16237	S00096	Autism Disorder	Missense	Coding	7	15393678	CCATCACTTA (T/C)TTTCCATATG	F279L	0.472
NRXN1	S02959	Schizophrenia	Indel	Coding	2	50002821	CAGCACACGG (-/ACGG)GTATGGTCGT	G1402DfsX29	--
MAP2K1	S00237	Schizophrenia	Intronic	Intronic	15	64561310	CTTCTTGTAC (G/T)GTCAGGGAGA	Within intron	--
SHANK3	S00161	Childhood Onset Schizophrenia	Missense	Coding	22	49484091	GCATGACACA (C/T)GGCCTGGTGA	R536W	0.051
GRIN2B	S05650	Paranoid Schizophrenia	Missense	Coding	12	13611351	CTTCTACATG (T/G)TGGGGGCGGC	L825V	<0.001
SHANK3	S00285	Schizoaffective, mild MR	Nonsense	Coding	22	49506476	TGCCCGAGAG (C/T)GAGCTCTGGC	R1117X	--

Table 4-1 Continued

KIF17	S00215	Schizophrenia	Nonsense	Coding	1	20886681	GGAGCAGATA (C/A)TTCCTGGATG	Y575X	--
BSN	S00237	Schizophrenia	Silent	Coding	3	49666988	GCACTGCAGT (G/C)GTAGACCTCC	V1665V	--
ATP2B4	S00182	Disorganized Schizophrenia	Silent	Coding	1	201935404	TCATCCGAAA (C/T)GGTCAACTCA	N195N	--
SHANK3	S04261	QNTS - unknown	Missense	Coding	22	49507364	GCCACCAGTG (C/T)CTCCCAAGCC	P1429S	0.107

Table 4-2. Cell line mutations not observed in the blood sample of patients

Gene	Sample	Status	Mutation Type	Mutation Location	Chr	Position	Nucleotide Change & Genomic Context	Amino Acid change	Cell Line Origin
WWC1	S00056	AUT	Silent	Intronic	5	167788404	CAGAAGGAAC (G/A)GTCTGTGTGG	--	UMontreal
PLCB1	S00068	AUT	Missense	Coding	20	8585862	GATTTCACTC (C/T)AGAAGTGTAC	P209L	UMontreal
PLXNB3	S00009	AUT	Missense	Coding	X	152694010	GGTGACCTGG (C/T)GGCCCATTAC	A1431V	UMontreal
DRP2	S00093	AUT	Missense	Coding	X	100383370	AAGCAGGCGA (C/T)GGTGGCCAGT	T203M	UMontreal
PSMD10	S00016	AUT	Silent	Coding	X	107217953	TTGCGGCTTC (T/A)GCTGGCCGGG	S82S	UMontreal
MCF2	S00204	SCZ	Silent	Intronic	X	138512219	TACAGTAATT (A/C)TTCAAGTATT	--	Krebs
SLC7A3	S00218	SCZ	Indel	Intronic	X	70064365	CAGGTCAGTAT (-/A)CAAATGTTTG	InsA 3' of exon	UMontreal
CAMK2A	S00191	SCZ	Missense	Coding	5	149598465	CGAGGATGAA (G/A)ACACCAAAGG	D342N, D353N	UMontreal
GRPR	S00264	SCZ	Missense	Coding	X	16080316	TCCCGGAAGC (G/T)ACTTGCCAAG	R261L	DeLisi/Coriell
ADAM22	S00193	SCZ	Missense	Coding	7	87660477	AAAGTGAACC (G/A)ACAAAGTGCC	R860Q, R889Q, R896Q	UMontreal
ARHGAP6	S00261	SCZ	Missense	Coding	X	11592643	GAGAGTCTCG (G/A)CCCTCGCTTG	G76D	UMontreal
ADD2	S00067	SCZ	Nonsense	Coding	2	70744118	AAAGAAATTC (C/T)GAACCCCCTC	R404X, R710X	UMontreal
RPS6KA6	S00274	SCZ	Silent	Coding	X	83206731	ATCAGCGGTA (T/C)ACTGCTGAAC	Y672Y	DeLisi/Coriell

Table 4-3. **Base-pairs and DNMs surveyed among ASD and SCZ trios with no-family history.**
All statistics are calculated with base-pairs and DNMs calculated for only trios with unaffected families. [a] estimated 2.8% of initial coding sites and 1% of initial intronic sites are CpG sites. [b] (non-CpG sites) + (10 x CpG sites) to account for increased mutation rates. [c] P(X ≥ # functional DNMs | nonfunctional rate)

Site and Mutation Class		**Initial Count***	**CpG[a]**	**Non-CpG**	**Effective Count[b]**
Functional	Non-synonymous bases	96,065,492	2,689,834	93,375,658	120,273,996
	Non-Synonymous DNMs	6	3	3	6
Neutral	Synonymous bases	61,481,915	1,721,494	59,760,421	76,975,357
	Intronic bases	218,145,769	2,181,458	215,964,311	237,778,888
	Silent (synonymous and intronic) DNMs	4	2	2	4
***p* value**	One-Tail Binomial Test[c]	0.003	0.161	0.031	0.008
	Fisher's Exact Test	0.022	0.4041	0.1067	0.032

Table 4-4. Clinical Information for ASD and SCZ Individuals where DNMs were confirmed.
M, male; F, female; IQ, Intelligence Quotient; NA, not available; [1]ASQ: Autism Screening Questionnaire (score >15 = ASD); [2]ADI-R; Autism Diagnostic Interview-Revised (Total cut-off score for the communication and language domain is 8 for verbal subjects and 7 for nonverbal subjects. For all subjects, the cut-off for the social interaction domain is 10, and the cut-off for restricted and repetitive behaviors is 3).

Sample	Final diagnosis	Sex	IQ	Clinical Information	Comorbidity	Familial history of psychiatric illness	Age of Father at Birth (years)
S00004	Autism Disorder	M	NA	ASQ[1] score = 23	None	None	30
S00015	Asperger Syndrome	F	NA	No physical dimorphism. ADI-R[2] scores: social = 17, communication = 13, behaviour = 7	Moderate scoliosis, hypo-pigmented skin patch	None	27
S00036	Autism Disorder	M	NA	ADI-R scores social = 23, communication = 14, behaviour = 4	None	None	31
S00044	Autism Disorder	M	NA	ADI-R scores social = 24, communication = 10, behaviour = 6	Minor anomaly: skull broad and flat on posterior aspect	None	40
S00096	Autism Disorder	M	NA	ASQ score = 19	None	None	38
S00161	Schizo-affective	F	67	Childhood onset, age of onset 11 years. Patient with normal growth, no dysmorphic feature, speech impairment and poor academic and social performance. ASQ score = 1	None	Father has lifetime depression and compulsive behaviour.	NA
S00285	Schizo-affective	M	NA	Schizoaffective disorder with age of onset of 19 years.	Mild mental retardation	Parents are unaffected, two brothers are diagnosed with atypical chronic psychosis	NA
S00215	Schizophrenia	M	NA	No dysmorphic feature, moderate to severe emotional withdrawal	None	None	41

Table 4-5. Comparisons of constraint for genes expressed at the synapse in the ASD and SCZ cohort.

Shown are the counts of point and indel mutations or segregating SNPs for the different categories of variation. *p*-value is the result of Fisher Exact Test comparisons for DNMs versus the two allele frequency classes of SNPs (unique or all).

	Functional	**Non-functional**	**Functional/Non-functional**	***P*- value**
Point and Indel DNMs	10	4	2.5	-
Unique SNPs	785	1652	0.48	0.003*
All SNPs	1306	4878	0.27	<0.001*

Chapter 5 A Population Genetic Approach to Mapping Neurological Disorder Genes using Deep Resequencing

Rachel A. Myers[1,2,3,§], Ferran Casals[1,§], Julie Gauthier[4], Jon Keebler[1,2,3], Adam R. Boyko[5], Carlos D. Bustamante[5], Amelie M. Piton[4], Dan Spiegelman[4], Edouard Henrion[4], Martine Zilversmit[1], Julie Hussin[1], Jacki Quinlan[1], Yan Yang[4], Ron Lafreniere[4], Alexander R. Griffing[3], Eric A. Stone[3], Guy A. Rouleau[2,4]*, and Philip Awadalla[1,2,3,4§]*

[1]Department of Pediatrics, University of Montreal, Montreal, Quebec, Canada; [2]CHU Sainte-Justine Research Centre, University of Montreal, 3175 Cote Sainte-Catherine, Montreal, Quebec, Canada; [3]Bioinformatics Research Centre, North Carolina State University, Raleigh, North Carolina, USA; [4]Centre of Excellence in Neuromics of Université de Montréal, Centre Hospitalier de l'Université de Montréal, and Department of Medicine, Université of Montréal, Montreal, Quebec, Canada; [5]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

*Corresponding authors: Philip Awadalla, PhD, Department of Pediatrics, Faculty of Medicine, University of Montreal, Ste Justine Research Centre, 3175 Cote Sainte-Catherine, Montreal, PQ, Canada H3T 1C5, tel: 514-345 4931 ext. 3393 Email: philip.awadalla@umontreal.ca

Guy A. Rouleau, MD, PhD, FRCPC, OQ, Ste Justine Research Centre, Department of Medicine, Faculty of Medicine, 3175 Cote Sainte-Catherine, Montreal, PQ, Canada H3T 1C5,Montreal, Quebec, Canada H2L 2W5, Tel: 514-890-8000 x 24699, Fax: 514-412-7602 , Email: guy.rouleau@umontreal.ca

[§]these authors provided an equal contribution

Submitted and in Revision at *PLoS Genetics*

Abstract

Deep resequencing of functional regions in human genomes is key to identifying potentially causal rare variants for complex disorders. Here, we present the results from a large sample resequencing (*n*=285 patients) study of candidate genes coupled with population genetics and statistical methods to identify rare variants associated with Autism Spectrum Disorder and Schizophrenia. A consensus among methods identified three genes, *MAP1A*, *GRIN2B* and *CACNA1F* having significant excess of rare missense mutations in either one or both disease cohorts. In a broader context, we also found that the overall site frequency spectrum of variation in these cases is best explained by population models of both selection and complex demography rather than neutral models or models accounting for complex demography alone. Mutations in the three disease-associated genes explained much of the difference in the overall site frequency spectrum among the cases versus controls. This study demonstrates that genes associated with complex disorders can be mapped using resequencing and analytical methods, using sample sizes far smaller than those required by GWA studies. Additionally, our findings support the hypothesis that rare mutations account for a proportion of the phenotypic variance of these complex disorders.

Author Summary

It is widely accepted that genetic factors play important roles in the etiology of neurological diseases. However, the nature of the underlying genetic variation remains unclear. Some of the main questions currently in the field of human genetics relates to the frequency and size effects of genetic variants associated with disease. For instance, the common disease-common variant model states that a reduced set of common variants explains a substantial fraction of cases and forms the basis of GWAS. The rare allele – major effects hypothesis proposes high genetic heterogeneity where a collection of variants, each with a strong effect, produce the disease phenotype. Such variants are kept at low frequencies due to their high penetrance and reduced fitness, limiting detection in GWAS. Resequencing approaches ensure detection of both common and rare variants, but describing significant association with disease remains a challenge. Our approach was to use deep resequencing to capture the full frequency spectrum range of variation, capturing variants at frequency of 1%, and use population genetic methods to test hypotheses of excesses of rare variants in the disease cohorts, and identify genes associated with autism and schizophrenia. We detected an excess of rare variants in the disease cohorts and showed that some genes have negative (deleterious) selection coefficients, which indicates an accumulation of detrimental variants. Our results support the rare variant model describing a component of the genetic etiology of autism and schizophrenia.

