Proxy Genotypes and Phenotypes for Human Genetics

by
Roman Yelensky

M.S. Computer Science, Stanford University, 2003
B.A. Computer Science, Cornell University, 2000

SUBMITTED TO THE DIVISION OF HEALTH SCIENCES & TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN
BIOINFORMATICS AND INTEGRATIVE GENOMICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2008

Copyright 2008 Massachusetts Institute of Technology. All rights reserved.

Signature of Author:________ ________
Harvard-MIT Division of Health Sciences & Technology
August 4th, 2008

Certified by (Thesis Supervisor):__ ________
David Altshuler MD, PhD
Professor of Genetics and Medicine
Harvard Medical School, Massachusetts General Hospital

Certified by (Co-Advisor):______ ________
Mark Daly, PhD
Assistant Professor of Medicine
Harvard Medical School, Massachusetts General Hospital

Accepted by:________ ________
Martha Gray, PhD
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Sciences and Technology
Massachusetts Institute of Technology

Proxy Genotypes and Phenotypes for Human Genetics

by

Roman Yelensky

Submitted to the Division of Health Sciences & Technology on August 8th, 2008 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Bioinformatics and Integrative Genomics

Abstract

Genetic mapping by association is an unbiased approach to discover genes and pathways influencing disease traits and response to drugs and environmental exposures. There are two key obstacles to mapping in humans: (1) The full sequence of study subjects cannot yet be obtained; and (2) There are substantial limits to the phenotypes that can be safely elicited or measured. Geneticists thus rely on practically measurable sets of genotypes to proxy for the sequence and human *in-vitro* models that proxy for *in-vivo* genetics and physiology while allowing for perturbation and characterization in high throughput. This thesis presents the development of one important class of proxy genotypes, those that capture most common genetic variation, as well as an evaluation and refinement of proxy phenotypes offered by one commonly used *in-vitro* model, the lymphoblastoid cell-line.

Capturing common human genetic variation for genome-wide association studies requires genotyping a feasible subset of proxy (or "tag") SNPs. We investigated selection and analysis of tag SNPs, examined the relationship between investment in genotyping and statistical power, and evaluated whether power is compromised when tags are selected from an incomplete resource such as HapMap. We demonstrate an efficient haplotype-based tagging approach and other methods that dramatically increase tagging efficiency. Examining all observed haplotypes for association increases power to detect rare causal alleles, while reducing power for common alleles. Power is robust to completeness of the reference panel and holds across demographically related groups.

Lymphoblastoid cell lines (LCLs) are being developed into an *in-vitro* model where genetics of human gene expression, drug response, and other traits can be studied under controlled conditions. However, the impact of the immortalization process, the relative influence of non-genetic factors, and reproducibility of measured traits are not yet understood. We addressed these questions while mapping loci for response to chemotherapy and found that traits in LCLs are subject to substantial confounders and are only modestly reproducible in independent experiments. Despite this, RNA expression of many genes is affected by genetic variation and predicts response to drugs; integrating SNPs, RNA, and drug response can identify novel pharmacogenetic variation mediated by RNA.

Thesis Supervisor: David Altshuler MD, PhD
Title: Professor of Genetics and Medicine, Harvard Medical School
Co-Advisor: Mark Daly, PhD
Title: Assistant Professor of Medicine, Harvard Medical School

Acknowledgments

I would like to first acknowledge my key collaborators on the research described in this thesis: Paul de Bakker and Itsik Pe'er for the Proxy Genotypes section and Edwin Choy and Sasha Bonakdar for the Proxy Phenotypes. You were not only colleagues, but also friends, and working with you made the last few years productive and enjoyable.

I am forever indebted to my mentors, David Altshuler and Mark Daly. I have learned to identify the important questions, pursue answers with rigor, and to effectively communicate my results. These skills will serve me every day of my career.

None of this would be possible without the unwavering love and support of my family. Thanks to my wife Abby for making such a wonderful home for us and putting up with me through the ups and downs of research. Thanks to my parents Vladimir and Irina Yelensky and my grandparents Semyon and Sofia Grinberg for always believing in me and helping in every possible way.

Finally, thanks to my department, HST, and to my committee chairman, Zak Kohane, who got me to come here in the first place. HST is a unique place, where engineers get to understand and solve the most important medical problems, and I could not imagine earning my PhD anywhere else.

This thesis is dedicated to my grandfather:
Semyon Grinberg, Apr 12th, 1933 - March 29th, 2008.
We love you and we miss you.

Table of Contents

Abstract 3
Acknowledgments 4
Table of Contents 5
Chapter 1: Introduction 7
  Proxy Genotypes: Capturing common genetic variation 11
  Proxy Phenotypes: Genetic analysis of human traits in-vitro 17
  References 22
Chapter 2: Efficiency and power in genetic association studies 27
  Introduction 27
  Results 29
    Disease association studies based on empirical genotype data 29
    Capturing all sites observed in a complete reference panel 30
    Increasing efficiency by relaxing thresholds for tag SNP selection 32
    Tags selected from an incomplete reference panel 33
    The impact of LD on tag SNP selection and power 34
    Exhaustive haplotype tests to detect less common alleles 36
    Software 37
  Discussion 37
  Methods and Materials 40
    Data sets 40
    Genetic model and simulation of case-control panels 40
    Reference panels for tag SNP selection 40
    Selection of tag SNPs and definition of tests 40
    Region-wide test statistic and power calculations 41
    Exhaustive haplotype tests 41
  Figures 42
  References 47
Chapter 3: Transferability of tag SNPs in genetic association studies in multiple populations 49
  Introduction 49
  Results 49
  Discussion 54
  Methods and Materials 55
    DNA samples 55
    SNP discovery 55
    SNP genotyping 55
    Simulation of case-control association studies 56
    Tag SNP selection 56
    Power calculations 56
  Figures and Tables 57
  References 62
Chapter 4: Evaluating and Improving Power in Whole Genome Association Studies using Fixed Marker Sets 65
  Introduction 65
  Results 65

Discussion........70
Methods and Materials........71
Data sets........71
Choosing multi-marker predictors........71
Simulating case-control panels........71
Power calculations........71
Derivation of weights for allelic tests........72
Figures........74
References........76
Chapter 5: Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants........79
Introduction........79
Results........80
Discussion........86
Figures........87
References........89
Chapter 6: Genetic Analysis of Human Traits *In-Vitro*: Drug Response and Gene Expression in Lymphoblastoid Cell Lines........91
Introduction........91
Results........93
Data Collected........93
Cell line sensitivity to chemotherapeutic drugs........94
Variability in RNA expression........97
Inter- and Intra- individual variance component analysis........101
Integrating data from eQTLs and drug response in LCLs........102
Discussion........105
Methods and Materials........108
Cell Culture........108
Drug Response Assay........108
Modeling Drug response........110
Growth Rate Measurements........111
FACS Analysis........111
Luminex Assay........112
RNA preparation and Affymetrix expression profiling........113
Relative EBV and mtDNA Copy Number........113
Fraction of RNA variance explained by cellular phenotype or eQTL (Figure 4)........114
Decomposing gene expression into inter- and intra- components (Figure 5)........115
GWAS for drug response........115
Figures and Tables........117
References........128
Chapter 7: Summary and Discussion........131
References........135
Appendix........136

Chapter 1: Introduction

Much of the progress of biomedical science over the last several hundred years can be viewed as the gradual refinement of gross phenotypic observations of anatomy and disease into a detailed mechanistic description of organ, tissue, and cellular functions, inter-relationships, and pathophysiology. The tools and methods of molecular biology developed over the last few decades now open the "final frontier" of this advance and make possible a detailed understanding of how bio-molecules interact to produce behaviors at every level of aggregation above. When this frontier is sufficiently explored, an integrative view of human biology can emerge that will revolutionize pharmaceutical development and the practice of medicine. The scientific questions framing and motivating this research often focus on the study of differences: Why do some develop heart disease and others do not? What changes a previously healthy tissue to diseased? Why are some fatally susceptible to infection, while others are immune? Uncovering the molecular basis of this phenotypic variation is thus one of the great undertakings of biomedical research today.

A major component of this enterprise is the investigation of *inherited* genotype-phenotype relationships, or genetics. The existence of tremendous natural, heritable variation in most traits of interest, coupled with the current detailed understanding of the molecular mechanisms of inheritance makes fertile ground for scientists interested in how molecular changes lead to differences between whole organisms and what happens at every level in-between. For example, the most practical first step in understanding unknown biology of a disease is often the genetic mapping of susceptibility or risk, followed by identification of the mutations responsible and characterization of the role implicated genes play in pathophysiology. Genetic discoveries can often lead to disease prevention (i.e. Tay-Sachs), more effective treatments (cystic fibrosis [1]), and may even hold promise for cures. A complementary application of human genetic research is the study of natural variation in drug toxicity and response. Identification of genetic variants resulting in impaired drug metabolism have lead to screening tests that prevent substantial morbidity/mortality associated with standard treatments for a variety of conditions (i.e. 6-mercaptopurine for leukemia and inflammatory bowel disease [2], abacavir for HIV [3]). Further research will help explain differences in efficacy [4] encountered with commonly used drugs and suggest new pathways and targets for pharmacologic development.

For several decades, human genetic research was carried out using the linkage approach (pioneered by Morgan and Sturtevant for drosophila [5]) with sparse collections of markers throughout the genome for overt clinical phenotypes demonstrated by severe mendelian disorders [6]. While the approach clearly made tremendous contributions in uncovering the genetic basis of rare conditions such as Huntington's disease, it generally fell short when applied to common, complex, though still heritable, conditions such as obesity, diabetes, or cardiovascular disease. [7-9] Among the many reasons linkage was less effective was its relative lack of statistical power to detect common variants of modest effect [10], locus heterogeneity, and the inherent difficulties in fine-mapping causal genes from wide linkage peaks. Most importantly, the key assumption on which the success of linkage rests, that disease is caused by high penetrance alleles in few genes, appeared to not hold for most traits. On the phenotype side, investigators have traditionally been limited to low-throughput, limited scope physical exam, history, clinical chemistry and pathology, making determination of mechanism difficult and time consuming (or impossible) even if the causal variant could be found.

To clarify next steps necessary to overcome these obstacles and realize the full promise of human genetic research for all heritable traits, it is helpful to imagine what the "best case" scenario for a human geneticist would look like. Aside from the usual desire of infinite sample size, clearly we would like to have the full sequence of every individual under study, preferably combined with a genome-wide understanding of epigenetic and other relevant genome variation. On the phenotype side, we would like to know the RNA expression, post-translational modifications, and intra-cellular localization of every gene in every cell in the body, the expression and activation state of every cell-surface receptor, as well as complete proteomic and metabolomic measurements in all extra-cellular compartments. Finally, to understand variation in drug toxicity and response, we would like the ability to harmlessly dose any study subject with any drug or toxin and measure relevant changes in physiologic state.

While some of these requirements may conceivably be within reach (like full sequence), others are still quite far off (such as detailed proteomic measurements of remote compartments), and the ability to safely give any person any drug for the sake of genetic research will likely never be attained. Nevertheless, how close we can get to these goals will define how much we can learn from human genetics (and, indeed, from biomedical research in general!) This thesis presents the development of analytical and experimental tools and resources that moves us towards this "best case", while expanding the horizons of discoveries possible right now.

Recent advances in sequencing and genotyping technology, as well as significant investment, have given us an unprecedented understanding of the human genome and genetic variation [11], and with it the capability to search for the genetic basis of complex traits using the association approach. [12] At the same time, -omic tools such as RNA micro-arrays have broadened and deepened our conception of "phenotype." [13] Histology based cancer staging being enriched by detailed expression profiles [14], and classical definitions of diabetes being supplemented by detailed views of the dysregulated pathways [15], are just two of a myriad examples. This confluence of genotypic and phenotypic advances opens exciting avenues for progress [16] in understanding the links between genes and traits and every step in-between; however, substantial practical obstacles and open methodological questions in the application of these tools remain.

On the genotype side, the 3GB human sequence still far dwarves our ability to measure it routinely in more than a handful of individuals; full re-sequencing genetic mapping experiments are still likely several years away. Even the millions of markers (SNPs) [17] of human variation that have recently become available are beyond present-day financial and technological means. An even more important challenge is that unlike model organisms that can be exhaustively phenotyped and readily exposed to drugs and toxins in the laboratory, there are substantial limits to the phenotypes that can be safely elicited or measured in human subjects. Geneticists thus rely on practically measurable sets of genotypes to proxy (as well as possible) for the full sequence and human *in-vitro* models that can proxy for *in-vivo* genetics and physiology while allowing for systematic perturbation and characterization in high throughput.