Introduction

Genome-wide interrogation approaches for mapping genes often are designed to detect the common variants associated with common phenotypes or disease (CD-CV), generally leaving rare variants undetected or untested [1,2,3,4,5]. The Rare Allele–Major Effects (RAME) model postulates rare (minor allele frequency < 0.01 to 0.05) penetrant variants are key to the genetic etiology of common disease. Functional mutations that lead to an altered amino acid are often deleterious and potentially disease causing, while natural selection either removes these alleles from the population or maintains them at low frequencies relative to the neutral expectations [3,4,6,7,8,9]. Partial genome or candidate gene resequencing in a large numbers of individuals holds the promise of finding both common and rare variants associated with clinically relevant phenotypes [3,4]. While a number of studies have mapped mutations or structural variants associated with neurological disorders like Autism Spectrum Disorder (ASD) (e.g. [10]) and Schizophrenia (e.g. [11]), the RAME model is perhaps better suited for modelling diseases that occur sporadically in families, and for testing whether rare variants contribute to these disorders.

ASD is a neurodevelopmental disorder characterized by stereotyped and repetitive behaviours and impairments in social interactions. Schizophrenia is a chronic psychiatric syndrome characterized by a profound disruption in cognition, behaviour and emotion, which begins in adolescence or early adulthood. The incidence of both ASD and schizophrenia is higher in males than in females [12,13], which points to an important

role of X-chromosome genes in the two diseases. There is significant clinical variability among ASD and schizophrenia patients, suggesting that they are etiologically and genetically heterogeneous. For ASD, genetics clearly plays an important role in the etiology, as revealed by twin and familial studies [14,15,16]. Some susceptibility regions have been identified through whole genome linkage analyses [17,18], although they rarely coincide among the different studies [19].Additionally, *de novo* mutations have been observed [20] and are candidate variants in sporadic cases of either ASD or schizophrenia, but they explain a small percentage of the phenotypic variation. Together, these observations suggest that disruption of numerous genes by rare but penetrant mutations could represent an important cause of ASD. Schizophrenia has an estimated heritability of 80% [21], and it has been recently associated with common variants at the MHC locus [22,23,24]. There are also several observations suggesting a causal link between rare mutations and schizophrenia, such as familial cases showing rare inherited copy number variants (CNVs) [25] and associations of both paternal age and decreased fertility[2].Recent studies also reported *de novo* CNVs in schizophrenia, providing further support for the rare variant hypothesis in non-familial cases [26,27,28,29].

We hypothesized that capturing genetic variation at low frequencies (rare variants ≥ 1% frequency) in a large set of genes expressed in the brain will contribute significantly to our understanding of the genetic basis of ASD and schizophrenia. If the RAME model is relevant to these two diseases, the expectation is an enrichment of rare deleterious mutations among individuals diagnosed with ASD and schizophrenia. Here,

we use population genetic and other statistical methods to analyze a resequencing dataset to map genes related to these two disorders. With this approach, we identify candidate genes by testing for genes hosting an excess of rare missense variants among individuals affected with ASD and schizophrenia, test for selection at the gene level in each disease cohort, and assess the impact of the candidate genes on the distribution of missense allele frequencies for each disease cohort.

Results

Variants Discovered from Deep Resequencing

We resequenced 408 selected brain-expressed genes in 142 ASD and 143 schizophrenia-affected individuals. The ASD and schizophrenia cohorts had global ethnicity representation yet were predominantly European with a large French Canadian sub-group. It is crucial to exclude ethnic and genetic outliers when analyzing rare variants because such samples contain private alleles from other populations. While the results from the software *structure* [30] revealed no population structure, we identified and removed potential ethnic outliers from the analysis using self reported ethnicity and a principle component analysis (Figure 5-1). The result was a sample of European ancestry individuals: ASD (n=102) and schizophrenia (n=138). Two hundred eighty-five samples from the Québec Newborn Twin Study (QTNS) [31] were screened for self reported European ancestry (n = 240) and were used as controls for the ASD cohort. Thirty-eight (19 autosomal, 19 X-linked) of the 408 brain expressed genes were sequenced in the

QTNS controls, including any gene with *de novo* mutations previously described in one of the disease cohorts [32] or with potential protein disrupting mutations.

We identified a total of 5,396 segregating sites in the disease cohorts, including 1,111 missense and 11 nonsense variants (Tables 5-1 and 5-2), from lymphoblastoid cell DNA. As expected, there was a reduction in nucleotide diversity (π) on the X chromosome relative to autosomes by a ratio of 0.76 (Table 5-1), consistent with neutral expectations for the reduced effective population size of the X chromosome [33]. The ratio of the male to female population mutation rate [34] was estimated to be 6.37, slightly higher but similar to previous estimates of the male mutation rate being four times the female mutation rate [35]. We estimated the population nonsense mutation (θ_W per base pair) rate to be 3.9×10^{-7} for ASD and 7.5×10^{-7} for schizophrenia.

Rare Missense Variants Show Increased Predicted Detrimental Effects

Given both the high heritability of ASD and schizophrenia and the low replication of common variant associations, rare variants may explain a component of the genetic etiology for these diseases. Based on this hypothesis, we expect the site frequency spectrum (SFS) of missense variants to show an excess of deleterious, low frequency variants in our disease cohorts relative to either neutral expectations or controls. We analyzed the proposed detrimental effects of missense variants by estimating the potential functional effect of the missense variants observed using *MAPP* [36]. *MAPP* predicts the severity of a mutation based on conservation in a multispecies protein alignment and the

physiochemical properties of the amino acids. Severity scores indicated that 19% of the nonsynonymous variants, in 47% of the genes with one or more missense variants in the ASD and schizophrenia cohorts were likely to adversely affect protein function, considering a threshold for *MAPP* scores of ten. *MAPP* scores are significantly higher for rare versus common variants (average 7.59 vs. 5.91, Mann-Whitney test $P = 1.45\times 10^{-4}$, Figure 5-2), and the proportion of variants with high *MAPP* scores (>10) is also significantly higher in the rare variants (21%) than in the common variants (11%) ($\chi^2 = 8.3$, $P = 0.0039$). These results are consistent with the presence of deleterious alleles being maintained at low frequency by selection. However, we did not observe an excess of high *MAPP* scores in the cases relative to those in the QTNS controls suggesting that the above observations are merely indicative that rare mutations are likely subject to selection across all populations and not an indication of an overall excess of deleterious mutations across all cognitive related genes screened here.

Individual Genes with Excess of Missense and Rare Missense Variants

We applied three different methods to identify genes harboring an excess of rare missense variants in the disease cohorts. First, within each disease cohort, we tested the ratios of missense to silent variants for each gene relative to the remaining genes pooled; separately for X-linked and autosomal loci. The excess of missense variants was assessed using Fisher's Exact Test. In both cohorts, *MAP1A* exhibited a significant excess of missense compared to silent variants (Figure 5-3, Table 5-3) (ASD $P = 0.04$,

schizophrenia $P = 0.03$ after Bonferroni correction) and 23 of the 29 missense variants described in this gene are rare. *MAP1A* also exhibits an excess of missense variants when compared to the total counts of autosomal missense and silent variants in the control cohort (ASD-Control $P = 0.009$, Schizophrenia-Control $P = 0.008$). We further tested the total predicted effect of each gene by summing the *MAPP* scores for all missense variants within a gene, and testing the ratio of summed *MAPP* scores to the counts of silent variants relative to all other genes. While *MAP1A* was not significant after multiple testing correction, *GRIN3A* showed an increased ratio of summed *MAPP* score to silent variant count, relative to all other genes ($P = 0.011$) in the ASD cohort.

Second, we tested for an excess of individuals bearing rare missense variants, using Li and Leal's collapsing method [37]. We i) contrasted the ASD and schizophrenia cohorts to each other ($n = 277$ genes with one or more missense variant), ii) compared ASD to the QTNS controls ($n = 26$ genes with one or more missense variant), and iii) compared the schizophrenia cohort to the QTNS controls ($n = 26$ genes with one or more missense variant), and corrected for multiple testing. Here, for every gene, the number of individuals carrying at least one rare missense variant is compared to the number of individuals without rare missense variants (see Materials and Methods). Although the ASD–Schizophrenia comparison revealed no significant genes, three genes having an excess of individuals with rare missense variants in the disease cohorts relative to their respective controls. *GRIN2B* and *CACNA1F* exhibited an excess of individuals with rare variants in the ASD cohort relative to QTNS controls (Bonferroni adjusted $P = 0.026$,

and $P = 0.031$ respectively, Table 5-4). The schizophrenia cohort showed significant excess of individuals hosting rare missense variants *GRIN2B* ($P = 0.044$). We repeated this analysis considering only missense variants predicted to have detrimental effects on protein function. When evaluating rare missense mutations with *MAPP* predicted severity > 10, we found *GRIN3A* to exhibit an excess of individuals hosting rare detrimental missense variants in the ASD cohort relative to QTNS controls ($P = 0.034$).

Finally, a population genetic method was used to estimate per-gene selection parameters. An extension of the McDonald Krietman test, implemented in *mkprf* [38], was used to obtain estimates of γ per gene, based on our observed polymorphisms within humans and substitutions between humans and an out-group (*Pan troglodytes*, see Materials and Methods). Negative γ values are estimated when an excess of missense polymorphisms relative to divergent sites is observed [39]. Here, we compared selection coefficients between genes to detect genes enriched for missense variants in one cohort compared to another. The overall distribution of γ values among genes was similar between ASD, schizophrenia and the QTNS controls (Figures 5-4A-C). When we contrasted ASD and schizophrenia γ estimates to those estimated from a Western-European population[39], we found both the ASD and schizophrenia cohorts have significantly more negative γ distributions (Wilcox paired test, ASD $P = 0.0026$, schizophrenia $P = 0.0005$). However, the QTNS controls do not show significant differences with the European populations ($P = 0.18$). While the ASD and schizophrenia γ distributions are similar, we do find some genes to vary by cohort. *NOS1* had

significantly more positive γ estimates in ASD compared to schizophrenia (Figure 5-4), and had a positive γ estimate in the European samples. Both results suggest an excess of missense variants in the schizophrenia cohort. We also observed a significant difference in γ for *CACNA1F* between the QNTS and ASD cohorts (Figure 5-5), with the disease cohort having the lower γ estimate.

Excess of Rare Deleterious Variants at Autosomal Loci Among ASD and Schizophrenia Individuals

Having identified a number of individual genes with an excess of deleterious rare alleles, we examined the contribution of these individual genes to genome-wideγ estimates. We used the Poisson Random Fields method implemented in *prfreq* [40] to ask 1) whether the site frequency spectrum of the missense variants showed evidence of selection relative to the SFS of silent and intronic variants, 2) if the strength of selection in the missense SFS differed significantly between the disease cohorts and the control cohort, and 3) if removing the disease associated genes affects γ estimates. The *prfreq* approach is used to infer the demographic and selection parameters of variants from the overall site frequency spectrum of variation among all loci. Estimates of population selection parameters, $\gamma = 2N_e s$, where N_e is the effective population size, and s is the selection coefficient, will be negative when an excess of low frequency variants are observed, suggesting an accumulation of deleterious variants[39] through negative selection.

We first estimated the demographic parameters for the different cohorts using the silent and intronic variants discovered among all autosomal genes with one or more variant (265 for disease cohorts, 19 for controls). While fixing the estimated demographic parameters, the selection coefficients (γ) were inferred from the missense variants SFS (198 genes with one or more missense variant in the disease cohorts, 15 such genes for the controls; Table 5-5 and Figure 5-6, Supplementary Material) thus generating likelihood measure of selection model. Additionally, we estimated the likelihood of the missense SFS given the demographic model and the Wright-Fisher neutral model. We contrasted (Table 5-5) the demographic model first to a neutral model (two times the difference in log likelihood, $P< 0.001$) and second to a demographic with selection model to the demographic only model ($P< 0.001$), noting that the model with selection and demography best fits the observed SFS. Estimates of γ were more negative in the two disease cohorts than for 15 autosomal coding regions sequenced in the controls ($\gamma_{control}$ = -740, γ_{ASD} = -920, and γ_{SCZ} = -1,100, Table 5-5 and Figure 5-6 at Supplementary material), specifically γ_{SCZ} was significantly more negative than the $\gamma_{control}$ ($P = 0.01$), although this was not observed among ASD samples ($P = 0.12$) (see Materials and Methods). These observations indicate an excess of low frequency missense variants in our disease cohorts.

We developed an empirical distribution of γ estimates by removing each gene from the SFS and re-estimating γin each cohort. For ASD, when excluding *MAP1A*, *GRIN3B* and *RPGRIP1*, either individually or combined, the γ values became more

positive (less deleterious; $\gamma = -880$, $P[\gamma \geq -880|$ empirical distribution$] = 0.015$ and -800 respectively) than when estimated for all genes combined ($\gamma = -920$ for all the genes) (Figure 5-7). These values are also closer to the γ value estimated in controls (-740, see above) and are at the most positive end of the empirical distribution of γ values (Figure 5-7A). In the schizophrenia cohort, we estimated γ values when removing *MAP1A* and *GRIN3B* were $\gamma = -1{,}020$ ($P[\gamma \geq -1020 |$ empirical distribution$] = 0.01$) and $\gamma = -980$ ($P[\gamma \geq -920 |$ empirical distribution$] = 0.005$), respectively (Figure 5-7B). Excluding from the SFS simultaneously *MAP1A* and *GRIN3B*, the estimated γ decreases to -880, again closer to the γ in controls, and are the most positive values when compared to the empirical distribution. These results indicate that the presence of a few genes, enriched with rare missense variants in the disease cohorts, is enough to alter the global SFS among all of our candidate genes. In our case, the overall excess of deleterious missense variants and the reduction in γ observed in the two disease cohorts as compared to a neutral cohort is mainly caused by a very few candidate genes.

Discussion

Through the analysis of candidate genes using population genetic approaches that specifically analyze both the function and accumulation of rare variants in disease cohorts, we have mapped a number of candidate genes associated with ASD and schizophrenia in two cohorts with samples sizes substantially smaller than those required

for association studies. Given the difficulty associated with detecting genes exhibiting excesses of rare variants, multiple methods are required to map genes associated with disease. In essence, population genetic model-based approaches are designed to deal with this type of data even when sample sizes are limited compared to the sizes required in most GWAS studies. In both the ASD and schizophrenia cohorts, *GRIN2B* and *MAP1A* harbor a statistically significant excess of rare missense variants, while the excess of rare missense variants in *CACNAF1* was restricted to the ASD cohort. The involvement of a particular gene (e.g. *GRIN2B* or *MAP1A*) in the etiology of both disorders may reflect a critical role in neurodevelopment [41]. Population genetic models incorporating demographic and selection processes, implemented in *prfreq,* corroborated this result for *MAP1A* and *CACNA1F*, and identified three new candidate genes: *GRIN3B*, *NOS1*, *RPGRIP1* (Table 5-6).

Most of our mapped genes (Table 5-6) have been previously implicated in neurological disorders or in neurodevelopment. *MAP1A*, a member of the microtubule-associated MAP1 proteins family, is predominantly expressed in adult neurons and is involved in axon and dendrite development. Among other interactions, *MAP1A* participates in the linking of DISC1 to microtubules [42]. DISC1 is a protein that was described as related to the pathogenesis of schizophrenia by linkage analysis of a Scottish family [43,44] and confirmed among Finnish cohorts [45,46]. The association between schizophrenia and eleven genes interacting with DISC1 (including *MAP1A*) was explored in Finish families, with significant results in three of the candidates but not for *MAP1A*

[47]. Among calcium channel classes of genes, such as *CACNA1F*, we found a significant excess of rare variants and two segregating inframe indels at CDS position 807 falling in a glutamic acid-rich coiled domain. *CACNA1F* has been previously associated to schizophrenia [48] and mutations have also been described for other neurological disorders [49]. Finally, two independent meta-analyses corroborate our findings for a role of *GRIN2B* in the etiology of schizophrenia [50,51]. *GRIN2B* codes a subunit of the glutamate and N-methyl-D-aspartate (NMDA) receptor. There exist several NMDA receptors that are constructed with one or more isoforms of the NR1 subunit (*GRIN1*) in different combinations with NR2 (*GRIN2A*, *GRIN2B*, *GRIN2C* and *GRIN2D* genes) and NR3 subunits (*GRIN3A* and *GRIN3B*) [52]. Some studies have suggested a relationship of decreased expression and abnormalities in NMDA receptors with schizophrenia [53,54]. Additionally, we observed key results for other NMDA related genes. *GRIN3B* exhibited an excess of detrimental variants, two different analyses revealed significantly higher *MAPP* scores in *GRIN3A* in the ASD cohort, while we observed a nonsense mutation in both *GRIN2C* and *GRIN3A*in our cohorts. In the case of *GRIN2B*, we also observed a coding indel which results in an amino acid insertion at cds position 1,353.This collection of rare and functional changes in the *GRIN* gene family points to an important role of NMDA receptors in these neurological disorders.

Evidence for disease association using all methods use is ideal, however it is unlikely, as each statistical test evaluates a different aspect of the data and the ability to use each test varies by chromosome and cohort. The Li and Leal collapsing method

evaluates the accumulation of rare missense variants in genes in cases relative to controls, and is independent of silent variants. The ratio of missense to silent mutations is cohort specific and used to identify genes that have an excess of missense mutations conditioned on the number of silent mutations relative to the cohort-wide average. Since these two methods test different aspects of the data, congruent results are not expected. For example, a significant result using the Li and Leal collapsing method may not be significant when testing the ratio of rare missense mutations to the number of silent within a gene. A gene can have a minimum of two rare missense variants and generate a significant result in the collapsing method, but will not reach significance when testing the ratio of rare missense to silent variants. We expect genes with high missense to silent variant ratios to also have negative gamma estimates using *mkprf* and to alter *prfreq* gamma estimates. Additionally, genes showing statistical significance in collapsing method results are also expected to alter gamma estimates using *prfreq*. We observe these trends in *GRIN2B* (Table 5-6) in both diseases. This gene has a significant collapsing method results and a negative gamma estimate using *mkprf*, and this estimate is more negative than the control cohort. *MAP1A* has a similar story, with a significant ratio of missense to silent variants and a strong impact on *prfreq* gamma estimates in both disease cohorts. Due to the limited resequencing data in controls, we were unable to apply the collapsing method to this gene, and lack of divergence data in Bustamante *et al* [39] prohibited us from estimating gamma with *mkprf*. Like *GRIN2B and MAP1A, CACNA1F* also shows concordance among the methods in the ASD cohort; including a nominally

significant missense to silent variant ratio, a significant collapsing method result, and a negative shift in the gamma estimate relative to a control cohort. Since *CACNA1F* lies on the X chromosome, it could not be evaluated using *prfreq*. For our three main candidate genes, *GRIN2B, MAP1A*, and *CACNA1F,* we see concordant results among multiple methods when data availability allows testing by the multiple methods.

Previous studies have shown reduced deleterious selection (less negative γ) in genes related to complex diseases compared to those genes associated with Mendelian disease or cancer [55]. This unexpected pattern maybe explained by a late-onset effect of these genes, or by a potential enrichment of positively selected genes among the genes involved in complex disease [55,56]. Perhaps a more plausible explanation is that genes which accumulate either rare inherited or de novo mutations are also more likely to accumulate rare mutations which individually have not just high impact, as in the case of Mendelian disorders, but either have intermediate impact either alone or in aggregate. In our case, estimated γ values in the two disease cohorts are lower than those estimated in the neutral cohort, likely due to the excess of missense polymorphic variants and reduced reproductive fitness of the disease cohorts. A proportion of the rare missense is likely to have a negative functional impact and natural selection prevents them from reaching higher frequencies. The question then remains as to whether such mutations work independently or whether there are hidden interactions that are not captured, or that we lack sufficient power to detect.

In this paper we hypothesized that rare variants in neurologically expressed genes are associated with disorders such as ASD and schizophrenia. In our candidate gene resequencing survey, we identified multiple rare functional variants in genes specific to either or common to both disorders. Our findings support a rare allele-major effect model as we have uncovered significant excess of rare variants in our disease cohorts. It remains an open question if ASD and schizophrenia are caused by variants found in a reduced set of genes such as DISC1 or NMDA receptor related genes, or in a high number of genes defined by a common functional class or pathway/network. In this case lack of replication between populations could be observed for individual genes yet particular pathways or gene families could arise as having a main role in the etiology of the disease.

Materials and Methods

Candidate gene selection

In total, 408 genes were selected for sequencing: 122 in the X chromosome and 286 autosomal genes from a comprehensive list of potential synaptic genes (n=5,000) based on published studies and databases [57,58,59,60,61,62]. X-chromosome synaptic genes were chosen due to the excess of affected males as compared to females in schizophrenia [12] and ASD [13], and since many genes affecting neurodevelopmental brain diseases have been found on the X-chromosome [63]. Autosomal genes implicated in synapse function, including those encoding glutamate receptors and their interactors were also chosen, because glutamate signalling is strongly implicated in synapse function

[64]. A total of 38 genes (19 autosomal and 19 X-linked) showing extreme Tajima's D values and or had *de novo* mutations in the disease cohorts were chosen for sequencing in the controls.

Samples

ASD subjects: Subjects diagnosed with autism spectrum disorders and both of their parents were recruited from clinics specializing in the diagnosis of Pervasive Developmental Disorders (PDD), rehabilitation centers, and specialized schools in the Montreal and Quebec regions, Canada [65]. Subjects with ASD were diagnosed by child psychiatrists and psychologists specialized in the evaluation of ASD. All subjects were diagnosed using the Diagnostic and Statistical Manual (DSM) of Mental Disorders criteria from patients in the Montreal and surrounding area, and depending on the recruitment site, either the Autism Diagnostic Interview-Revised or the Autism Diagnostic Observation Schedule was used. In addition, the Autism Screening Questionnaire (ASQ) was also completed for all of our subjects. All samples were French-Canadian. Furthermore, all proband medical charts were reviewed by a child psychiatrist expert in PDD to confirm diagnoses. Exclusion criteria were: (1) an estimated mental age <18 months, (2) a diagnosis of Rett Syndrome or Childhood Disintegrative Disorder and (3) evidence of any psychiatric and neurological conditions, specifically: birth anoxia, rubella during pregnancy, fragile-X disorder, encephalitis, phenylketonuria, tuberous sclerosis, Tourette and West syndromes. Subjects with these

conditions were excluded based on parental interview and chart review. However, participants with a co-occurring diagnosis of semantic-pragmatic disorder (due to its large overlap with PDD), attention deficit hyperactivity disorder (seen in a large number of patients with AD during development) and idiopathic epilepsy (which is related to the core syndrome of AD) were eligible for the study.