This thesis presents the development of one important class of proxy genotypes, those that capture most common genetic variation, as well as an evaluation, refinement, and application of proxy phenotypes offered by one commonly used *in-vitro* model, the lymphoblastoid cell-line. The section titled "Proxy Genotypes: Capturing common genetic variation" below offers a detailed introduction and historical background to our treatment of the genotype challenge, contained in Chapters 2-5. The following section, titled "Proxy Phenotypes: Genetic analysis of human traits *in-vitro*", likewise reviews the context for the phenotype challenge, addressed in full in Chapter 6. We then recap with a roadmap for the reader to follow, before diving in and exploring the material in depth.

Proxy Genotypes: Capturing common genetic variation

In an influential perspective published in Science in 1996 [10] and soon echoed by other [18], Risch and Merikangas argued that the path forward for uncovering the genetic basis of common, complex human diseases lay with the genome-wide association, and not the linkage, approach. Linkage uses sparse marker genotypes throughout the genome to identify shared chromosome segments in closely related affected individuals and then aggregates this information over many families to detect increased sharing in particular (hopefully small) regions of the genome. This region is then hypothesized to harbor a causal mutation for the disease. Association, in the case-control formulation, counts alleles at a potentially causal polymorphic site in unrelated cases (affected) and controls (unaffected) and declares association if there is statistically significant enrichment in either set. [12] In it's transmission disequilibrium form, association counts transmissions of an allele from heterozygous parents to affected offspring and declares success if there is significant over- (or under-) transmission of the allele. The authors demonstrated that association (because it tested the polymorphism directly) was a more statistically powerful approach and, in particular, that association can detect variants of modest effect suspected of having an important role in common disease at attainable sample sizes, whereas linkage could not.

A crucial assumption and key criticism of the above analysis was that the polymorphism evaluated for association was the causal site itself. It was pointed out immediately [19] that power would fall dramatically if the polymorphism evaluated for association with the disease trait was not in strong linkage disequilibrium (non-random association of alleles in unrelated individuals) with, or substantially differed in frequency from, the causal mutation. It was generally understood that genome-wide catalogues of variation available at the time (libraries of 100s or 1000s RFLPs or microsatellites), while adequate for capturing recombination events in families for linkage scans, would not be linked tightly enough (or at all) to most possible causal mutations for association to work. Thus, calls were issued for the creation of much larger catalogues of human genetic variation that may contain many putative causal alleles directly (i.e. coding changes), and be potentially dense enough to detect association indirectly through linkage disequilibrium (LD) with most others. [20]

These calls were answered by several efforts, most notably by groups participating in the ongoing Human Genome Project and the newly formed SNP Consortium. [17,21] The genome project confirmed previous estimates [22] that humans were identical at >99.9% of their sequence and that by far the most abundant class of variation in the human genome were single nucleotide polymorphisms (or SNPs); any two genomes differed at ~8 SNPs for every 10 kilobases of sequence. The first genome-wide map of 1.4M SNPs, derived from overlapping reads from large-inset clones and shotgun sequences from a panel of individuals was published along-side the consensus genome in 2001 and was hailed by human geneticists as an equally (if not more) important accomplishment [23-25]. Despite this impressive advance in increasing the density of genome-wide catalogues of variation by 1000-fold and making >60,000 coding SNPs available for candidate gene association studies, a crucial questions for medical genetics remained: Can putative disease causing variants not on the map (still an overwhelming majority) now be detected indirectly through LD with variants on the map? Practically, even cutting-edge genotyping technology circa 2001 could only hope to measure on the order of 1000s variants in the sample sizes required for association scans, so the relationships between polymorphic sites (LD) would need to be understood (and technology improved) before well-powered unbiased genome-wide association studies could be designed and carried out.

While there were many linkage maps that detailed where recombination was expected to occur during meioses in families, the history of mutation, recombination, bottlenecks, and drift of all chromosomes present currently in the population (the determinants of linkage-disequilibrium) was not known genome-wide. If this history was such that LD extended for only short distances in the genome (as simulations predicted [26]), and if most linked polymorphisms were not well frequency matched, then the map would need to get much denser and more putative causal sites would need to be discovered and tested for association to disease directly. If, on the other hand, LD was long and (relatively) distant sites carried correlated alleles, indirect association methods could be designed to find mutations responsible for common, complex disease.

The effort to understand LD in the human genome took decades and proceeded in steps, from original studies looking at single exons and genes [27], to bigger regions, to multiple regions [28], to eventually a genome-wide view. One of the earlier region-wide SNP-based studies, published in 2001 [29], described LD across ~100 SNPs in ~250 chromosomes in a 500kb region on 5q31; it was noted that variation in the region existed in 11 10-100kb sized "blocks" of limited haplotype diversity, punctuated by sites of historical recombination. Most chromosomes in the population indeed appeared to be mosaics of 2-4 common haplotypes in each block. This study was soon extended, in 2002 [30], to a survey of 54 autosomal regions, spanning 13Mb, in 400 total independent chromosomes from several population groups. The block-like nature of variation was confirmed, with over half the genome surveyed appearing in blocks with limited evidence for historical recombination of length >40kb and >20kb in out-of-Africa and African samples respectively. Again, limited haplotype diversity was observed, with most chromosomes consisting of 3-5 common haplotypes per block. Just as importantly, a haplotype framework defined on only a subset of SNPs correlated well with the rest of variation in the region. Thus, excitement began to build in the field that at least all common genetic variation was within reach and the International Haplotype Map project was launched to catalogue common SNPs genome-wide.

The HapMap project consisted primarily of two goals: (1) To obtain an evenly spaced, densest map possible of SNP with MAF>5% across the human genome and (2) To obtain all (or nearly all) SNPs with MAF>5% in 10 unlinked 500kb (ENCODE) regions. Crucially, the project would provide genotypes for these variants in 100s of individuals so that LD could be assessed. Together, the two components would allow geneticists to develop tools for the selection and analysis of polymorphisms to test the common variant contribution to disease and apply these tools genome-wide. Phase 1 of the project, containing genotypes for 1M SNPs – 1 per 5kb – with MAF>5% in 270 people, as well as 18K SNPs – 1 per ~300bp – in the 5MB ENCODE subset, was published in 2005 [11] and was hailed by some as an "unprecendented gift" for the genetics community. [31]

Indeed, the data offered great promise for those planning candidate-gene and whole-genome association scans, as well as commercial vendors designing technology to support these activities. Analysis revealed extensive allelic correlation between common-variant polymorphic sites, especially in out-of-Africa populations. For instance, in the CEU panel, nearly 90% of all common SNPs in the fully ascertained ENCODE regions were strongly correlated (r^2>0.8) to at least one other common SNP, while almost 60% were strongly correlated to (also termed "proxied for") at least 10 others. Even in the 10-times sparser Phase 1 map, 75% of SNPs had at least one proxy at a level of r^2>0.8. On the whole, it became apparent that most common variation in the population could likely be captured (or "tagged") by genotyping only a fraction of all sites and appropriately designed methods to select and test these tags for association to disease may yield substantial statistical power at feasible technological and economic cost.

Therefore, contemporaneously with the development of the HapMap/ENCODE resource, attention also focused on the development and evaluation of methods to facilitate its use in association scans. The fundamental challenge facing investigators attempting GWAS is how to balance statistical power to detect associations with financial/technological "efficiency" of the study; i.e. how can we get the best bang for our buck? At the time of HapMap Phase 1 release, it was not possible to genotype all 1M common variants known in high throughput; even now its not feasible to genotype all >3M common variants eventually made available by HapMap Phase 2. More importantly, the majority of variants with MAF>5% are not even known, making it imperative that tagging/testing methods maximize power to detect *all* potential causal sites.

A variety of tagging/testing were proposed in support of previous smaller scale candidate gene association studies and in anticipation of the HapMap resource. These differed primarily in the treatment of haplotypes: Some argued that tag SNPs be selected to capture haplotypic diversity and then tested against traits as proxies for the haplotype. [32,33] Others thought that haplotypes be ignored entirely and tag SNP selection simply focus on capturing as many single snps as possible. [34,35] A compromise approach was then advanced where some number of snps was genotyped and tested in a multiple regression framework. [36] One suggestion was even made to test all possible haplotypes exhaustively, in the hope that undiscovered SNPs could come to the fore. [37] A review and comparison of the main methods proposed can be found in [38].

The availability of the HapMap/ENCODE resource finally allowed a comprehensive empirical evaluation of key ideas and underlying principles in the selection and analysis of tagSNPs: Do pair-wise or multi-marker methods maximize efficiency and power? Should all observed haplotypes be examined for association, rather than those that are proxies for known SNPs? What is the general trade-off between investment in genotyping and power to detect effects? Finally, to what extent is power compromised when tags are selected from an incomplete resource such as HapMap? The answers to these questions are important at every stage of study design and analysis and are critical determinants of success.

We examined the questions above using genotype data from the HapMap ENCODE project, association studies simulated under a realistic genotype-phenotype causative model, and empirical correction for multiple hypothesis testing; the results are presented in Chapter 2 [39]. Our study demonstrated that whole genome association was practical with straight-forward methods using the HapMap resource and can reasonably hope to capture most of the genetic information about common variants in a sample of individuals. An important criticism of the work was our reliance on a single sample (the HapMap individuals) to both select tag-SNPs and simulate causal variants. Practically, tags are selected in some reference panel of individuals (such as the HapMap), but then genotyped/tested in a new disease study cohort. We addressed this issue in Chapter 3 by repeating the simulation study across multiple samples and found that our conclusions did not change under this more realistic scenario [40]. As many investigators will be using commercially available genotyping technology, we extended aspects of the analysis to fixed-marker platforms such as those manufactured by Affymetrix or Illumina, and found they too capture a significant fraction of common genetic variation; these results are presented in Chapter 4 [41]. Finally, in Chapter 5 [42], we apply the simulation framework used in previous chapters to derive an empirical genome-wide significance threshold for genome-wide scans for investigators to apply when evaluating their study results.

Taken together, the results in Chapters 2-5 show that a tiny fraction (1/100th of 1%) of the genome can represent (or proxy for) an entire class of genetic variation (common variants with MAF>5%) and provide excellent relative power to detect association with disease. Available genome-wide catalogues of variation (i.e. the

HapMap) and simple analytical approaches are thus sufficient to achieve favorable tradeoffs between power and efficiency in the design and execution of genome-wide association scans. This optimistic view of the potential success of GWAS appears to have borne out, with new genotype-phenotype association findings now being published nearly every week. [43]

Proxy Phenotypes: Genetic analysis of human traits in-vitro

Identification of a genetic variant that confers risk to disease or influences another whole-organism phenotype is, of course, only the beginning of a long process to elucidate its function, place it in context of other genetic and non-genetic variation, and understanding the relevant (patho-) physiology. Unless the variant is an obvious coding mutation in an exon of a well-studied gene, pointing the way to the mechanism by which it causes disease, it may not be clear how to proceed from a "hit" in a genome-wide scan. Often, even after extensive fine-mapping, the LD signal cannot be precisely localized to a gene at all and no "causal" variant is evident. [44] This, in fact, is the result of a GWAS we might expect for a complex trait influenced by non-coding regulatory variation. More work must now be done to generate further leads about which genetic polymorphisms are responsible and exactly how they act.

There are many potential approaches to this challenge, each more or less applicable depending on the specifics of the GWAS result to follow up. In the minority of cases where a promising causal variant is suggested by the genetics, transgenic animals can be made to study genotype-phenotype relationships in depth, with the expectation (and hope) that findings will carry over to human beings. [45] If no coding or simple regulatory (i.e. splice site or poly-A tail) mutation is obvious, but the association signal can be narrowed down to a relatively small region, then other tools such as databases of transcription factor binding sites [46] may be brought to bear to identify potentially functional polymorphic sites. In the general case, however, when the exact polymorphism or even its precise location is not clear after a GWAS, we will need more information about study subjects than just the sequence and disease status to make progress.

Traditional clinical traits (sometimes also referred to as "endo-phenotypes" in the genetics literature) can certainly make important contributions to elucidating mechanisms of disease-risk conferring genetic variation. As one example, recent studies have shown that SNPs at nine loci raised LDL cholesterol, lowered HDL cholesterol and conferred greater propensity to cardiovascular events, implicating previously known pathways of pathogenesis. [47] Similarly, association with disease and an important endo-phenotype can offer a new perspective on the disease state and associated complications as did variation in the glucokinase receptor, which turned out to both confer risk to developing diabetes and affected levels of serum triglycerides. [48] Although many clinical endo-phenotypes are easily measured and widely available, they suffer from important limitations: They are likely far removed, both spatially and temporally, from the causal pathophysiological process. For instance, while reality may be that a regulatory variant in hepatocytes alters signaling cascades and induces a myriad important changes in the liver that eventually lead to increased serum LDL and atherosclerosis, measuring only LDL and disease risk does not get us close enough to untangling what really happened. Clearly, much more detailed phenotypic measurements will be necessary to leverage genetic discoveries in the quest to understand pathogenesis and devise targeted interventions.