Schizophrenia subjects: The schizophrenia subjects were collected from among several large schizophrenia clinical genetic research centers worldwide. These include: (A) L.E. Delisi cohort collected in the USA and Europe [66]. Dr. DeLisi and her collaborators had identified and collected over 500 families with schizophrenia or schizoaffective disorder in at least two siblings over the last two decades. Diagnoses were made by using the DSM-III-R criteria on the basis of structured interviews, review of medical records from all hospitalizations or other relevant treatments, and structured information obtained from at least one reliable family member about each individual. Two independent diagnoses (one made by L.E.D.) were made for each individual in the study. In cases of disagreement between the diagnosing clinicians, a third diagnostician was consulted, and final diagnoses were made by consensus after discussion. In cases where there was a sibling diagnosed with schizophrenia, the schizophrenia with earlier age of onset and more definite schizophrenia diagnosis was selected for the initial screening. (B) R. Joober cohort: Dr. Joober has collected over 300 schizophrenia families in Montreal in the past 10 years[67]. The same clinical assessment procedures have been followed as in (A). In addition, extensive pharmacological data have been collected in

this cohort. (C) J. Rapoport cohort (USA): collection includes Childhood Onset Schizophrenia cases (COS) [68]. Individuals in this cohort known to carry the VCFS deletion on chromosome 22q11 were excluded. All patients met DSM-IIIR/DSM-IV criteria for schizophrenia or psychosis not otherwise specified (NOS), had premorbid full-scale IQ scores of 70 or above and onset of psychotic symptoms by age 12 years. (D) Marie-Odile Krebs cohort (4 cases) was collected in France: All subjects were examined according to the standardized Diagnostic Interview for Genetic Studies (DIGS 3.0) [69]. Family histories of psychiatric disorders were also collected using the Family Interview for Genetic Studies (FIGS). All DIGS and FIGS have been reviewed by two or more psychiatrists for a final consensus diagnosis based on DSM-IIIR or DSM-IV at each centre. Exclusion criteria for all subjects included neurologic hard signs (referring to any symptoms or neurological conditions that can come with psychosis (and not related to schizophrenia such as Parkinson, Alzheimers, etc.), a history of head trauma and substance abuse or dependence. Institutional ethical approval for the study and informed consent was obtained for all study participants.

From over 500 ASD and 1,000 schizophrenia families, 122 males and 20 females were selected for the ASD cohort, and 95 males and 48 females were selected for the schizophrenia cohort. We selected patients where blood samples were available so that *de novo* mutations can be accurately validated. All parentage was tested using 17 microsatellite markers. The random population cohort (150 males; 135 females) consist of unrelated blood DNA samples collected for the Quebec Newborn Twin Study (QNTS)

[31] where DNA samples were available for both parents and both twins (either monozygotic or dizygotic), however only one sibling was chosen randomly for sequencing.

DNA preparation, Sequencing and SNP calling

DNA samples were available for all affected and unaffected individuals and parents. For certain individuals where blood DNA was limited, we used DNA isolated from an Epstein-Barr Virus transformed lymphoblastoid cell line derived from the individual for the screen. The ASD cell lines samples had been frozen or regrown a maximum of two times. Genomic DNA was extracted from peripheral blood lymphocytes for each individual using Puregene extraction kits (Gentra System, USA). In all cases, rare variants were confirmed by sequencing both parents and using blood-derived DNA to rule out variations having arisen during production or growth of the lymphoblastoid cell line. Primers were designed using the Exon Primer program from the UCSC genome browser. PCR products were sequenced at the Genome Quebec Innovation Centre in Montreal, Canada (www.genomequebecplatforms.com/mcgill/) on a 3730XL DNA Analyzer System. PolyPhred (v6.0) and Mutation Surveyor (v3.10, Soft Genetics Inc.) were used for mutation detection analysis. Initial screens were done on cell-line DNA from samples to conserve blood sample DNA. PolyPhred (v6.0) scores of 40 or higher were used as the threshold cut-off for all sequencing reads. When reads did not meet this

criterion they were resequenced. Chromatograms for all rare variants (singletons or homozygous doubletons) were manually checked.

Population Structure

Structure [30] was used to analyze the degree of population structure. Analysis of all samples showed no significant population structure as a large proportion of samples were of the same ethnicity. Principal Component Analysis (PCA) is more susceptible to samples with excess of private alleles and revealed genetic variability between individuals. We removed samples of non-European ancestry (self-reported ethnicity, ethnic outliers) and used *eigensoft* [70] to identify and remove remaining genetic outliers (defined below). All autosomal variants excluding those with calls in less than 20 per cohort were used for PCA analysis (4,645 SNPs). We used the LD correction and calculated the top 10 principal components (PCs) and removed individuals with PC projections >two standard deviations or more from the mean, for all significant principal components, using 10 iterations. Individuals exhibiting excess of rare variants genome-wide are likely to be genetic outliers and readily identifiable with PCA and removed, while individuals with an excess of rare variants in specific genes are retained. To ensure the PCA outliers were true outliers, we used a one-sided *t*-test to assess if the proportion of missense singletons in each PCA outlier was higher than in PCA retained samples across all autosomal genes. *Structure* (admixture model) was used to reassess levels of population structure within the final sample set (results not shown).

Computational Inferences of Mutation Severity

MAPP was used to predict severity and assign scores to each missense variant. Orthologous protein sequences were obtained using the Galaxy Browser (http://main.g2.bx.psu.edu/) and the UCSC Human Genome Browser (hg18, http://genome.ucsc.edu/cgi-bin/hgGateway) to generate columns of aligned orthologous amino acids. *MAPP* scores and p values were calculated as shown in Stone and Sidow [36]. *MAPP* assesses variation observed at each amino acid position with respect to six physicochemical properties (hydropathy, polarity, charge, side-chain volume, free energy in alpha-helical conformation, and free energy in beta-sheet conformation) after weighting each protein sequence to mitigate the influence of phylogeny.

Statistical Analysis

Estimating population mutation rates and selection coefficient: *Hclust* [71] was used to calculate Tajima's D, θ_W, and π for each gene in each cohort. Demographic parameters were inferred from the folded site frequency spectrum (SFS) of silent and intronic variants using *prfreq* [40]. Conditioning on the demographic parameters from the silent and intronic SFS, the likelihood of the missense SFS was estimated for three models: a Wright-Fisher neutral population, the demographic model inferred form the silent and intronic variants, and a demographic with a selection coefficient (γ) model. *P*

values of model improvement were estimated by comparing the likelihoods, assuming a χ^2 distribution of two times the differences in the likelihoods between the models. To compare the γ values estimated in the ASD and schizophrenia cohort to the γ values estimated in the control cohort, we computed the likelihood of the control parameters estimates in the schizophrenia and ASD cohorts, and compared these likelihoods to the likelihoods of the ASD and schizophrenia parameters estimates. We analyzed the effect of each gene with at least one missense variant (n=198 autosomal genes) on the SFS by estimating γ in all the SFSs resulting from excluding the missense variants of each individual gene, in each cohort.

Estimating individual gene selection coefficients: The gene specific selection coefficient, γ, was calculated with the *mkprf* program [38] for each cohort (ASD, Schizophrenia, and QTNS Controls). The number of synonymous and nonsynonymous changes between humans and chimpanzees was obtained from Bustamante, *et al*. [39] for 244 genes. Additionally, the European ancestry samples also reported in from Bustamante, *et al*. [39] were used as a secondary control data set for this analysis.

Within gene excess of (rare) missense variants: Fisher's exact test was used to detect deviations in the missense to silent variant ratio within genes. For each gene, the ratio of missense (or rare missense) to silent variants was contrasted to the same ratio in all remaining genes. This analysis was conducted in the ASD and schizophrenia cohorts testing autosomal and X-linked genes separately. We used the Bonferroni correction for multiple tests (n = 277). Excess of predicted deleterious load within genes was evaluated

by summing the *MAPP* scores for all missense variants within a gene and testing the ratio of summed *MAPP* scores to silent variants within the gene relative to all other genes. This was done autosomes and X-linked genes separately, and in each disease cohort.

Genes with excess of individuals bearing rare missense variants: To identify genes with an excess of individuals bearing rare missense variants, we used Li and Leal's collapsing method [37]. For ASD vs. schizophrenia, ASD vs. QTNS, and schizophrenia vs. QTNS controls, the number of individuals with at least one rare missense mutation and the number of individuals with no rare missense variants was determined for each cohort, and these counts made up the cells of the two by two table. We assessed statistical significance using Fisher's Exact test and used Bonferroni's correction for multiple tests ($n_{\text{ASD-SCZ}} = 277$, $n_{\text{ASD – QTNS Controls}} = 26$, $n_{\text{SCZ- QTNS Controls}} = 26$). This analysis was repeated, considering only missense variants with a *MAPP* score > 10.

Acknowledgements

We would like to thank all the families involved in this study. This work was supported by Genome Canada and Génome Québec, and received co-funding from Université de Montréal for the 'Synapse to Disease' (S2D) project as well as funding from the Canadian Foundation for Innovation to both G.A.R and P.A and co-funding from the MDEIE of Quebec. G.A.R. holds the Canada Research Chair in Genetics of the Nervous System; P.A. holds the Genome Quebec Award in Population and Medical Genomics.

References

1. Ma DQ, Cuccaro ML, Jaworski JM, Haynes CS, Stephan DA, et al. (2007) Dissecting the locus heterogeneity of autism: significant linkage to chromosome 12q14. Mol Psychiatry 12: 376-384.

2. McClellan JM, Susser E, King MC (2007) Schizophrenia: a common disease caused by multiple rare alleles. Br J Psychiatry 190: 194-199.

3. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124-137.

4. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11: 2417-2423.

5. Smith DJ, Lusis AJ (2002) The allelic structure of common disease. Hum Mol Genet 11: 2455-2461.

6. Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet 1: 40-47.

7. Eyre-Walker A, Keightley PD (1999) High genomic deleterious mutation rates in hominids. Nature 397: 344-347.

8. Giannelli F, Anagnostopoulos T, Green PM (1999) Mutation rates in humans. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from hemophilia B. Am J Hum Genet 65: 1580-1587.

9. Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80: 727-739.

10. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, et al. (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459: 528-533.

11. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, et al. (2009) Common variants conferring risk of schizophrenia. Nature 460: 744-747.

12. Lewis DA, Levitt P (2002) Schizophrenia as a disorder of neurodevelopment. Annu Rev Neurosci 25: 409-432.

13. Skuse DH (2000) Imprinting, the X-chromosome, and the male brain: explaining sex differences in the liability to autism. Pediatr Res 47: 9-16.

14. Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, et al. (1995) Autism as a strongly genetic disorder: evidence from a British twin study. Psychol Med 25: 63-77.

15. Steffenburg S, Gillberg C, Hellgren L, Andersson L, Gillberg IC, et al. (1989) A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. J Child Psychol Psychiatry 30: 405-416.

16. Bolton P, Macdonald H, Pickles A, Rios P, Goode S, et al. (1994) A case-control family history study of autism. J Child Psychol Psychiatry 35: 877-900.