Indeed, from the point of view of understanding the impact of genetics on disease intervention (i.e. drug development), the challenges are even greater. There is great inter-individual variability in response to drugs[49,50], some of it undoubtedly genetic, but only a few of the causal variants are currently known and even fewer are in active clinical use. [51] At the same time, genetic variation can have a profound impact on efficacy and safety, for instance genotyping variants in the TMPT enzyme can be (cost-effectively) used to adjust 6-mercaptopurine dose in the treatment of leukemia and avoid potentially fatal myelosuppression events. [2] However, the discovery and characterization of these variants is impeded by significant ethical and practical concerns: *In-vivo* response can only be studied in patients that were prescribed a drug for a necessary indication, and these patients may vary in many parameters (i.e. dose, diet, age, other conditions, other drugs) that we cannot control, making sufficiently sized well-matched cohorts difficult or impossible to assemble and maintain. Moreover, as above, we may not be able to measure important sub-phenotypes of response, such as expression of P450 enzymes in the liver, that may be necessary to understand the mechanisms of the variant even if it is found.

In sum, human geneticists are faced with the challenge that unlike model organisms that can be exhaustively phenotyped and readily exposed to drugs and toxins in the laboratory, there are substantial limits to the phenotypes that can be safely elicited or measured in human subjects. Thus, there would be great value in a human *in-vitro* model that faithfully reflects (or "proxies" for) both *in-vivo* genetics and physiology while allowing for systematic perturbation and characterization in high throughput.

To address this need, recent work in human genetics has seen a substantial evolution in the role of lymphoblastoid cell lines (LCLs) from their traditional use as a renewable source of DNA into an *in-vitro* model system where the genetics of human gene expression, drug response, and other traits can be studied under controlled conditions. LCLs from hundreds of individuals have already successfully been used to map the genetic determinants of RNA expression. (eQTLs) [52,53] Some of these eQTLs have even been found to confer risk to disease, including lupus and asthma. [54] [55] Attempts have likewise been made to use LCLs to understand the genetic/genomic basis of drug toxicity and response. [56] Much larger projects, involving cell-lines from thousands of individuals and ever-more extensive catalogues of genotype/phenotypes are currently under way. But, while early promising results have emerged, little has been written about the impact of the immortalization process and cell culture conditions, the relative influence of non-genetic as opposed to genetic factors on cellular trait variation, and reproducibility of measured traits. As LCLs are a potentially important adjunct to WGAS of human disease and afford opportunities to study the genetics of traits that cannot easily be assayed in human beings, we set out to address several of the pressing questions above using LCLs generated by the HapMap project, while also attempting to map novel loci for response to several commonly used chemotherapeutic drugs.

Our results are presented in full in Chapter 6. In brief summary, we show that: (1) Drug response in LCLs can be technically well measured, but is subject to substantial experimental confounders such as baseline growth-rate and metabolic properties of the cell line (both non-heritable traits) and is only modestly reproducible in independent experiments. After correcting for these confounders, no loci for drug response emerge in a GWAS in our sample with genome-wide significance; this observation is consistent with prior literature. (2) RNA expression in LCLs can likewise be well-measured, but is also only modestly reproducible, and is significantly affected by levels of EBV titer and the confounders of drug response above. Despite this, the expression of many genes is affected by genetic variation (eQTLs), but the effect sizes are usually small. (3) Baseline RNA expression predicts response to several drugs and an approach that integrates SNPs, RNA, and drug response can potentially identify novel pharmacogenetic loci acting through influence on RNA. Overall, our study indicates that it is critical to understand both non-genetic as well as genetic factors influencing *in-vitro* trait variation in LCLs. We offer practical recommendations regarding experimental design and analysis, and demonstrate suggestive evidence of three novel loci influencing response to drugs through their effects on expression of RNA.

As the need for functional follow-up of results from WGAS grows and as investigators branch out to study the genetics of ever-more complex traits, the need for reliable, human *in-vitro* model systems to proxy for *in-vivo* phenotypes will only increase. We hope our detailed analysis of one such promising model, the LCL, will serve as an important building block in the quest to create and productively use such models.

To recap, this thesis is comprised of two halves, dealing with each of the two "proxy" challenges raised above. The first half, consisting of Chapters 2-5 presents my work (jointly with wonderful colleagues) addressing the first challenge of capturing and testing common human genetic variation in association studies. Chapter 2 [39] describes the theoretical relationship between a given investment in genotyping and statistical power to detect association with a common disease trait, as well as offers practical approaches to optimize this tradeoff. Chapter 3 [40] extends the findings in Chapter 2 to realistic scenarios where marker selection and disease study takes place in different samples. Chapter 4 [41] evaluates the statistical power of current-generation genotyping platforms offered by commercial vendors. Chapter 5 [42] then concludes with an estimate of the multiple-testing burden imposed when all common variants in the genome are tested. The second half of the thesis, containing Chapter 6, addresses the second challenge by attempting to build and use *in-vitro* models for human genetics, where any (observable) trait of interest can be evaluated in depth, linked to genetic variation, and potentially expose biology not accessible in human beings. Chapter 7 then offers a summary and discussion, as well as some reflection on further lines of research. In all, the thesis represents the development and use of proxy genotypes and phenotypes that we must rely on for genetic discoveries until we reach the best case scenario imagined above. As the ideal will likely never be attained, I hope that the findings in this thesis will be useful for some time to come.

References

1. Wilschanski, M. et al. Gentamicin-induced correction of CFTR function in patients with cystic fibrosis and CFTR stop mutations. *N Engl J Med* **349**, 1433-41 (2003).
2. van den Akker-van Marle, M.E. et al. Cost-effectiveness of pharmacogenomics in clinical practice: a case study of thiopurine methyltransferase genotyping in acute lymphoblastic leukemia in Europe. *Pharmacogenomics* **7**, 783-92 (2006).
3. Mallal, S. et al. HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med* **358**, 568-79 (2008).
4. Kirsch, I. et al. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* **5**, e45 (2008).
5. Sturtevant, A.H., Bridges, C.B. & Morgan, T.H. The Spatial Relations of Genes. *Proc Natl Acad Sci U S A* **5**, 168-73 (1919).
6. Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**, 314-31 (1980).
7. Guan, W., Pluzhnikov, A., Cox, N.J. & Boehnke, M. Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium. *Hum Hered* **66**, 35-49 (2008).
8. Saunders, C.L. et al. Meta-analysis of genome-wide linkage studies in BMI and obesity. *Obesity (Silver Spring)* **15**, 2263-75 (2007).
9. Liu, W., Zhao, W. & Chase, G.A. Genome scan meta-analysis for hypertension. *Am J Hypertens* **17**, 1100-6 (2004).
10. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-7 (1996).
11. HapMap_Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
12. Balding, D.J., Bishop, M. & Cannings, C. *Handbook of statistical genetics*, (John Wiley & Sons, Chichester, England; Hoboken, NJ, 2003).
13. Schadt, E.E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).
14. Samani, N.J. et al. Genomewide association analysis of coronary artery disease. *N Engl J Med* **357**, 443-53 (2007).
15. Mootha, V.K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-73 (2003).
16. Schadt, E.E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37**, 710-7 (2005).
17. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).
18. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536-9 (1996).
19. Muller-Myhsok, B. & Abel, L. Genetic analysis of complex diseases. *Science* **275**, 1328-9; author reply 1329-30 (1997).

20. Collins, F.S., Guyer, M.S. & Charkravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580-1 (1997).
21. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
22. Li, W.H. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513-23 (1991).
23. Stoneking, M. Single nucleotide polymorphisms. From the evolutionary past. *Nature* **409**, 821-2 (2001).
24. Chakravarti, A. To a future of genetic medicine. *Nature* **409**, 822-3 (2001).
25. Cardon, L.R. & Watkins, H. Waiting for the working draft from the human genome project. A huge achievement, but not of immediate medical use. *Bmj* **320**, 1223-4 (2000).
26. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**, 139-44 (1999).
27. Antonarakis, S.E., Boehm, C.D., Giardina, P.J. & Kazazian, H.H., Jr. Nonrandom association of polymorphic restriction sites in the beta-globin gene cluster. *Proc Natl Acad Sci U S A* **79**, 137-41 (1982).
28. Ardlie, K.G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**, 299-309 (2002).
29. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229-32 (2001).
30. Gabriel, S.B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
31. Goldstein, D.B. & Cavalleri, G.L. Genomics: understanding human diversity. *Nature* **437**, 1241-2 (2005).
32. Stram, D.O. et al. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered* **55**, 27-36 (2003).
33. Johnson, G.C. et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**, 233-7 (2001).
34. Roeder, K., Bacanu, S.A., Sonpar, V., Zhang, X. & Devlin, B. Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol* **28**, 207-19 (2005).
35. Weale, M.E. et al. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* **73**, 551-65 (2003).
36. Chapman, J.M., Cooper, J.D., Todd, J.A. & Clayton, D.G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**, 18-31 (2003).
37. Lin, S., Chakravarti, A. & Cutler, D.J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* **36**, 1181-8 (2004).
38. Stram, D.O. Software for tag single nucleotide polymorphism selection. *Hum Genomics* **2**, 144-51 (2005).

39. *de Bakker, P.I. et al. Efficiency and power in genetic association studies. *Nat Genet* **37**, 1217-23 _**(*equal contributors)**_ (2005).
40. de Bakker, P.I. et al. Transferability of tag SNPs in genetic association studies in multiple populations. *Nat Genet* **38**, 1298-303 (2006).
41. Pe'er, I. et al. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* **38**, 663-7 (2006).
42. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* **32**, 381-5 (2008).
43. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nat Genet* **39**, 813-5 (2007).
44. Haiman, C.A. et al. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* **39**, 638-44 (2007).
45. Reddy, P.H. et al. Behavioural abnormalities and selective neuronal loss in HD transgenic mice expressing mutated full-length HD cDNA. *Nat Genet* **20**, 198-202 (1998).
46. Berger, M.F. et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429-35 (2006).
47. Kathiresan, S. et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med* **358**, 1240-9 (2008).
48. Saxena, R. et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-6 (2007).
49. Kimmel, S.E. Warfarin therapy: in need of improvement after all these years. *Expert Opin Pharmacother* **9**, 677-86 (2008).
50. Ranganathan, P. An update on methotrexate pharmacogenetics in rheumatoid arthritis. *Pharmacogenomics* **9**, 439-51 (2008).
51. Evans, W.E. & Relling, M.V. Moving towards individualized medicine with pharmacogenomics. *Nature* **429**, 464-8 (2004).
52. Cheung, V.G. et al. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365-9 (2005).
53. Stranger, B.E. et al. Population genomics of human gene expression. *Nat Genet* (2007).
54. Graham, R.R. et al. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc Natl Acad Sci U S A* **104**, 6758-63 (2007).
55. Moffatt, M.F. et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470-3 (2007).
56. Watters, J.W., Kraja, A., Meucci, M.A., Province, M.A. & McLeod, H.L. Genome-wide discovery of loci influencing chemotherapy cytotoxicity. *Proc Natl Acad Sci U S A* **101**, 11809-14 (2004).
57. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
58. Barrett, J.C. & Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat Genet* **38**, 659-62 (2006).

59. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11**, 241-7 (1995).
60. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* **32**, 227-34 (2008).

Chapter 2: Efficiency and power in genetic association studies

Material in this chapter is reprinted by permission from Macmillan Publishers Ltd from:

de Bakker PI*, ***Yelensky R****, Pe'er I, Gabriel SB, Daly MJ, Altshuler D.
Efficiency and power in genetic association studies.
Nat Genet. 2005 Nov;37(11):1217-23. Epub 2005 Oct 23. PMID: 16244653, Copyright 2005
*these authors contributed equally

in accordance with the following license (and Nature Group author rights):
Licensee: Roman Yelensky, License Date: Jun 03, 2008, License Number: 1961470797234
Publication: Nature Genetics, Title: Efficiency and power in genetic association studies
Type Of Use: Thesis / Dissertation

Introduction

Complete genome sequencing offers a comprehensive approach to test all human genetic variation for association to clinical traits. While routine sequencing of thousands of genomes remains impractical, it has become possible to test systematically the vast majority of human heterozygosity that is due to common genetic variations[1,2]. Correlations among nearby variants (linkage disequilibrium, LD) can improve the cost-effectiveness of such studies[3-5], guiding selection of informative "tag" SNPs[6], and providing information about nearby variants not genotyped in disease samples. The International HapMap Project is a resource that provides empirical genome-wide data to support such analyses[7,8] (see also Ref. 9).

Given practical limitations on genotyping in patient samples, investigators are forced to make a number of practical decisions, including: (*a*) selecting and prioritizing tag SNPs[10-19]; (*b*) deciding which tests of association to perform[20-26]; and (*c*) evaluating statistical significance of putative findings[27-29] (see Box for terminology). While genotyping a higher density of tag SNPs increases the fraction of sites captured through LD[30], the quantitative relationship between additional genotyping and increased power in association studies is not well described. The use of multi-marker haplotypes shifts this relationship towards greater efficiency[31], but can be a double-edged sword: if haplotype testing were to increase degrees of freedom or numbers of tests in statistical analysis, it has the potential to decrease, rather than increase, overall power[24]. Many studies will rely on data from the International HapMap Project, which is an extensive but incomplete inventory of common genetic variation[8]. Thus, it is critical to understand how tags selected from HapMap compare in power to those selected from a more comprehensive resource.

We set out to study the tradeoffs between efficiency and power for different tagging and testing approaches. Since expected power in disease association studies is the most relevant figure of merit (as compared, for example, to the distribution of correlation coefficients (r^2) between tag SNPs and untyped variants), we explicitly model disease association studies. Second, since varying both the density of tag SNPs and statistical testing procedure can influence the number of statistical tests (and many of these tests are not independent), we empirically assess significance thresholds. Finally, as results are intimately dependent on the true properties of human LD — which are not necessarily well modeled by population-genetic simulations[32] — we perform these evaluations in empirical (rather than simulated) genotype data from human samples.

BOX: TERMINOLOGY

Variants to be tested for association to phenotype are **putative causal alleles**.

The hypothesized relationship between alleles and phenotype(s) is termed the **genetic model**.

Genotype data used to guide experimental design (tag SNP selection and definition of statistical tests to be performed) is termed the **reference panel**; HapMap is one such panel.

Tags are the subset of variants genotyped in a disease study. SNPs that are not typed in the study, but whose effect can be studied through LD with a tag, are termed **proxies.** Where the correlation between a tag and untyped putative causal alleles is perfect (r^2 = 1.0), we refer to a **perfect proxy.**

The allelic hypotheses examined for association to disease (based on genotypes of the tags) are termed **tests**. A test can simply be the allele of a tag — termed a **single-marker test**. Tests based on combinations of tags are **multi-marker tests**. A **specified multi-marker test** examines a particular allelic combination (a haplotype) of multiple tags based on its observed correlation to a putative causal (untyped) allele in the reference panel. An **exhaustive multi-marker test** searches over many or all allelic combinations of tags in the hope of finding a test that captures a hitherto unseen putative causal allele.

Results

Disease association studies based on empirical genotype data

We started by creating case-control panels based on empirical genotype data from the HapMap-ENCODE project[8]. These ten 500 kb regions were sequenced in 48 individuals with all SNPs discovered (as well as any others in dbSNP) genotyped in 269 HapMap samples: 30 trios of the Yoruba in Ibadan, Nigeria (abbreviation: YRI), 30 trios from Utah, USA, with European ancestry (CEU), 45 Han Chinese from Beijing, China (CHB), and 44 Japanese from Tokyo, Japan (JPT). This data set contains 16,970 SNPs (one every ~300 bp) with an allele frequency distribution that is near-complete for common alleles, and is available for download at http://www.hapmap.org.

To simulate a case-control panel, one SNP from this data set was nominated as "causal". An effect size was calculated so that if this SNP were directly tested in 1,000 cases and 1,000 controls, power would be 95% to achieve nominal $P = 0.01$. Since our concern was the relative effect on power of tagging and analysis strategies (rather than absolute power), and to make it possible to average results over all putative causal alleles, we fixed the absolute power for each putative causal SNP. Constant power requires minor allele frequency to be inversely correlated to penetrance: in this model, rare alleles are assigned a stronger effect than common alleles (Supplementary Fig. 1). This approach further avoids consideration of uninformative scenarios where power is uniformly high (such that any tagging strategy might suffice), or non-existent (such that tagging is irrelevant).

To simulate the case/control studies, chromosomes spanning each 500 kb region were drawn at random from the phased empirical data, conditional on the genotype and effect size at the nominated "causal" SNP. This was repeated until 1,000 cases and 1,000 controls were obtained in each panel, and then 25 such panels were created for each causal SNP. Finally, the entire process was iterated over all SNPs in the data, resulting in a large collection of case-control panels in which each SNP has an equal chance of being causal.

Tag SNPs were selected and statistical tests defined from a reference panel under a variety of scenarios as described below. Association was evaluated for each statistical test using standard 2x2 chi-square comparisons of cases and controls. The significance threshold for declaring association was based on the empirical null distribution: the tags and statistical tests selected in each scenario were examined in a set of null panels (in which no SNP is causal), with the maximum χ^2 value exceeded in 1% of null panels chosen as the threshold to declare a positive result (region-wide corrected $P = 0.01$). In the figures we report the proportion of case-control panels in which an association was detected, averaged over all putative causal SNPs and over all ENCODE regions.

Capturing all sites observed in a complete reference panel

We began by examining the relationship between the number of SNPs genotyped and statistical power in the best-case scenario: where complete resequencing has been performed in a reference panel, such that all putative causal alleles have been observed. In the first instance we examined only common alleles: tags were selected to capture alleles ≥ 5% in frequency in the reference panel, and the set of putative causal alleles in the simulations were limited to those present at ≥ 5%.

Figure 1 shows the distribution of the maximum χ^2 values under the null and over all causal panels. While nominal power is set to 95% if each causal site is examined as a single test, the average power after testing all common sites in each 500 kb region falls to 60% (YRI) and 68% (CEU and CHB+JPT). This decline simply represents the power loss resulting from an empirical correction for having tested many hundreds of SNPs within each 500 kb region, with the decline in power tracking with the extent of LD in each set of DNA samples.

The simplest and most conservative approach to tag SNP selection is to select a subset of non-redundant SNPs from the reference panel such that every common allele is either directly genotyped, or has a perfect proxy ($r^2 = 1.0$) among the tags. The reduction in the number of genotypes required (as compared to testing all common SNPs directly) was 46% (YRI) and 65% (CEU and CHB+JPT) (Figure 2). Of course, since all sites are perfectly captured, power remains at 100% as compared to testing all common causal alleles directly. (From this point onward we will report the "relative power" of each tagging strategy — that is, power under a given tagging/testing strategy as compared to that obtained by testing all common sites directly.)

We next asked if multi-marker (haplotype) tests can improve the genotyping efficiency, as proposed elsewhere[10,31]. Because we were concerned about loss of power due to the introduction of additional statistical tests, we developed a strategy (see Methods) in which an identical set of 1 d.f. tests of association are performed, except that we allow a haplotype of tags to serve as surrogate for an untyped SNP (rather than restricting statistical tests to genotypes of single tags). That is, if a specific multi-marker combination (i.e., haplotype of tag SNPs) can serve as an effective proxy for another tag SNP, then that latter tag need not be included for genotyping. In this method, each single tag, as well as each specific haplotype defined above, is tested for association. To avoid over-fitting, we require that the tags in a specified multi-marker test are themselves in strong LD (LOD > 3.0) to the allele predicted.

Using this tagging procedure in simulated disease association studies as above, we computed power and the number of tag SNPs required. In comparison to pairwise tagging, power remains unchanged at 100%, but the number of tag SNPs that need be genotyped is reduced by another 26% (YRI), 30% (CEU) and 28% (CHB+JPT) (Figure 2). Thus, simply by removing redundancy from the complete set of SNPs in an efficient haplotype-based manner, we can reduce the genotyping burden by 60-77% while maintaining complete power.

Increasing efficiency by relaxing thresholds for tag SNP selection

The strategies above require that tags be selected to capture perfectly every common site observed in the reference panel. To the extent that this is unaffordable, investigators may be forced to reduce the density of genotyping by relaxing the criteria for tag selection. Two possibilities we examined are (*a*) capturing all common alleles, but at a less stringent r^2 threshold[14], or (*b*) by choosing to capture only a subset of sites, each at a high r^2 threshold.

For example, relaxing the threshold from perfect correlation to a slightly lower level ($r^2 \geq 0.8$) decreases the number of tags required substantially (a further decrease of 36% in YRI, 47% in CEU and 55% in CHB+JPT), and yet relative power remains nearly complete at 96%. Moreover, this approach can straightforwardly be combined with the multi-marker method described above, resulting in even greater efficiency (Figure 3a). While even lower r^2 thresholds result in less and less genotyping, relative power begins to decline rapidly. In fact, we find that lowering the r^2 threshold too far (while still requiring that all sites be captured at or above this threshold) can result in performance no better than random collection of SNPs (Figure 3a).

An alternative approach is to rank order potential tags based on the number of other SNPs for which they proxy, and then type the SNPs in this priority order (we term this method "best *N*"). This approach is significantly more efficient than lowering the r^2 threshold: for example, choosing a SNP every 10 kb in this manner (only ~5% of all common SNPs) provides relative power of 77% (YRI), 95% (CEU) and 92% (CHB+JPT). Of course, any such pairwise list can be made more efficient by replacing single-marker tests with appropriate multi-marker haplotypes (as above), resulting in the most efficient method of those we examined (Figure 3a).

In summary, if a complete reference panel is available, multi-marker haplotype tests are more efficient than pairwise tests, and prioritizing SNPs based on their LD properties allows impressive reductions in the genotyping burden while maintaining excellent power.

Tags selected from an incomplete reference panel

At present, only incomplete reference panels are available genome-wide[8,9]. It is therefore important to ask how power and efficiency decline when tags are selected from an incomplete, rather than complete, reference panel. To this end, we created a "pseudo" 5 kb HapMap by thinning the ENCODE data to achieve the spacing and frequency distribution of Phase I HapMap[8]. We selected tags and designed tests using this incomplete resource, evaluating performance in simulated case-control panels where *all* alleles (not just those from the incomplete HapMap) were allowed to be causal.

We observe two major changes, both unsurprising. First, a much smaller set of tags is selected for genotyping as compared to when tags are picked using the complete data (Figure 3b). Second, a subset of common variants have no good proxies in the reference panel: 55% (YRI), 26% (CEU) and 28% (CHB+JPT) of *all* common SNPs are not captured at $r^2 \geq 0.8$, because they are not observed in the pseudo HapMap, nor are they in LD with any other SNP that happened to be included[8].

Given these characteristics, it is noteworthy that power is largely undiminished relative to testing tags chosen from a reference panel of *all* common sites: tags selected from the pseudo Phase I HapMap (pairwise $r^2 \geq 0.8$) provide 91% relative power in CEU (73% in YRI; 89% in CHB+JPT), despite requiring less than half as many tags compared to tagging from complete data. In absolute terms, while a set of "best *N*" tags every 10 kb (on average) selected from complete data provides 95% relative power in CEU (77% in YRI; 92% in CHB+JPT), the same density of tags selected from the pseudo Phase I HapMap retains 88% power in CEU (64% in YRI; 85% in CHB+JPT).

We also asked whether the power provided by different tagging strategies was similar when performed on incomplete as compared to complete reference panels. Interestingly, whereas "best *N*" clearly outperformed lowering the r^2 threshold in complete data, this is no longer the case for tagging from the pseudo HapMap (Figure 3b). Here, the two methods perform similarly, with an apparent slight edge to lowering the r^2 threshold.

The impact of LD on tag SNP selection and power

We were initially surprised by the relatively high power obtained when tags were selected from incomplete reference panels or when the "best *N*" method was used to trim a complete tag set, as in both cases these tags fail to capture a substantial proportion of putative causal alleles. This behavior is illuminated, however, by the highly variable extent of LD in the human genome, and the impact of LD on the power obtained from each statistical test.

The completeness of the reference panel — and the strategy for tagging and testing — affects not only the distribution of the test statistic for causal SNPs, but also the significance thresholds under the null. Figure 4a displays the distribution of the maximum χ^2 test statistic under two scenarios: tags selected from complete and from incomplete reference panels. When tags are selected from incomplete data, as expected the causal distribution is shifted towards lower χ^2 values, since some causal SNPs are not well captured. But in addition, the null distribution shifts to lower thresholds due to a marked reduction in the number of tests performed. That is, although some alleles are poorly captured and not discovered — most notably those alleles with few proxies (Supplementary Fig. 2), the power for the majority of putative causal alleles remains high due to (*a*) inclusion of a good proxy for most causal alleles, and (*b*) a less stringent significance threshold for declaring association. While overall power is similar in both scenarios (Figure 4b), the mix of causal alleles discovered shifts toward those in LD with many other SNPs, at the cost of discoveries due to SNPs with few proxies.

Put another way, tests that capture many putative causal alleles add the same amount to the multiple testing burden as do independent tests that capture only a single site. The chance of encountering a true association, however, is much greater when many putative causal alleles are examined per test. While the first tag from the incomplete reference panel captures only a small fraction of all sites, it does so at the cost of only a single hypothesis test, and results in relative power that is 15-25% of that obtained by testing all common sites in the region (data not shown). Of course, adding more tags captures an ever larger fraction of putative causal alleles, and power rises. But the yield of each additional test falls monotonically as it examines a smaller slice of the prior distribution than the test before it.