17. Gupta AR, State MW (2007) Recent advances in the genetics of autism. Biol Psychiatry 61: 429-437.

18. Klauck SM (2006) Genetics of autism spectrum disorder. Eur J Hum Genet 14: 714-720.

19. Yang MS, Gill M (2007) A review of gene linkage, association and expression studies in autism and an assessment of convergent evidence. Int J Dev Neurosci 25: 69-85.

20. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007) Strong association of de novo copy number mutations with autism. Science 316: 445-449.

21. Tandon R, Keshavan MS, Nasrallah HA (2008) Schizophrenia, "just the facts" what we know in 2008. 2. Epidemiology and etiology. Schizophr Res 102: 1-18.

22. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748-752.

23. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, et al. (2009) Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460: 753-757.

24. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, et al. (2009) Common variants conferring risk of schizophrenia. Nature 460: 744-747.

25. Xu B, Woodroffe A, Rodriguez-Murillo L, Roos JL, van Rensburg EJ, et al. (2009) Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. Proc Natl Acad Sci U S A 106: 16746-16751.

26. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320: 539-543.

27. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, et al. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 40: 880-885.

28. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, et al. (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455: 232-236.

29. The_International_Schizophrenia_Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455: 237-241.

30. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

31. Lemelin JP, Boivin M, Forget-Dubois N, Dionne G, Seguin JR, et al. (2007) The genetic-environmental etiology of cognitive school readiness and later academic achievement in early childhood. Child Dev 78: 1855-1869.

32. Awadalla P, Gauthier J, Myers RA, Casals F, Hamdan FF, et al. (2010) Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. Am J Hum Genet 87: 316-324.

33. Hartl DL, Clark AG (2007) Principles of population genetics. Sunderland, Mass.: Sinauer Associates. xv, 652 p. p.

34. Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T (1987) Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harb Symp Quant Biol 52: 863-867.

35. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297-304.

36. Stone EA, Sidow A (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 15: 978-986.

37. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311-321.

38. Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, et al. (2002) The cost of inbreeding in Arabidopsis. Nature 416: 531-534.

39. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.

40. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4: e1000083.

41. Carroll LS, Owen MJ (2009) Genetic overlap between autism, schizophrenia and bipolar disorder. Genome Med 1: 102.

42. Halpain S, Dehmelt L (2006) The MAP1 family of microtubule-associated proteins. Genome Biol 7: 224.

43. Blackwood DH, Fordyce A, Walker MT, St Clair DM, Porteous DJ, et al. (2001) Schizophrenia and affective disorders--cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. Am J Hum Genet 69: 428-433.

44. St Clair D, Blackwood D, Muir W, Carothers A, Walker M, et al. (1990) Association within a family of a balanced autosomal translocation with major mental illness. Lancet 336: 13-16.

45. Ekelund J, Hovatta I, Parker A, Paunio T, Varilo T, et al. (2001) Chromosome 1 loci in Finnish schizophrenia families. Hum Mol Genet 10: 1611-1617.

46. Ekelund J, Lichtermann D, Hovatta I, Ellonen P, Suvisaari J, et al. (2000) Genome-wide scan for schizophrenia in the Finnish population: evidence for a locus on chromosome 7q22. Hum Mol Genet 9: 1049-1057.

47. Tomppo L, Hennah W, Lahermo P, Loukola A, Tuulio-Henriksson A, et al. (2009) Association between genes of Disrupted in schizophrenia 1 (DISC1) interactors and schizophrenia supports the role of the DISC1 pathway in the etiology of major mental illnesses. Biol Psychiatry 65: 1055-1062.

48. Wei J, Hemmings GP (2006) A further study of a possible locus for schizophrenia on the X chromosome. Biochem Biophys Res Commun 344: 1241-1245.

49. Hope CI, Sharp DM, Hemara-Wahanui A, Sissingh JI, Lundon P, et al. (2005) Clinical manifestations of a unique X-linked retinal disorder in a large New Zealand family with a novel mutation in CACNA1F, the gene responsible for CSNB2. Clin Experiment Ophthalmol 33: 129-136.

50. Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK, et al. (2008) Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nat Genet 40: 827-834.

51. Li D, He L (2007) Association study between the NMDA receptor 2B subunit gene (GRIN2B) and schizophrenia: a HuGE review and meta-analysis. Genet Med 9: 4-8.

52. Cull-Candy S, Brickley S, Farrant M (2001) NMDA receptor subunits: diversity, development and disease. Curr Opin Neurobiol 11: 327-335.

53. Funk AJ, Rumbaugh G, Harotunian V, McCullumsmith RE, Meador-Woodruff JH (2009) Decreased expression of NMDA receptor-associated proteins in frontal cortex of elderly patients with schizophrenia. Neuroreport 20: 1019-1022.

54. Kristiansen LV, Huerta I, Beneyto M, Meador-Woodruff JH (2007) NMDA receptors and schizophrenia. Curr Opin Pharmacol 7: 48-55.

55. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, et al. (2008) Natural selection on genes that underlie human disease susceptibility. Curr Biol 18: 883-889.

56. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614-1620.

57. Brachya G, Yanay C, Linial M (2006) Synaptic proteins as multi-sensor devices of neurotransmission. BMC Neurosci 7 Suppl 1: S4.

58. Collins MO, Yu L, Coba MP, Husi H, Campuzano I, et al. (2005) Proteomic analysis of in vivo phosphorylated synaptic proteins. J Biol Chem 280: 5972-5982.

59. Croning MD, Marshall MC, McLaren P, Armstrong JD, Grant SG (2009) G2Cdb: the Genes to Cognition database. Nucleic Acids Res 37: D846-851.

60. Takamori S, Holt M, Stenius K, Lemke EA, Gronborg M, et al. (2006) Molecular anatomy of a trafficking organelle. Cell 127: 831-846.

61. Trinidad JC, Specht CG, Thalhammer A, Schoepfer R, Burlingame AL (2006) Comprehensive identification of phosphorylation sites in postsynaptic density preparations. Mol Cell Proteomics 5: 914-922.

62. Zhang W, Zhang Y, Zheng H, Zhang C, Xiong W, et al. (2007) SynDB: a Synapse protein DataBase based on synapse ontology. Nucleic Acids Res 35: D737-741.

63. Laumonnier F, Cuthbert PC, Grant SG (2007) The role of neuronal complexes in human X-linked brain diseases. Am J Hum Genet 80: 205-220.

64. Grant SG, Marshall MC, Page KL, Cumiskey MA, Armstrong JD (2005) Synapse proteomics of multiprotein complexes: en route from genes to nervous system diseases. Hum Mol Genet 14 Spec No. 2: R225-234.

65. Gauthier J, Bonnel A, St-Onge J, Karemera L, Laurent S, et al. (2005) NLGN3/NLGN4 gene mutations are not responsible for autism in the Quebec population. Am J Med Genet B Neuropsychiatr Genet 132B: 74-75.

66. DeLisi LE, Shaw SH, Crow TJ, Shields G, Smith AB, et al. (2002) A genome-wide scan for linkage to chromosomal regions in 382 sibling pairs with schizophrenia or schizoaffective disorder. Am J Psychiatry 159: 803-812.

67. Joober R, Rouleau GA, Lal S, Dixon M, O'Driscoll G, et al. (2002) Neuropsychological impairments in neuroleptic-responder vs. -nonresponder schizophrenic patients and healthy volunteers. Schizophr Res 53: 229-238.

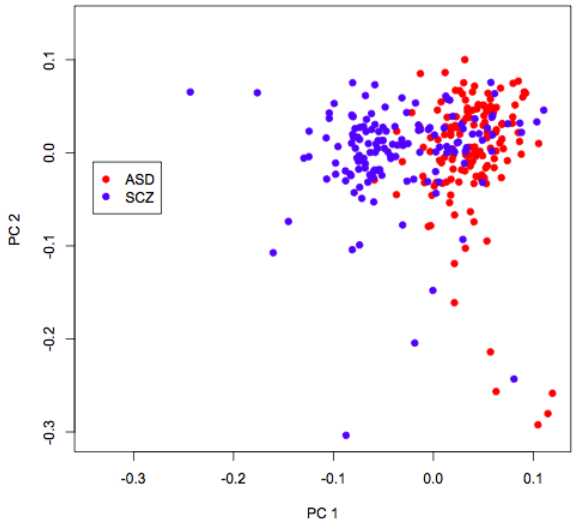
68. Gochman PA, Greenstein D, Sporn A, Gogtay N, Nicolson R, et al. (2004) Childhood onset schizophrenia: familial neurocognitive measures. Schizophr Res 71: 43-47.

69. Gourion D, Goldberger C, Bourdel MC, Bayle FJ, Millet B, et al. (2003) Neurological soft-signs and minor physical anomalies in schizophrenia: differential transmission within families. Schizophr Res 63: 181-187.

70. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

71. Luca D, Ringquist S, Klei L, Lee AB, Gieger C, et al. (2008) On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82: 453-463.

A)



B)

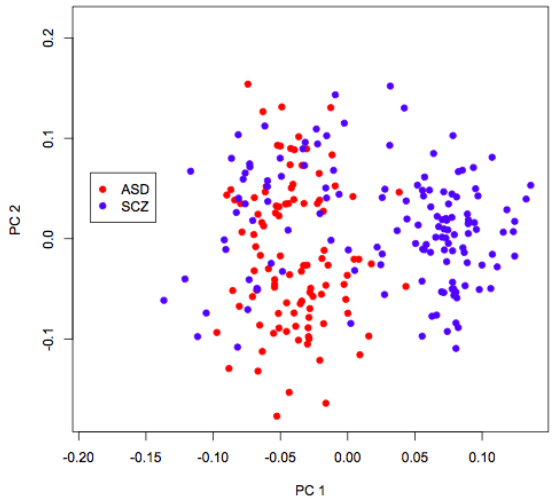


Figure 5-1. Principal Component Analysis (PCA) of Disease Cohorts.
Each dot denotes a sample's position in the top 2 principal components for A) all samples sequenced and B) samples after excluding ethnic and PCA outlier filters.

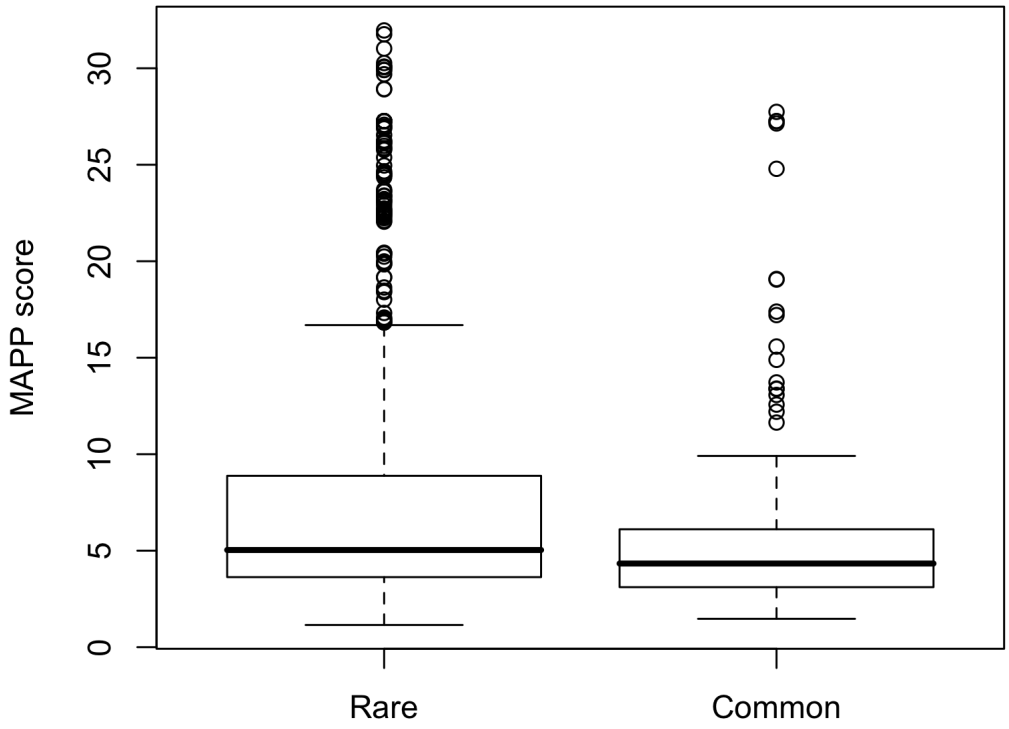


Figure 5-2. MAPP values in common versus rare variants.
Rare is defined as minor allele frequency < 0.03.