This simple idea underlies the "best *N*" method for tag SNP selection, as it preferentially excludes those SNPs that have no proxies, and which offer the least marginal power per hypothesis test. Similarly, an incomplete reference panel (HapMap) has also preferentially (but imperfectly) dropped SNPs with no proxies – such SNPs can only be tested for association if they are included on the HapMap, while SNPs with many proxies will almost always be tested as only one of its proxies needs to be present on HapMap. The "best *N*" approach fails at sparser densities (in complete and incomplete data, Figure 3), however, because the set of SNPs with no proxies has been depleted, and thus the tags being dropped carry with them information about an increasingly larger number of putative causal alleles. The "best *N*" method suffers further when run on incomplete reference panels, because from such data it is not possible to distinguish which SNPs truly have no proxies, and which actually have proxies that have not yet been typed. Empirically, about 50% of the SNPs on the pseudo HapMap have no observed proxies (at $r^2 = 1$), and thus are preferentially dropped using "best *N*". Of these, a large number actually do have proxies in the complete data, but it is impossible to tell which are which without more complete data. Thus, where complete data is available (as in selected candidate genes[33]), and as denser versions of HapMap become available (such as the pending Phase II), the impact of the "best *N*" method should become more significant, particularly for choosing marker densities of more than 1 SNP per 10 kb.

Exhaustive haplotype tests to detect less common alleles

Above we considered only scenarios in which the causal alleles are common. Of course, less common SNPs also influence disease, and might be discovered incidentally even if tags are selecting and tests designed only to capture common variation. We thus examined power under the scenario that the causal allele is < 5% in frequency with the same 95% nominal power (and thus a larger magnitude of effect). Interestingly, while power for < 5% alleles is reduced compared to that enjoyed for common alleles, it remains substantial: relative power of 29% (YRI), 23% (CEU) and 15% (CHB+JPT).

Exhaustive haplotype testing has been suggested as an approach to capture alleles not observed in the reference panel. This approach tests many or all local haplotypes in the hope that one or more might correspond to an *unobserved* causal allele[25]. The chance of capturing an unobserved allele is likely to be increased with exhaustive haplotype testing, because a better proxy for the putative causal allele is obtained. However, this benefit comes at the cost of numerous additional statistical tests, many of which do not correspond to any actual variant.

We first evaluated the scenario where exhaustive haplotype testing is performed on tags picked to capture all common alleles in the complete reference panel ($r^2 = 1.0$), but where the universe of causal alleles was limited to those with < 5% frequency. As described previously[25], exhaustive haplotype testing increases relative power: 59% (YRI), 58% (CEU) and 45% (CHB+JPT) (Figure 5a). That is, for less common alleles, the benefit of finding a better proxy outweighs the cost of multiple comparisons, and results in substantial power even as compared to testing the less common alleles directly.

In contrast, when the causal alleles were common (≥ 5%), relative power is reduced by exhaustive haplotype testing to ~85% (Figure 5a). This penalty is not surprising: the testing burden is increased with no possibility of true benefit, since all putative causal alleles are already captured.

It seemed more likely that exhaustive haplotype tests might improve power for tags selected from incomplete data, or at random. When we selected tags from the incomplete (pseudo Phase I HapMap) reference panel, or at random at lower densities (one common SNP per 10 kb and 30 kb), exhaustive haplotype tests continued to boost power for less common alleles, but failed to improve power for common alleles (Figure 5b, Supplementary Fig. 3). We conclude that in empirical genotype data the benefit of exhaustive haplotype tests is very real, but primarily limited to lower frequency alleles.

Software

The optimal tradeoff between power and efficiency depends on the resources available and assumed characteristics of allele frequency and LD for putative causal alleles. Since investigators will want to make their own decisions, we have implemented these methods in the web server Tagger (http://www.broad.mit.edu/mpg/tagger/), and the program Haploview[34] (http://www.broad.mit.edu/mpg/haploview/). The software enables investigators (*a*) to select tags from empirical data, using single-marker or specified multi-marker tests, (*b*) to rank-order the tags based on proxy count, (*c*) to record the statistical tests to be performed on these tags (single-marker tests, specified multi-marker tests, or exhaustive tests). Haploview can perform association tests based on these selections, including permutation testing. The software also makes it possible (*d*) to force in or exclude specific sets of SNPs as tags based on other considerations, such as the existence of previous data or a working assay; (*e*) to incorporate genotyping platform design scores to pick tags based on the likelihood of success; (*f*) to evaluate the coverage with respect to a reference panel (based on r^2) for an existing set of user-specified tags; and (*g*) to derive specified multi-marker tests from a static list of tags to extend coverage with respect to a reference panel (such as HapMap).

Discussion

In summary, our analyses indicate that (*a*) specified multi-marker tests substantially increase tagging efficiency as compared to single-marker approaches, without loss of power; (*b*) when selecting SNPs from very dense reference panels, a method such as "best *N*" which rank orders SNPs based on their number of proxies allows dramatic reductions in genotyping with limited loss of power—substantially outperforming a method based on relaxing r^2 thresholds; (*c*) sparser sets of tags selected from a pseudo Phase I HapMap are nearly as powerful as equally sized sets chosen from complete reference panels; and (*d*) exhaustive multi-marker tests improve power for less common causal alleles, but are neutral or reduce power when the causal SNP is common. These relationships hold for each of the different population samples studied by HapMap, although the number and performance of tags varies as expected based on the general extent of LD in each sample.

It has become common practice to select tags until a high threshold for the correlation coefficient (often $r^2 \geq 0.8$) is exceeded for all observed sites[14]. The use of multi-marker tests and prioritization of tags permits cost to be substantially reduced, with little loss of power. Whether it is attractive to take advantage of this tradeoff of efficiency for power will be different for each investigator, depending on available resources for genotyping, the sample size and power of the patient sample, the perceived cost of a false negative study, and the anticipated value of a true positive result.

Whether exhaustive haplotype testing is justified depends on assumptions about the relative balance of rare and common causal variants, and the completeness of the reference panel from which tags are picked. Given the current Phase I HapMap, there appears little cost and evident gain to employing the exhaustive haplotype test[25]. As reference panels become more complete (particularly for less common alleles), however, the balance may shift towards the specified haplotype-based method that limits tests to only those that predict the increasingly complete inventory of putative causal sites.

A limitation of our study is that we do not evaluate whether tags and tests defined in the HapMap samples are transferable across populations, and how this varies for single-marker and haplotype-based methods. In preliminary analysis we observe minimal loss of power when tags and tests are transferred to a variety of disease studies (PIWdB, Noel Burtt, Rob Graham, MJD, DA, manuscript in preparation), and similar findings have been reported elsewhere[11,35,36]. Much more work is needed on this topic, and the answer will likely vary depending on the population studied.

In our minds, the most significant observation in this study is that SNPs that capture many putative causal alleles have different statistical properties than tests capturing only a single site — at least, under the frequentist approach to setting statistical thresholds. An implication is that rather than using a universal significance threshold for all tests, power may be increased by a Bayesian approach in which a prior for each test is established as a function of the number of sites captured, integrated over each site's individual likelihood of being causal. Incorporating such ideas into study design may lead to greater efficiency in use of genotyping resources, and maximize the yield of discoveries for a given investment in such research.

Methods and Materials

Data sets

We used phased genotype data for ten chromosomal regions, each spanning 500 kb, generated as part of the HapMap-ENCODE project (http://www.hapmap.org/downloads/encode1.html.en). This data set (release 16c.1) is based on genotyping all variable sites observed after resequencing 48 unrelated individuals (as well as any additional SNPs in dbSNP) in the 269 DNA samples used in HapMap: 30 parent-offspring trios from the Yoruba people of Ibadan, Nigeria (YRI); 30 parent-offspring trios from Utah residents with northern and western European descent (from the Centre d'Etude du Polymorphisme Humain; CEU); 45 unrelated Han Chinese from Beijing, China (CHB); and 44 unrelated Japanese from Tokyo, Japan (JPT). We have combined the CHB and JPT samples for all analyses performed, yielding three analysis panels: YRI, CEU (both 120 unrelated chromosomes) and CHB+JPT (178 chromosomes).

Genetic model and simulation of case-control panels

From the ENCODE data, we generated almost 10 million case-control panels to evaluate study-wide power as a function of a number of tagging/testing strategies. A multiplicative disease model was employed in which we nominate all non-singleton SNPs in the complete data to be causal, one by one, reflecting a uniform prior probability of any of the SNPs contributing to the phenotype. For each causal SNP, 250 replicate case-control panels were made by sampling with replacement from the ENCODE chromosomes to give 1,000 cases and 1,000 controls (4,000 chromosomes in total). The frequency of the causal allele (minor/major chosen at random) in the cases is determined by the genotype relative risk, calibrated so that we obtain 95% nominal power to detect an association with the 1 d.f. chi-square test (at $P < 0.01$), if that causal SNP was tested directly (Supplementary Fig. 1). Thus, all causal SNPs are assigned to have equal nominal power. We also created control-control (null) panels by randomly sampling from the ENCODE chromosomes; these were used to define statistical significance thresholds.

Reference panels for tag SNP selection

Reference panels were constructed at two densities: (1) "complete" reference panels based on all ENCODE data (120 unique chromosomes for YRI and CEU; 178 for CHB+JPT), where complete refers to the ascertainment of common (≥5%) variation; and (2) "incomplete" reference panels by thinning the data as follows. To mimic the ascertainment scheme of the 5 kb HapMap (Phase I), we randomly picked SNPs present in dbSNP build 121 (excluding "non-rs" SNPs in HapMap release 16a) for every 5 kb bin until a common (MAF ≥ 5%) SNP was picked (allowing up to three attempts per bin).

Selection of tag SNPs and definition of tests

We have developed a computer program called Tagger for selecting tag SNPs and defining tests from a reference panel. Tags can be picked in different ways: (1) greedy pairwise tagging[14], in which alleles of interest are captured by single-marker tests at the prescribed r^2; (2) prioritizing tags ("best *N*") by the number of alleles they can proxy for at a given r^2. In addition, Tagger can perform an aggressive search to attempt to replace each tag with a specific multi-marker predictor (on the basis of the remaining tags) to improve efficiency. This predictor will be accepted only if it can capture the alleles originally captured by that discarded tag at the required r^2; otherwise, that tag is considered indispensable. As a result of this "peel back" approach, we end up with fewer tags that specify a similar (identical if $r^2 = 1$) set of 1 d.f. statistical tests as the original set of single-marker tests. In this study, we allow up to three tags to form a specified multi-marker test, and limit the search to evaluate at most 10,000 allelic predictors. The maximum allowed physical distance between an allele and a tag was 200 kb. To minimize risk of overfitting, tags within a specified multi-marker test are forced to be in strong LD (here defined as the LOD score on $|D'| > 3$) with one another and with the predicted allele.

Region-wide test statistic and power calculations

For every explored tagging/testing scenario, we generate a set of 1 d.f. chi-square allelic tests. Our region-wide test statistic for association is the maximum of these chi-square values. The null distribution of the test statistic was generated by performing the same allelic tests in the random null panels, and used to derive the significance threshold corresponding to a region-wide $P = 0.01$ (see Supplementary Note for a brief discussion on how this compares to explicit permutation testing). The absolute power to detect association is computed as the fraction of the case-control panels in which the maximal chi-square test statistic exceeds the significance threshold (when a true association is declared). To normalize results for different strategies, we report power (for both common and rare causal alleles) relative to the power to detect common causal alleles (MAF ≥ 5%) when these are tested directly, averaged over all 10 ENCODE regions.

Exhaustive haplotype tests

Exhaustive haplotype tests were performed by enumerating all haplotypes corresponding to adjacent combinations of tags of all sliding windows of a maximum span. We applied this to pairwise tags (selected at $r^2 = 1$ from complete panel) forming haplotypes of up to 25 kb, and 17 and 50 random common markers per region (30kb and 10 kb average spacing, respectively) from incomplete reference panels forming haplotypes of up to 100 kb. Allelic chi-square tests were performed on these haplotypes, as above.

Figures

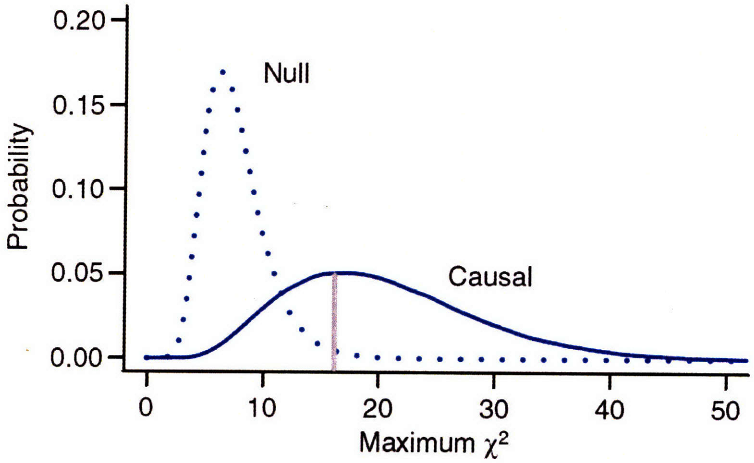


Figure 1: Distributions of the test statistic in a typical ENCODE region. Maximum chi-square statistics for association to disease status are evaluated in the simulated case-control panels (solid line) and random null panels (dotted line). The study-wide significance threshold (vertical grey line) is empirically determined such that the maximum chi-square test statistic exceeds it in 1% of the null panels (region-wide $P = 0.01$). True associations in the simulated case-control panels with a test statistic below the threshold are rejected (false negatives). Due to the empirical multiple testing correction, absolute power to detect an association drops from 95% (nominal) to 60% (YRI) and 68% (CEU and CHB+JPT), averaged over all 10 ENCODE regions.