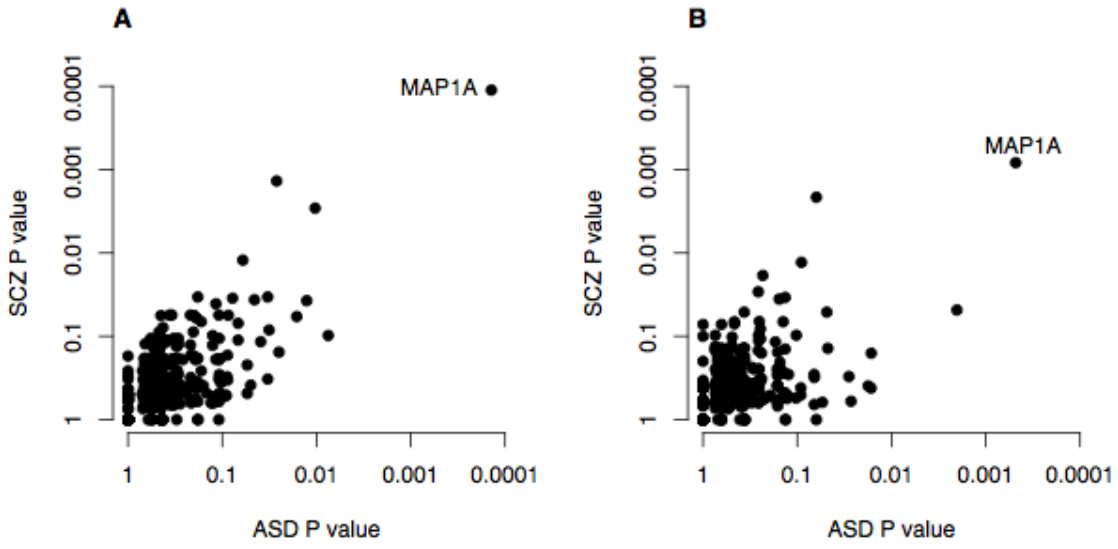
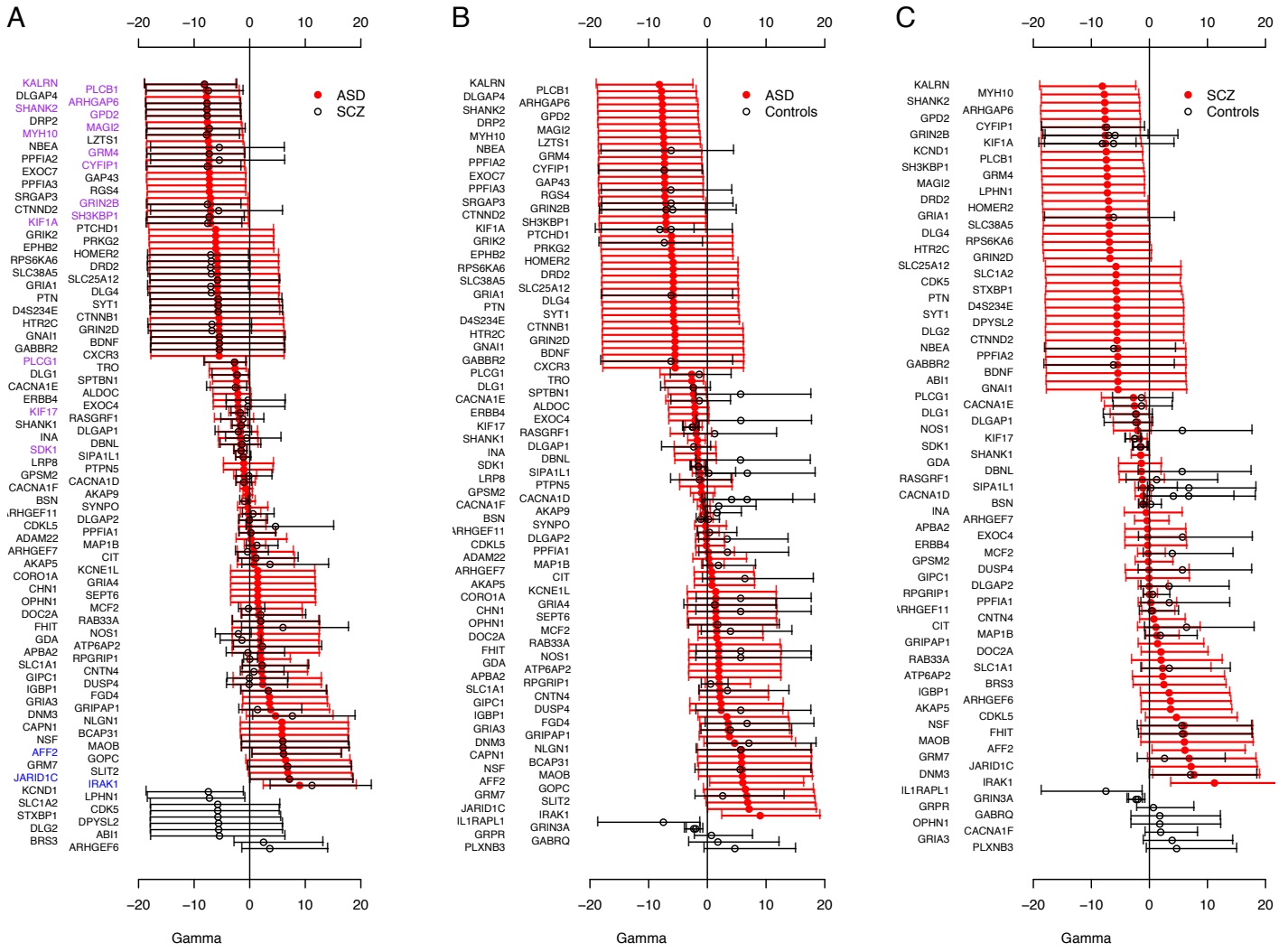


Figure 5-3. Excess of Missense Variants by Gene.
Fisher Exact Test *P* of the proportion of A) missense to silent variants and B) rare missense to all silent variants, at each individual locus compared to those proportions in the rest of the genes in the ASD and SCZ cohorts.

Figure 5-4. Distributions of Individual Locus Selection Coefficients
Posterior mean and 95% CI of γ values by gene in the A) ASD and Schizophrenia (SCZ) cohorts, gene names colored purple are negatively selected in both cohorts (γ 95% CI < 0) and gene names in blue are positively selected in both cohorts (γ 95% CI > 0) B) ASD and Controls (QTNS and Western European) and C) Schizophrenia and Controls (QTNS and Western European).


A
B
C
ASD
SCZ
Controls
Gamma
−20
−10
0
10
20

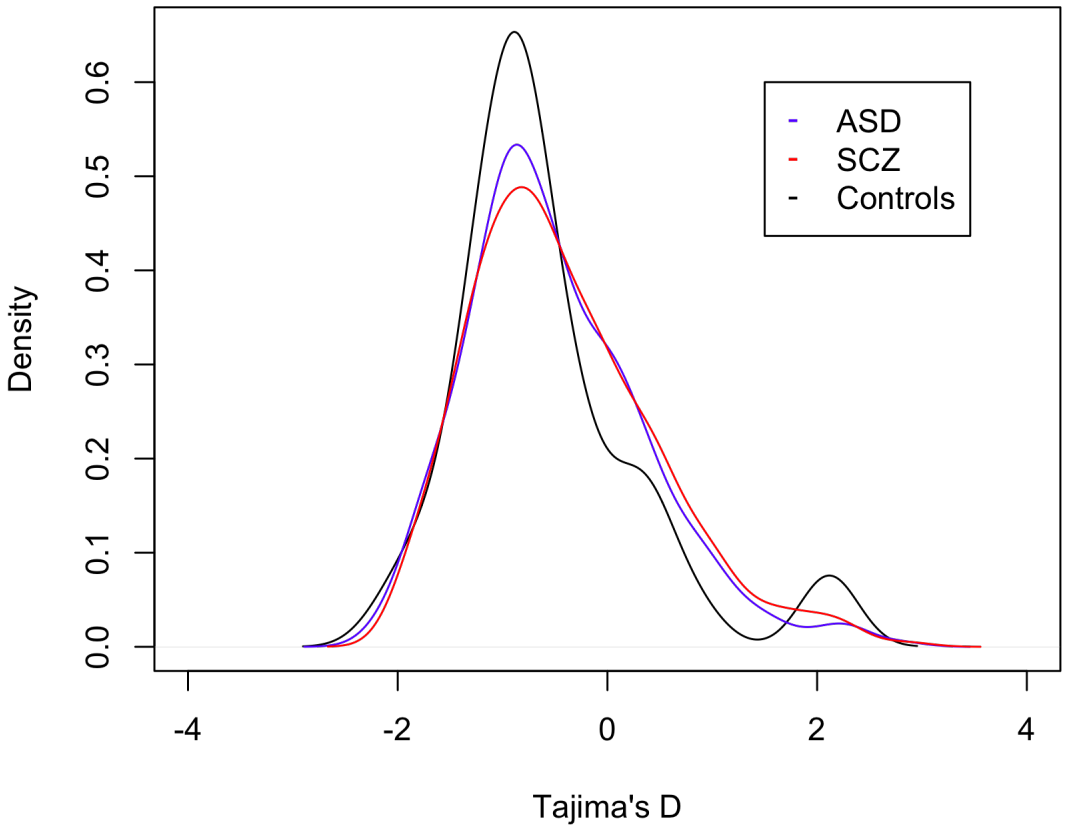


Figure 5-5. Distribution of per Gene Tajima's D.
The distribution of Tajima's D in all three cohorts, ASD (blue), SCZ (red), and Controls (black). The two-sided Kolmogorov-Smirnov test revealed the distribution of Tajima's D was not different between cohorts (ASD vs. Schizophrenia $P = 0.85$, ASD vs. Controls $P = 0.60$, and Schizophrenia vs. Controls $P = 0.29$).