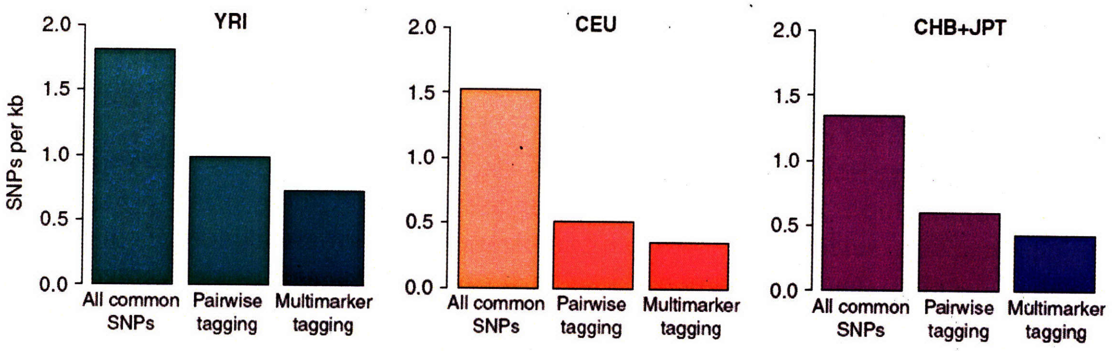


Figure 2: Efficiency afforded by a tagging approach. Using a pairwise tagging and single-marker analysis strategy, a non-redundant ($r^2 = 1$) subset of all SNPs provides 100% relative power to capture all common SNPs (≥ 5%) in the ENCODE data. Efficiency is increased further, while retaining 100% relative power, by the use of multi-marker haplotypes as described in the text.

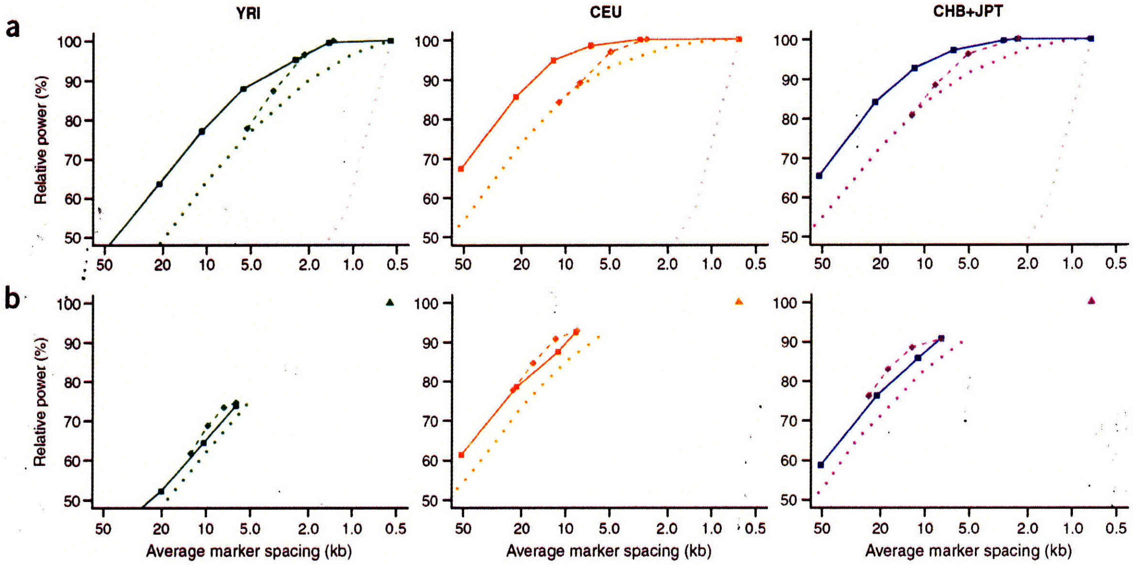


Figure 3: Efficiency and power for various tagging strategies. Relative power to detect associations due to common (≥ 5%) causal alleles is shown as a function of the average spacing of tags for multi-marker tagging from (*a*) complete reference panels and (*b*) incomplete reference panels (pseudo Phase I HapMap). Tags are picked by random selection of common SNPs (dotted line); by lowering the r^2 threshold (dashed line); and by prioritizing "best *N*" tags according to number of proxies (solid line). In the top panel, expected power is displayed for a hypothetical scenario in which there is no LD among SNPs, and all tests are independent (grey dotted line); the comparison of this line to the actual data shows the tremendous gain in efficiency and power offered by the extensive LD in the human genome.

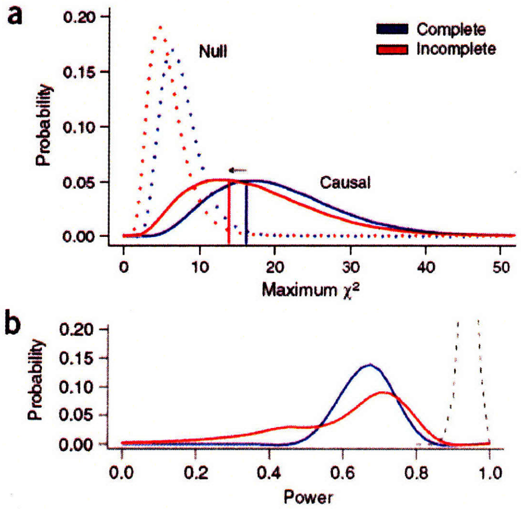


Figure 4: Effect of tagging from an incomplete reference panel on the testing burden and power. (*a*) Null (dotted lines) and causal (solid lines) distributions of the test statistic are plotted for two scenarios: tagging from complete (in blue) and incomplete (in red) reference panels. The causal distribution as well as the region-wide significance threshold are reduced concomitantly when tags are picked from the pseudo Phase I HapMap, thus preserving power. (*b*) Distribution of region-wide power for individual causal alleles is plotted for two scenarios: tagging from complete (in blue) and incomplete (in red) reference panels. Nominal power is centered around 95% (dotted grey line). Overall power is comparable in both scenarios, for reasons discussed in the text.

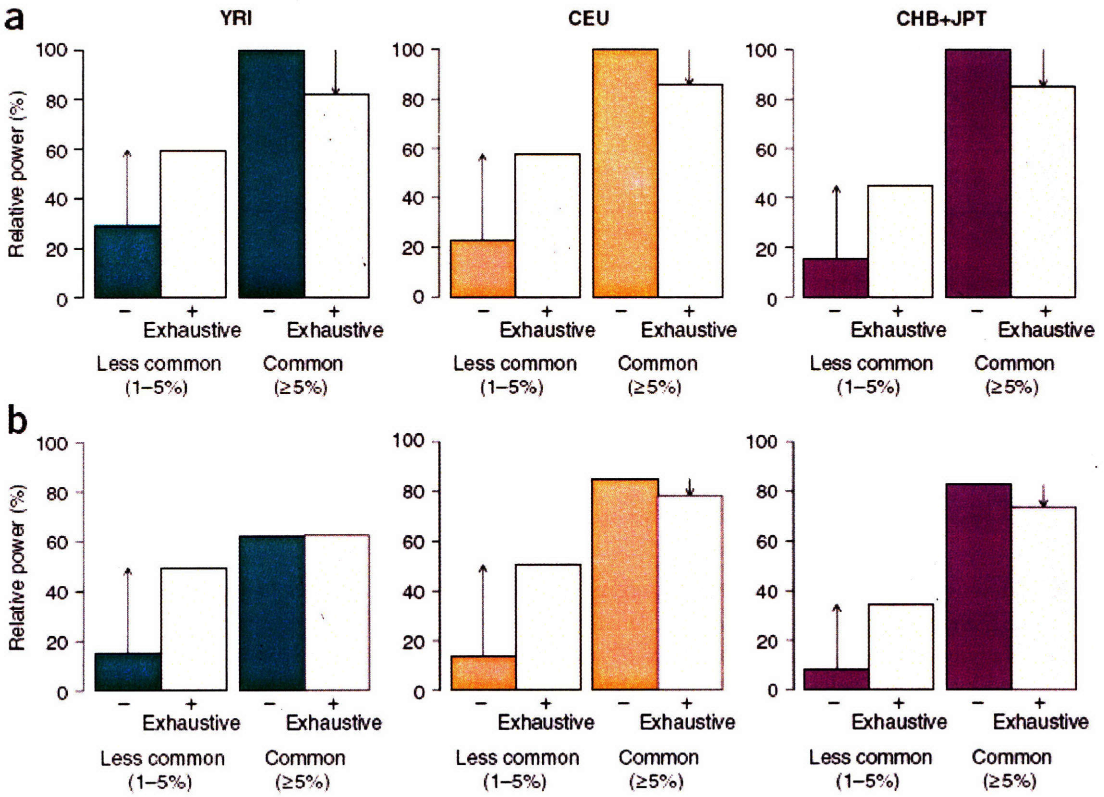
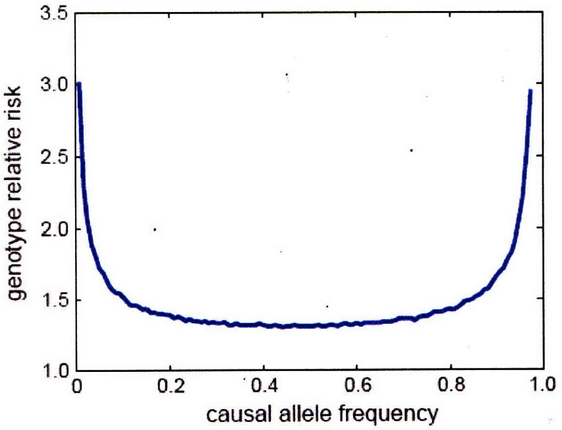
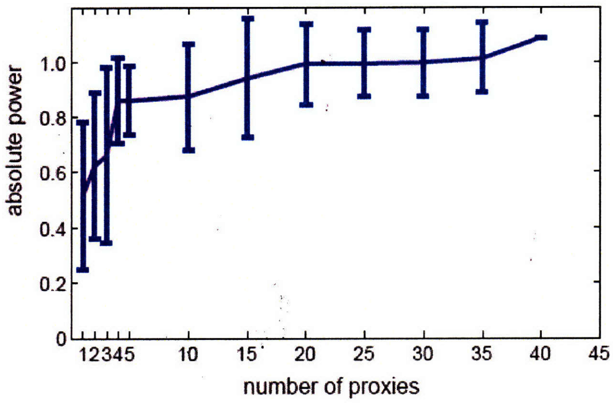


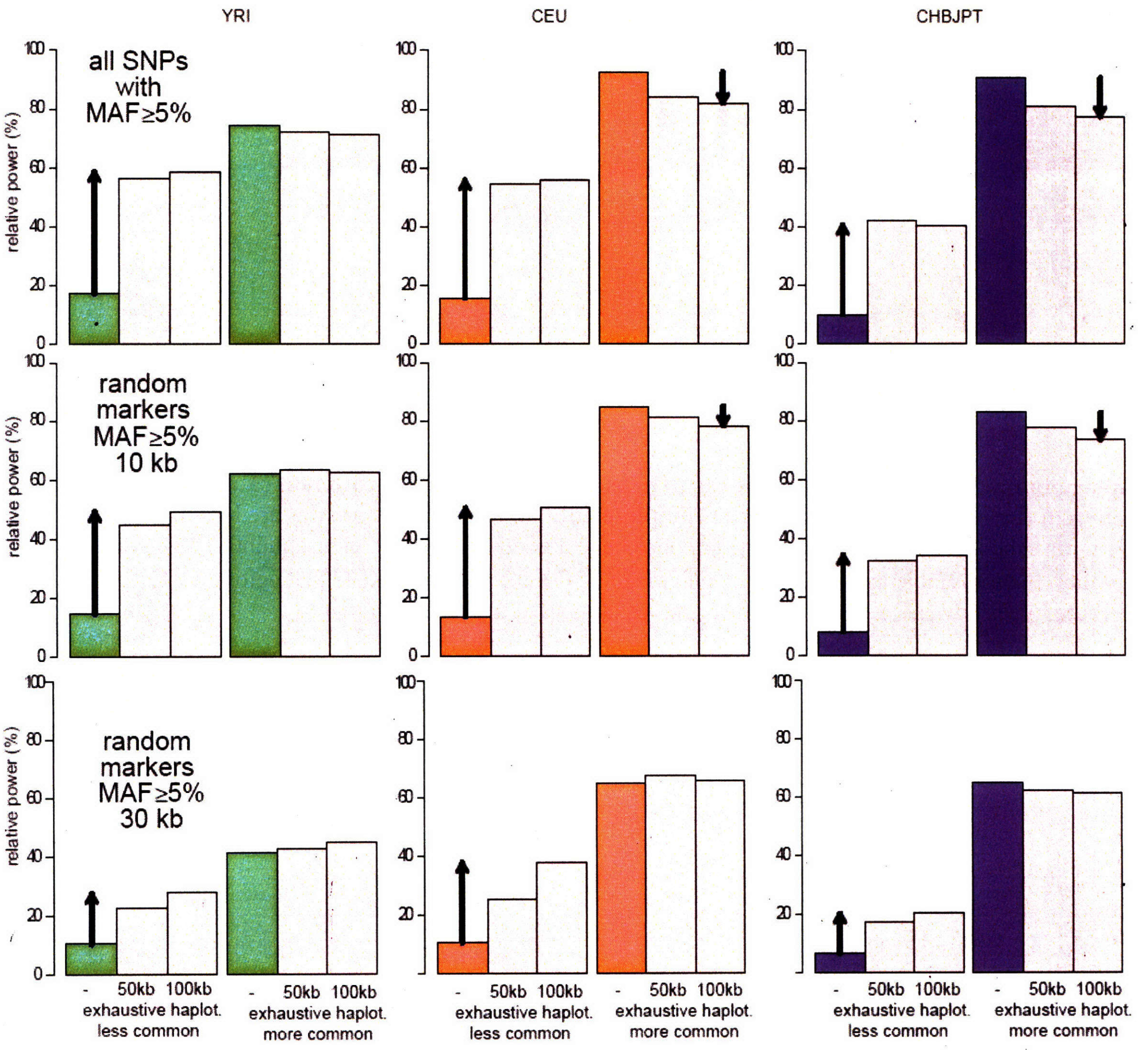
Figure 5: Effect of exhaustive haplotype tests on statistical power. Relative power is given for common (≥ 5%) and less common (< 5%) causal alleles for two scenarios: (*a*) when a non-redundant set of SNPs are used as tags from complete reference panels; and (*b*) when tag SNPs (MAF ≥ 5%) are randomly selected every 10 kb from pseudo Phase I HapMap (as in Ref. 25). Power is computed when each tag SNP is tested for association using single-marker tests (–), and when exhaustive haplotype tests are performed on the same data (+). Exhaustive haplotype tests increase power for less common alleles, but at a cost of reduced power for common alleles.



Supplementary Figure 1: Genotype relative risk as a function of the frequency of the causal variant. Less common alleles are assigned a greater effect size than common alleles to achieve constant nominal power of 95%.



Supplementary Figure 2: Absolute power to detect association for all common causal variants as a function of the number of proxies in the complete data. SNPs with few proxies are less likely to have made it onto the incomplete reference panel, or to have a good proxy on it. Consequently, tags picked from the pseudo Phase I HapMap (shown here: pairwise $r^2 \geq 0.8$ for a single ENCODE region in CEU) contribute lower power for SNPs with few proxies compared to those with many.



Supplementary Figure 3: Exhaustive haplotype testing on tags picked from incomplete reference panels. The effect of exhaustive haplotype testing on relative power is given for tags picked from pseudo Phase I HapMap at different levels of completeness (all common SNPs; 1 common SNP every 10 kb; 1 common SNP every 30 kb). Power to detect association for common (≥ 5%) causal alleles takes a modest hit in power, while improving power for rare (< 5%) causal alleles. Power is expressed relative to the power when all common sites in the complete reference panel are tested.

References

1. Wang, W.Y., Barratt, B.J., Clayton, D.G. & Todd, J.A. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* **6**, 109-18 (2005).
2. Carlson, C.S., Eberle, M.A., Kruglyak, L. & Nickerson, D.A. Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446-52 (2004).
3. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229-32 (2001).
4. Gabriel, S.B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
5. Patil, N. et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719-23 (2001).
6. Johnson, G.C. et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**, 233-7 (2001).
7. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-96 (2003).
8. The International HapMap Consortium. A haplotype map of the human genome. *Nature* (2005). **In the press.**
9. Hinds, D.A. et al. Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072-9 (2005).
10. Stram, D.O. et al. Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered* **55**, 27-36 (2003).
11. Weale, M.E. et al. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* **73**, 551-65 (2003).
12. Ke, X. & Cardon, L.R. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19**, 287-8 (2003).
13. Meng, Z., Zaykin, D.V., Xu, C.F., Wagner, M. & Ehm, M.G. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet* **73**, 115-30 (2003).
14. Carlson, C.S. et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **74**, 106-20 (2004).
15. Hu, X., Schrodi, S.J., Ross, D.A. & Cargill, M. Selecting tagging SNPs for association studies using power calculations from genotype data. *Hum Hered* **57**, 156-70 (2004).
16. Halldorsson, B.V. et al. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Res* **14**, 1633-40 (2004).
17. Ao, S.I. et al. CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics* **21**, 1735-6 (2005).
18. Zhang, K. et al. HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics* **21**, 131-4 (2005).
19. Rinaldo, A. et al. Characterization of multilocus linkage disequilibrium. *Genet Epidemiol* **28**, 193-206 (2005).

20. Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M. & Poland, G.A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* **70**, 425-34 (2002).
21. Zaykin, D.V. et al. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* **53**, 79-91 (2002).
22. Fan, R. & Knapp, M. Genome association studies of complex diseases by case-control designs. *Am J Hum Genet* **72**, 850-68 (2003).
23. Stram, D.O. et al. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered* **55**, 179-90 (2003).
24. Chapman, J.M., Cooper, J.D., Todd, J.A. & Clayton, D.G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**, 18-31 (2003).
25. Lin, S., Chakravarti, A. & Cutler, D.J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* **36**, 1181-8 (2004).
26. Roeder, K., Bacanu, S.A., Sonpar, V., Zhang, X. & Devlin, B. Analysis of single-locus tests to detect gene/disease associations. *Genet Epidemiol* **28**, 207-19 (2005).
27. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).
28. Nyholt, D.R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* **74**, 765-9 (2004).
29. Dudbridge, F. & Koeleman, B.P. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet* **75**, 424-35 (2004).
30. Wang, W.Y. & Todd, J.A. The usefulness of different density SNP maps for disease association studies of common variants. *Hum Mol Genet* **12**, 3145-9 (2003).
31. Goldstein, D.B., Ahmadi, K.R., Weale, M.E. & Wood, N.W. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet* **19**, 615-22 (2003).
32. Schaffner, S.F. et al. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* (2005).
33. Crawford, D.C. et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet* **74**, 610-22 (2004).
34. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-5 (2005).
35. Nejentsev, S. et al. Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. *Hum Mol Genet* **13**, 1633-9 (2004).
36. Ahmadi, K.R. et al. A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat Genet* **37**, 84-9 (2005).

Chapter 3: Transferability of tag SNPs in genetic association studies in multiple populations

Material in this chapter is reprinted by permission from Macmillan Publishers Ltd from:

de Bakker PI, Burtt NP, Graham RR, Guiducci C, ***Yelensky R***, Drake JA, Bersaglieri T, Penney KL, Butler J, Young S, Onofrio RC, Lyon HN, Stram DO, Haiman CA, Freedman ML, Zhu X, Cooper R, Groop L, Kolonel LN, Henderson BE, Daly MJ, Hirschhorn JN, Altshuler D.
Transferability of tag SNPs in genetic association studies in multiple populations.
Nat Genet. 2006 Nov;38(11):1298-303. Epub 2006 Oct 22.
PMID: 17057720 [PubMed - indexed for MEDLINE], Copyright 2006

in accordance with the following license (and Nature Group author rights):
Licensee: Roman Yelensky, License Date: Jun 03, 2008, License Number: 1961471110297
Publication: Nature Genetics, Title: Transferability of tag SNPs in genetic association studies in multiple populations, Type Of Use: Thesis / Dissertation

Introduction

The International HapMap Project provides empirical genotype data for >4 million SNPs in a limited sample of 270 individuals from four populations[1,2]. There are two fundamental questions with regard to a dense reference panel such as HapMap. First, to what extent is power compromised when tags are selected from incomplete genotype data in the reference panel? Second, how is power affected when tags are selected from a reference panel but then genotyped in another population sample? We addressed the first question in previous work, where we investigated the quantitative relationship between marker density and power in simulated association studies using HapMap ENCODE data[3]. Here, we characterize the extent to which tag SNPs picked from HapMap DNA samples are transferable across different population samples.

Results

To this end, we have collected dense genotype data uniformly across HapMap and non-HapMap population samples. As part of the Multiethnic Cohort (MEC) study[4,5], we first compiled a list of genes in the steroid hormone and growth factor pathways for a comprehensive study of genetic variation. We selected a dense set of SNPs from the public dbSNP database[6], augmented by SNP discovery through exon resequencing in each of these genes in 190 cases with breast and prostate cancer from five different ethnic groups (**Supplementary Table 1**). In total, we attempted genotyping for 3,302 SNPs in over one thousand DNA samples from 15 different population samples (**Table 1** and

Supplementary Table 2). Keeping all SNPs that were successfully genotyped in all population samples and polymorphic in at least one (**Supplementary Table 3**), the final data set contained 1,679 SNPs across 25 genes with a total span of 2.6 Mb (**Supplementary Table 4**). With an average marker density of 1 SNP per 1.6 kb (approaching that of Phase II HapMap), this data set provides good coverage of common variation, consistent with previous evaluations[2,3].

To assess the transferability of tags picked from HapMap samples for association studies in other population samples, two relevant measures are (a) the distribution of the correlation (r^2) between the allelic tests (based on the tags) and the collection of all "untyped" variants (that is, SNPs not selected as tags) present in these samples, and (b) how this translates into study-wide power to detect an association under a specified disease model. We prefer these measures to comparisons based on differences in LD structure[7-9], haplotype diversity[10] or allele frequencies[11], as the question of immediate interest is the impact of any such differences on the power in the disease study.

Using only the genotype data collected in each HapMap panel (YRI, CEU, CHB and JPT), we selected tags until every SNP observed with ≥5% allele frequency in that panel was captured with a pairwise $r^2 \geq 0.8$ by at least one tag. By definition, these tags capture the "untyped" common SNPs (≥5%) at a maximum $r^2 \geq 0.8$ in that HapMap panel. Using this strategy, the mean maximum r^2 between the tags and the "untyped" SNPs was 0.93-0.96, and many "untyped" SNPs had a perfect proxy (maximum $r^2 = 1$) (**Figs. 1a-d**). The number of selected tags tracked inversely with the extent of LD in these samples.

Before considering transferability across population samples, it is crucial to measure the effect of transferability to a second, independent sample from the same population. Specifically, we expect to see statistical fluctuation in allele frequencies around the 5% threshold for tag SNP selection. For example, SNPs with an estimated allele frequency just below the 5% threshold will not be targeted during tag SNP selection (and may not be captured), but may well have an allele frequency above this threshold in a second sample. Conversely, SNPs with an estimated allele frequency just above the 5% threshold will be captured by a tag but may fall below the threshold in a second sample, and may therefore not be included in the assessment. Furthermore, there is fluctuation in the estimated r^2 for pairs of SNPs in independent samples of limited size: SNPs captured with $r^2 \geq 0.8$ by a tag in one sample may be captured with $r^2 < 0.8$ in a second sample (and vice versa) due to random fluctuations in the chromosomes chosen for each sample. These effects are a natural consequence of sampling variability and employing strict allele frequency and r^2 thresholds.

We characterized the extent of sampling variation in the HapMap reference panels by evaluating the coverage of common SNPs in independent samples drawn from the same population. The vast majority of the "untyped" common SNPs were still captured with a maximum $r^2 \geq 0.8$: 74% in HGDP-YRI, 89% in CEPH-EXT, 82% in HGDP-CHB and 79% in HGDP-JPT (**Figs. 1e-h** and **Table 2**). Moreover, nearly all SNPs were captured with a maximum $r^2 > 0.5$. The observed loss is interpreted as statistical fluctuation caused solely by drawing independent samples of limited size from the same underlying population, resulting in modest r^2 overestimation.

A small fraction of "untyped" SNPs, however, are not well captured: between 1 and 6% of "untyped" SNPs had a maximum $r^2 < 0.5$ in the second, independent sample from the same population (**Figs. 1e-h**). Upon closer inspection, these poorly captured SNPs typically had a lower allele frequency. All SNPs with a maximum $r^2 < 0.5$ in CEPH-EXT had a frequency below 5% in the HapMap CEU panel (a few were monomorphic), and were consequently missed, as no tags were explicitly picked to capture them (**Fig. 2**). This minor loss is due to the fluctuations in the allele frequency estimates, and not due to differences in LD structure. In contrast, only half of the SNPs with a maximum $r^2 < 0.5$ in HGDP-YRI are due to this allele frequency effect, with the other half showing a substantial drop in r^2 even though these SNPs were present at ≥5% in both samples. This was primarily limited to regions with low LD.