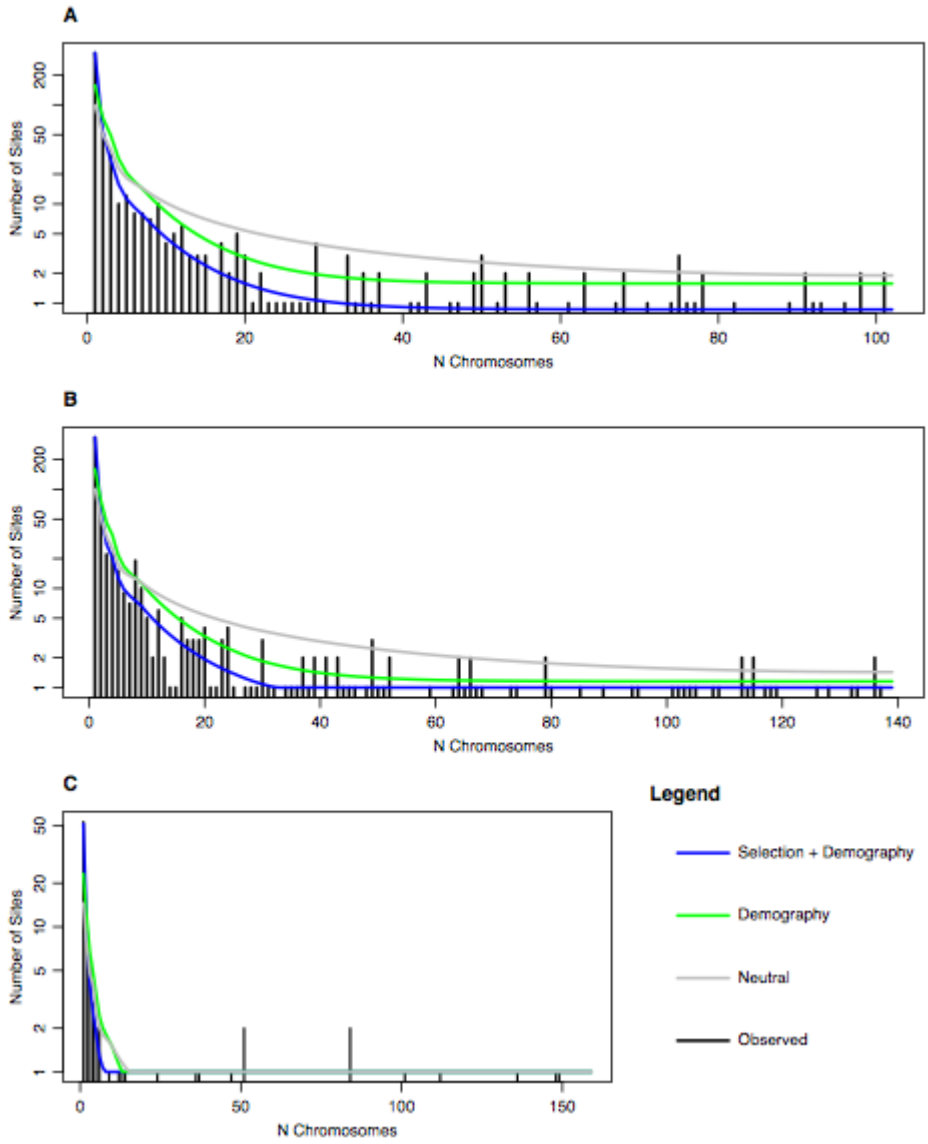


Figure 5-6. Observed and Expected Site Frequency Spectrum.
Observed and expected site frequency distributions for the SNPs the ASD (A), SCZ (B), and controls (C) cohorts. Expected distributions have been obtained under a neutral model not including demography, a model with demography, and a model with demography and selection.

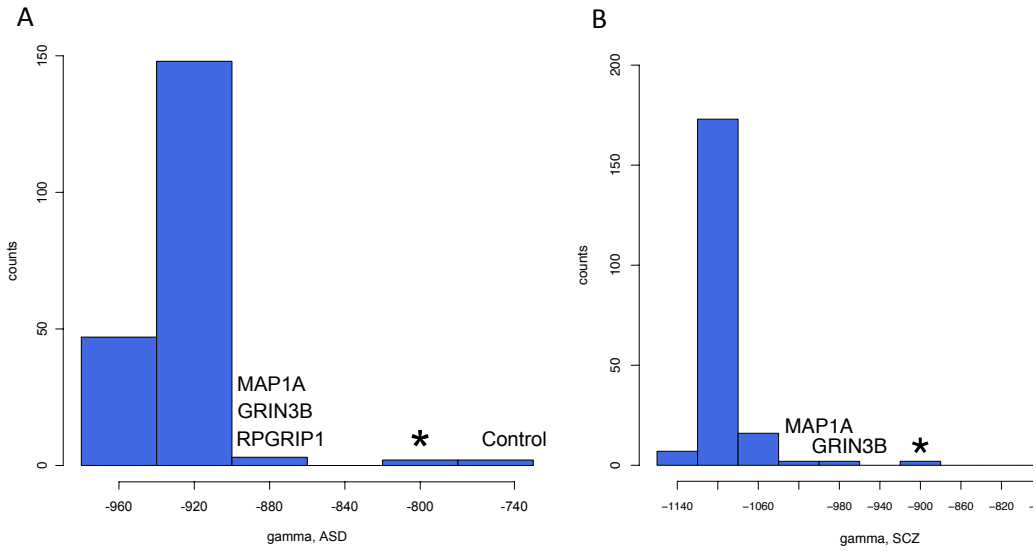


Figure 5-7. Distribution of Population Selection Coefficients Estimated Across All Loci.

Variation in the population selection parameter ($\gamma = 2N_e s$) obtained after excluding every gene with at least one missense variant (198) from the SFS for A) the ASD cohort; * denotes γ estimated after removing *MAP1A*, *GRIN3B*, and *RPGRIP1*, and B) the Schizophrenia (SCZ) cohort; * denotes γ estimated after removing *MAP1A* and *GRIN3B*. Genes producing the more important decrease in γ when excluded from the SFS are labeled. γ was estimated in the control cohort is also indicated for comparison.

Table 5-1. Segregating Sites and Diversity in the Disease Cohorts.
π is mean pair-wise nucleotide diversity across all samples; θ_W is Watterson's mutation rate estimator based on the number of segregating sites.
* Estimates were determined for each gene and then averaged across all genes.

Cohort		Missense	Nonsense	Silent	Intronic	UTR	Splice	Total	π* (per bp)	θ_W*(per bp)
ASD N = 102	Autosome	574	3	772	1,870	69	1	3,289	4.84×10^{-4}	3.58×10^{-4}
	X Chrom	124	1	153	508	40	0	826	3.35×10^{-4}	3.19×10^{-4}
	Total	698	4	925	2,378	109	1	4,115	4.39×10^{-4}	3.47×10^{-4}
SCZ N = 138	Autosome	614	6	818	1,722	70	0	3,230	4.45×10^{-4}	3.54×10^{-4}
	X Chrom	164	2	171	536	41	0	914	3.58×10^{-4}	3.10×10^{-4}
	Total	778	8	989	2,258	111	0	4,144	4.19×10^{-4}	3.41×10^{-4}
Total N = 240	Autosome	891	9	1,066	2,169	85	1	4,221	5.23×10^{-4}	3.48×10^{-4}
	X Chrom	220	2	228	675	50	0	1,175	3.96×10^{-4}	3.18×10^{-4}
	Total	1,111	11	1,294	2,844	135	1	5,396	4.84×10^{-4}	3.39×10^{-4}
QTNS N = 240	Autosome	108	0	103	217	5	1	434	4.72×10^{-4}	6.41×10^{-4}
	X Chrom	18	1	7	14	5	0	45	7.18×10^{-5}	1.08×10^{-4}
	Total	126	1	110	231	10	1	479	2.72×10^{-4}	3.75×10^{-4}

Table 5-2. Nonsense Mutations in Disease Cohorts

NA, not analyzed. [a] Human Mar. 2006 (hg18) assembly. * *de novo* mutations validated by resequencing parental tissue samples.

Chr	Position[a]	Gene	Counts ASD	Counts SCZ	Counts QTNS Controls
1	20,886,681	KIF17	0	1*	0
6	43,846,749	VEGFA	0	2	0
6	43,856,457	VEGFA	1	0	0
9	103,472,993	GRIN3A	0	1	0
14	71,241,195	SIPA1L1	0	1	NA
14	72,813,605	NUMB	0	1	NA
17	70,362,773	GRIN2C	1	0	0
19	956,224	GRIN3B	1	0	NA
22	49,506,476	SHANK3	0	1*	0
X	43,513,503	MAOB	0	1	NA
X	69,395,157	P2RY4	2	2	6

Table 5-3. Counts for Silent to Missense Ratios in MAP1A

	ASD		SCZ		Controls	
	Missense	Silent	Missense	Silent	Missense	Silent
MAP1A	20	5	25	8	NA	NA
Remaining Autosomal Genes	574	772	614	818	108	103

Table 5-4. Counts for Collapsing Method for Significant Results

	ASD		SCZ		Controls	
	# samples ≥ 1 rare missense	# samples 0 rare missense	# samples ≥ 1 rare missense	# samples 0 rare missense	# samples ≥ 1 rare missense	# samples 0 rare missense
GRIN2B	7	95	8	130	1	239
CACNA1F	9	93	5	133	3	237

Table 5-5. *Prfeq* Maximum Likelihood Estimates of Demographic and Selective Models.
* Autosomal mutations only. ** χ^2 (P value) with degrees of freedom being the difference of the number of estimated parameters. TAU is the time in generations since the non-stationary dynamics, scaled by 2*Ncurr. TAU B is the scaled time of the bottleneck. OMEGA is the ratio of ancestral to current Ne. OMEGA B is the ratio of bottleneck to current Ne.

Mutations*	Cohort	Model	maxLL	2 * lnL	Fixed parameters	Estimated parameters	P**
Silent And Intronic	ASD + SCZ	Stationary	13,629.28	-	-	-	-
Silent And Intronic	ASD + SCZ	Demography	13,771.17	238.78	-	OMEGA = 0.95, TAU = 0.1 OMEGA_B = 0.001 TAU_B = 0.001	<0.001
Missense	ASD	Demography	2,274.32	-	OMEGA = 0.95, TAU = 0.1 OMEGA_B = 0.001 TAU_B = 0.001	-	-
Missense	ASD	Demography + Selection	2,391.71	234.78	OMEGA = 0.95, TAU = 0.1 OMEGA_B = 0.001 TAU_B = 0.001	P = 0.04 γ = -920	<0.001
Missense	SCZ	Demography	2,368.64	-	OMEGA = 0.95, TAU = 0.1 OMEGA_B = 0.001 TAU_B = 0.001	-	-
Missense	SCZ	Demography + Selection	2,479.27	221.26	OMEGA = 0.95, TAU = 0.1 OMEGA_B = 0.001 TAU_B = 0.001	P = 0.05 γ = -1,100	<0.001
Missense	Controls	Demography	184.36	-	OMEGA = 0.95, TAU = 0.1 OMEGA_B = 0.001 TAU_B = 0.001	-	-
Missense	Controls	Demography + Selection	206.43	44.14	OMEGA = 0.95, TAU = 0.1 OMEGA_B = 0.001 TAU_B = 0.001	P = 0.06 γ = -740	<0.001

Table 5-6. Nominal *P* values of Identified Candidate Genes Exhibiting Excesses of Rare Variants

NA, not assessed. *Significant after Bonferroni Multiple testing. [a]Significant excess of missense and rare missense variants versus silent variants, compared to the rest of genes pooled together in each cohort and separately for X chromosome and autosomes. [b]Significant excess of individuals with rare missense variants in a given gene. QNTS controls used for the ASD cohort and schizophrenia negative controls used for the schizophrenia cohort. [c]Significant impact on the overall (autosomes) site frequency spectrum within each disease cohort. [d] Mean γ of disease cohort and control in parentheses. **denote mean γ significantly more negative in the disease cohort than among QTNS or Western-European controls. NA results from genes with no available divergence data.

A) ASD

	Missense vs Silent[a]		**Collapsing method[b]**	**prfreq[c]**	**mkprf[d]**
	All missense	**Rare missense**			
GRIN2B	0.057	0.139	0.0017*	0.96	-7.6 (-5.9)
GRIN3B	0.0029	0.018	NA	0.005	NA
CACNA1F	0.362	0.304	0.15	NA	NA
MAP1A	0.0001*	0.0008	NA	0.01	NA
NOS1	0.221	0.142	NA	0.96	-2.04** (5.68)
RPGRIP1	0.088	0.298	NA	0.91	0.015 (0.55)

B) Schizophrenia

	Missense vs Silent[a]		**Collapsing method[b]**	***prfreq*[c]**	***mkprf*[d]**
	All missense	**Rare missense**			
GRIN2B	0.016	0.047	0.001*	0.75	-7.311 (-5.9)
GRIN3B	0.01	0.23	NA	0.015	NA
CACNA1F	0.103	0.028	0.001*	NA	-0.691** (1.94)
MAP1A	0.0001*	0.0004	NA	0.015	NA
NOS1	0.429	0.536	NA	0.75	1.93 (5.68)
RPGRIP1	0.468	0.162	NA	0.015	2.02 (0.55)

Chapter 6 Conclusions

Motivated by the desire to better understand the genetic components of complex traits, we developed a combined approach to establishing genotype-phenotype associations that differs slightly from the classic approach in statistical genetics. Since natural selection acts on phenotype and alters the genomic landscape, we incorporated natural selection methods in parallel with standard association methods to better gauge the effectiveness of the respective tests in determining genetic contributions to complex traits. This hybrid approach is the logical progression for studying complex traits, since improving accuracy in detecting causal variants in diseases such as schizophrenia can lead to earlier diagnoses and personalized treatments.

When I began this research, microarray methods for genotyping were rapidly gaining popularity because they were far more affordable than sequencing. The widespread use of microarrays has lead to greater identification of genotype-phenotype associations at a genome-wide perspective, but limited tests of natural selection. These tests were limited because tests of natural selection generally rely on the distribution of allele frequencies and the ratios of functional to nonfunctional variants obtained from resequencing. Such tests could not be applied to genotyped data, because genotyped data is previously ascertained. As my research progressed, high-throughput and affordable sequencing technologies entered mainstream research resulting in genotyping being replaced by genome-wide resequencing. This progression enables joint testing of natural selection and genotype-phenotype associations, as well as increasing the types of complex trait models that can be tested. Chapters 2, 3, 4, and 5 reflect this shift from

genotyping to sequencing technology and the different complex trait models that can be tested with each of these technologies. I will conclude with key findings and challenges from the research chapters, and close with future directions for studying complex traits.

Summary of Main Results

In chapter two, we report the findings of a natural selection and drug resistance association study in a worldwide collection of *P. falciparum* isolates. During this study, we discovered that an interesting story develops when investigating genotype-phenotype associations and natural selection in the gene *pfcrt.* Chloroquine drug resistance associations and positive selection have been well characterized for this gene in previous studies, and our study also returned similar results. However, when *pfcrt* is compared to orthologous sequences from *P. reichenowi*, there is a large amount of variation maintained at this specific locus, indicating balancing selection. One explanation for this evidence of balancing selection is that positive selection acted on two different haplotypes from two geographically distinct populations. These haplotypes were targets of positive selection because they both confer resistance to chloroquine. As haplotypes undergo positive selection, variation in LD with the selection target will also increase in frequency. Since this process occurs independently in different populations, each population will have its own collection of variants. When viewed globally, the locus will exhibit an excess of variation. This explanation is valid if the evidence of balancing selection does not exist within each continent. However, even when independently testing

samples from Asia, South America and Africa, we found evidence of balancing selection. This evidence clearly indicates that prior to the widespread use of chloroquine, there existed a selection mechanism that favored variability in *pfcrt*, possibly the human host immune system. Nevertheless, our study clearly shows that incorporating natural selection methods into classic statistical genetics can provide a more holistic approach to the study of complex traits.

The experimental design utilized in chapter two is excellent for studying different types of natural selection but it is fundamentally limited in its smaller sample size and scope. The small sample size has limited power to find weak genetic determinants of drug response, particularly when samples come from multiple populations. Additionally, interrogating a single chromosome leaves the majority of the genome untested for signals of selection and association. These limitations are addressed in chapter three, where the sample size was increased, the scope was expanded to genome-wide analysis, and additional anti-malarial drugs were tested.

In chapter three, we report the findings of a genome-wide investigation of recombination, selective sweeps, and drug resistance associations in *P. falciparum*. A key result of the study was our discovery that there is inherent utility in studying positive selection in a collection of samples showing extensive population structure. The samples collected for this study showed a great amount of population structure, both globally and internally in the Southeast Asian population. Both of these population structures hinder interpretation of the genotype-phenotype association results since drug response also

varies by geography and is correlated with the respective population structures. We found that establishing controls for these population structures eliminated true association signals, while leaving these population structure unaccounted for resulted in false positive associations. Haplotype sharing-based tests, like iHS and XP-EHH, aided us in identifying selective sweeps and potential drug resistance associations. Since drug resistance is a strong selective force, it is reasonable to assume that loci associated with drug resistance should also reflect signals of positive selection. Positive selection signals can be separated from signals of population structure since positive selection signals will be localized to the genomic target of selection while population structure signals are genome-wide. Regions with extended haplotype sharing can be identified using iHS or XP-EHH, at which point variants in those regions can be tested for association with drug response. For example, *pfmdr1* showed evidence of positive selection in the Southeast Asian population and had significant associations with quinine response and nominally associated with chloroquine. The caveat to this approach, however, is that environmental pressures other than anti-malarial drug use may also drive selection. Also, evolution of anti-malarial resistance may not manifest as positive selection. These two factors clearly delineate that improvements in our experimental design, such as incorporating natural selection, are necessarily to expand our understanding of how complex traits evolve,

From the population genetics standpoint, this study is limited to tests of positive selection since a list of previously ascertained SNPs was interrogated. This is due to the limitation inherent within population genetics requiring all genetic variation data

originate from sequencing. From the statistical genetics standpoint, this study is an improvement over chapter two since the focus was expanded from a single chromosome to the entire genome. However, the study was still limited due to the extensive population structure in the data. Enhancements to this study should include resequencing rather than genotyping, improved sample collection with equal sampling from the distinct populations, and/or limiting the study to a single population. Chapters four and five learn from these lessons by utilizing resequencing rather than genotyping and also selecting samples from a more restricted ethnic background.

In chapter four, we present one of the largest human trio resequencing studies designed to detect *de novo* mutations, and describe their role in Autism Spectrum Disorder and schizophrenia. With this resequencing survey, we answer population genetic questions such as what are the neutral and functional human mutation rates, as well as medical genomics questions such as do disease affected cohorts show excess of deleterious *de novo* mutation. The neutral mutation rate estimate (1.36×10^{-8} per site per generation) is similar to approximations from phylogenetic estimates and also similar to estimates from other family based sequencing studies. While the functional mutation rate is nearly five times the neutral mutation rate, describing statistical excess of functional *de novo* mutations relative to neutral is more involved and subjective. Detecting an excess of functional *de novo* mutations is sensitive to such considerations as increased mutation rate in CpG sites relative to non-GpG sites, defining effective number of bases sequenced, and deciding which statistical test to use. The primary overarching factor

above all of the previously mentioned considerations is the source of DNA used for sequencing and *de novo* mutation detection. Nearly half of the *de novo* mutations could not be validated using DNA from original samples and were determined to be cell line artifacts. The prevalence of cell line artifacts is the single most important concern for most resequencing studies in which cell lines are the only available DNA sources. If resequencing is limited solely to cell lines, the number of *de novo* mutations will be inflated since the transformation of primary tissue samples to cell lines and the growth of cell lines causes mutations to accumulate. While sequencing the primary tissue sample is ideal, cell lines can be used for resequencing if the primary tissue sample is available for validation of the *de novo* mutations. Our study was careful to validate each and every *de novo* mutation using blood DNA, which allowed us to confidently bypass the above hurdles to determine functional mutation rates.

This study can be improved by including matched sequencing of unaffected samples. Sequencing control samples would provide a base line mutation rate for both functional and nonfunctional sites and allow direct comparisons of *de novo* mutation accumulation for different types of mutations. Sequencing of controls is often secondary in most studies, since most researchers find that discovering mutations in disease cohorts is ‘more interesting’. In the long run, a lack of emphasis on sequencing controls can hinder the study by limiting the researcher’s ability to make confident inferences about disease-associated *de novo* mutations.

A natural extension of the *de novo* hypothesis is that disease-affected cohorts will show accumulation of not just *de novo* mutations but also rare variants in key genes. This hypothesis is derived from population genetic theory, where deleterious selection on phenotype will keep disease-causing alleles at low frequency in the population. If a cohort is enriched for affected individuals, then key genomic regions involved in the disease will show excess of low frequency alleles. In chapter five, we present the methods and findings for such a study, using a combination of statistical and population genetic approaches to demonstrate accumulation of rare missense variants in the genes *GRIN2B*, *MAP1A,* and *CACNA1F*. An interesting result is that the GRIN gene family showed accumulation of rare variants, high impact missense variants, nonsense, coding indel, and *de novo* mutations. This accumulation of detrimental variants suggests gene families and perhaps gene pathways are the genomic unit of interest, rather than single genes.

Future Directions

As we gain a better understanding of the effect of natural selection on complex traits, we can incorporate selection information into association studies in a much more effective manner. For example, one such method is to jointly infer positive selection and associations when the trait of interest confers a selective advantage. Another example is refining the CDRV model to incorporate both the accumulation of deleterious variants with individual samples accumulating rare variants. Chapter three shows the promise of

using haplotype-sharing tests when studying a complex trait that confers a strong selective advantage. Haplotype sharing tests may also have a role to play when studying any complex trait in which selection alters the local LD structure. In contrast to the positive selection examples, genomic regions subject to deleterious selection will show decreased haplotype sharing in populations enriched for affected samples. Furthermore, contrasting haplotype sharing between cases and controls has the potential to identify genomic regions subject to different selective pressures.

Another direction I can see for this field of study in the future is to use lessons from the biological side of complex traits to improve our understanding of genotype-phenotype relationships. Biologists approach the complex disease hypotheses using a Bayesian approach. For example, genes known to be involved in brain function were resequenced in chapters four and five. This information could be incorporated as a "genome prior distribution," allowing certain regions of the genome to be more likely associated with complex disease than other regions. Other sources of prior information may include conservation, functional studies from model organisms, predicted effects of mutations, and the number of potentially deleterious mutations in a genome. Many of these sources of information are used after the fact, or as a way to rank association results. By using this prior information more proactively as a factor in statistical design, genotype-phenotype association studies can increase in power and decrease in error. As this genome-wide prior is better defined, disease risk or even status may become predictable given a person's genomic sequence.

Given the availability of affordable resequencing technologies and a greater awareness of population genetics theory, the future holds great promise for improved understanding of the genetic mechanisms of complex diseases and traits. My hope is that one day, the genetic diseases we currently consider debilitating will become better understood, better diagnosed, better treated and managed, and have only minor impacts on one's quality of life.