Having assessed the transferability within samples of the same population, we next examined transferability to samples with similar continental ancestry as the CEU and JPT HapMap panels, but sampled from different populations. Using the tags picked from HapMap CEU samples, we evaluated the coverage of common variants in self-described "White" individuals from Hawaii (MEC-W) and in individuals from the Botnia region of Finland (BOT). The performance of the tags was essentially unchanged, when compared with independent samples from the same underlying population as the HapMap panels (**Fig. 1i-j**). The mean maximum r^2 of "untyped" SNPs was 0.88-0.90 (**Table 2**). In both samples, only 3% of "untyped" variants were captured with a maximum $r^2 < 0.5$. We also evaluated the coverage in Japanese samples from Hawaii and Los Angeles, California (MEC-J) for tags picked from HapMap JPT samples. Performance was similar: 87% of "untyped" variants were captured with a maximum $r^2 \geq 0.8$ with a mean maximum r^2 of 0.91 (**Table 2**), and only 1% captured with a maximum $r^2 < 0.5$. Thus, there is very little additional loss in coverage beyond that observed previously within independent samples from the same population.

We next evaluated the performance of tags picked from the YRI HapMap samples in African-American samples from Los Angeles, California (MEC-AA), and from Chicago, Illinois (MAY). Of all "untyped" common variants, 62% were captured with a maximum $r^2 \geq 0.8$ with a mean maximum r^2 of 0.80-0.81 (**Table 2**). A comparatively larger (though still modest) fraction of SNPs (8-10%) was poorly captured with a maximum $r^2 < 0.5$ (**Fig. 1b**). This is not surprising: although African-Americans are estimated to have on average 80-85% African ancestry[12], a tagging strategy that takes into account the combined African and European ancestry of these samples would be expected to provide better coverage, as we indeed demonstrate below.

While these results are encouraging, we wanted to obtain a more direct estimate of statistical power in a disease study. Because of the correlated nature of dense SNP data and the number of statistical tests, it is not straightforward to estimate power directly from the r^2 distribution. Thus, we simulated case-control association studies for each non-HapMap population sample following a recently described procedure[3]. In these simulations, we nominated each common SNP with $\geq$5% allele frequency in that non-HapMap sample in turn to be "causal" (with modest effect) and generated a large number of simulated case-control panels. We evaluated power by performing the association tests (based on the tags selected from the HapMap samples) in these case-control panels and counting the number of panels in which we were able to detect an association at a gene-wide corrected *P* value of 0.01, averaged over all 25 genes. We report both the power to detect all common causal alleles (tags and "untyped" SNPs), and the power to detect only "untyped" common variants (that is, SNPs not selected as tags). We express power relative to that obtained by testing all common SNPs observed in the non-HapMap sample for association (as if we had genotyped directly all common SNPs in the case-control samples).

In independent samples from the same populations as the HapMap samples, power was 95-97% relative to the power obtained by testing all common SNPs in those samples, with slightly less power for the "untyped" common SNPs (**Table 3**). Power was essentially unchanged in MEC-W and BOT (using CEU tags) and in MEC-J (using JPT tags). Performance was somewhat lower in the African-American MEC-AA and MAY samples: relative power was 92% for tags picked in YRI. (For the sake of comparison, if we picked tags in CEU alone, relative power dropped to 79%.)

We attempted to improve the power for the African-American samples by picking tags from all four HapMap population samples combined (rather than picking tags from the YRI panel only). At an additional genotyping cost (22% more tags), this "cosmopolitan" tagging approach increased the relative power to 96% in both African-American samples with 89-90% relative power for the "untyped" common variation (**Table 3**). This result demonstrates that tags from the HapMap populations are able to provide good power in these samples. It is likely that tag SNP selection could be made more efficient by incorporating knowledge about the underlying local ancestry. Power in both African-American samples did not deteriorate when tags were picked from HapMap YRI and CEU panels (and not CHB and JPT), with a modest decrease in the number of tags. This is not unexpected: tags from CHB and JPT that are not redundant with tags from YRI and CEU include mostly SNPs that are unique to these two East-Asian populations.

For some population samples like Native Hawaiians (MEC-H) and Latinos (MEC-L) from the MEC, there is no obvious choice of HapMap reference panel from which to pick tags. Nevertheless, when we used CEU in our initial attempt, relative power was 94% (89% for "untyped" variants) in MEC-H, and power in MEC-L was only slightly worse (**Table 3**). The cosmopolitan tags improved relative power to 97-99% in both population samples, albeit at a greater genotyping cost.

Recently, we introduced specified multimarker (haplotype) tests as a means to improve upon pairwise tagging in terms of genotyping efficiency without sacrificing power[3]. In this approach, specific haplotypes act as effective surrogates for single tag SNPs. This keeps the multiple testing burden constant while decreasing the number of tags (but requiring greater genotyping quality and performance). When we used this "aggressive" tagging approach to capture all common SNPs with $r^2 \geq 0.8$, the genotyping burden was reduced by 15-23% compared with pairwise tagging. Power in simulated association studies in non-HapMap population samples remained essentially unchanged with this more efficient tagging approach. Hence, this multimarker tagging strategy is robust to transferability at least for the DNA samples tested here. An important implication of this result is that specified multimarker tests inferred from a dense reference panel (such as HapMap) can act as effective predictors for (some) untyped SNPs. We have recently demonstrated that this approach can provide a significant boost in the coverage of commercially available whole-genome products[13].

Discussion

Our work is broadly consistent with other assessments of tag transferability[14-24]. To our knowledge, this is the first systematic study to assess tag transferability using dense genotype data in all four HapMap population samples and many other samples. We estimated the effect of transferability on study-wide power and coverage of common variation, and we find that it is almost completely maintained in the non-HapMap samples. The minor loss in power that is observed is due largely to fluctuations around the allele frequency threshold (say, 5%) during tag SNP selection rather than true differences in LD between SNPs. These results indicate that tags selected from the HapMap samples can provide good power to study the role of common polymorphisms in complex traits in samples from many regions throughout the world.

Methods and Materials

DNA samples

We collected genotype data in the following DNA samples: 30 parent-offspring trios from the Yoruba people in Ibadan, Nigeria (YRI), 27 parent-offspring trios from Utah, USA, with northern and western European ancestry (from the Centre d'Etude du Polymorphisme Humain; CEU), 45 unrelated Han Chinese people from Beijing, China (CHB) and 44 unrelated Japanese people from Tokyo, Japan (JPT), also used in the International HapMap Project[2]; 25 unrelated individuals from Ibadan, Nigeria (HGDP-YRI), 40 unrelated Han Chinese from Beijing, China (HGDP-CHB), 31 unrelated Japanese from Tokyo, Japan (HGDP-JPT) from the Human Genome Diversity Project[25,26]; 62 trios from Utah, USA, with northern and western European ancestry from the CEPH collection (CEPH-EXT); 70 self-described African-American (MEC-AA), 69 self-described Native Hawaiian (MEC-H), 70 self-described Japanese (MEC-J), 70 self-described Latino (MEC-L) and 70 self-described White (MEC-W) samples from the Multiethnic Cohort study conducted in Hawaii and California (mainly Los Angeles), USA; 30 trios from Botnia, Finland (BOT); and 48 unrelated African-Americans from Chicago, Illinois, USA (MAY). These studies were approved by the Human Subject Institutional Review Boards at the respective institutions, and informed consent was obtained from all subjects.

SNP discovery

We performed exon resequencing in 95 cases of advanced breast cancer and 95 cases of advanced prostate cancer from the Multiethnic Cohort study. These are 19 samples from each of the five populations represented in the Multiethnic Cohort (see above) that do not overlap with the samples used to collect genotype data). Summary statistics are given in **Supplementary Table 1**.

SNP genotyping

A dense set of SNPs was selected for genotyping from two sources: (1) SNPs discovered by resequencing that were not in dbSNP (version 117) and that were located in exons or UTRs, and subsequently (2) SNPs from dbSNP (version 119) and Celera databases prioritizing "double-hit" and missense SNPs. Genotyping was performed to generate an initial map of roughly evenly spaced SNPs in the African-American MEC samples to classify regions according to their degree of LD, and to provide a guide for further genotyping. SNP density was preferentially increased in regions of low(er) LD as inferred from the initial map. In total, we attempted 3,302 SNP assays in 1,029 samples using the Sequenom MassArray and Illumina GoldenGate platforms (**Supplementary Table 2**). Concordance between the Sequenom and Illumina platforms was 98.2% (12,927 out of 13,170) for 863 markers typed in 16 identical samples. In the 15 population samples, on average, 84% (2,774) of the attempted assays passed quality control filters, defined as genotyping completeness >90%, no more than 1 concordance error, no more than 1 Mendel inheritance error, and $P > 0.001$ for the Hardy-Weinberg test (**Supplementary Table 3**). This resulted in a working set of 1,842 SNPs that passed QC in all 15 population samples, including 1,679 SNPs that are polymorphic in at least 1 population sample (1,473 SNPs with ≥5% frequency) (**Supplementary Table 4**). All genotype data were phased using the program PHASE 2.1.1 (ref. [27]) to produce phased chromosomes that were used in all analyses. For the purposes of estimating (high) r^2 values between SNPs, the impact of potential phasing errors is expected to be minimal[28].

Simulation of case-control association studies

We simulated case-control panels for every gene to evaluate study-wide power. We used a multiplicative genetic model in which we designated all common SNPs, one by one, to be "causal". For each causal SNP, we made case-control panels by sampling with replacement chromosomes from the phased data to give 1,000 cases and 1,000 controls (4,000 chromosomes in total). As a function of the allele frequency of the designated "causal" allele, we set the genotype relative risk to obtain a constant 95% power at a nominal P of 0.01 using a 2 x 2 chi-square test. We generated 75 replicate case-control panels per causal SNP, and all SNPs have an equal chance of being causal. We also generated 75,000 control-control (null) panels by sampling chromosomes at random from the phased data. These null panels have no causal SNP and were used to derive gene-wide significance thresholds (see below).

Tag SNP selection

We used the program Tagger to derive a set of tag SNPs from the HapMap reference panel such that each allele that satisfies the allele frequency threshold is captured at the given r^2 threshold either by a single tag (pairwise tagging)[29], or by a specified multimarker (haplotype) test ("aggressive" tagging)[3]. We noticed that the efficiency gain afforded by aggressive tagging was less than that observed in previous analyses of the HapMap-ENCODE data[2,3]; this can be explained by the fact that this study focuses on gene regions of ~100 kb size (compared to 500 kb ENCODE regions)[30]. We introduce a "cosmopolitan" tagging approach for picking tag SNPs that are maximally informative in multiple reference panels simultaneously. To this end, we implemented a greedy algorithm that maximizes, for every additional tag, the total number of alleles captured with the user-defined r^2 threshold, observed in the HapMap panels under consideration. This is similar in spirit to another approach that was recently described[31].

Power calculations

We evaluated power by performing the allelic tests (based on the selected tags) in the simulated null panels and the case-control panels. We derived significance thresholds from the null panels that correspond to a gene-wide corrected P of 0.01. We counted the fraction of case-control panels in which we observed a test statistic greater than the significance threshold. We report the average power, relative to that obtained by testing all common SNPs in that non-HapMap sample directly. Testing all common SNPs in the simulated case-control panels resulted in an absolute power of 82% in HGDP-YRI, 84% in CEPH-EXT and HGDP-CHB, and 85% in HGDP-JPT. This is substantially higher than the corresponding power in HapMap ENCODE data under the identical genetic model (60% in YRI and 68% in CEU and CHB+JPT)[3]. These differences correspond to a general reduction in the multiple testing burden. This is due to two effects. First, the genes in the current data set are on average 100 kb in size (much shorter than the 500 kb ENCODE regions). Second, the ascertainment of the current study was not as complete as that of the HapMap ENCODE project.

Figures and Tables

Table 1 - Population samples included in this study

Samples and origin	Abbreviation	Number of chromosomes in final data set
Utah, USA, European ancestry from CEPH (HapMap)	CEU	104
Utah, USA, European ancestry from CEPH	CEPH-EXT	248
Whites from Hawaii, USA, from Multiethnic Cohort	MEC-W	136
Bothnia, Finland	BOT	116
Han Chinese from Beijing, China (HapMap)	CHB	88
Han Chinese from Beijing, China, from Human Genome Diversity Project	HGDP-CHB	80
Japanese from Tokyo, Japan (HapMap)	JPT	88
Japanese from Tokyo, Japan, from Human Genome Diversity Project	HGDP-JPT	62
Japanese from Hawaii and Los Angeles, California, USA, from Multiethnic Cc	MEC-J	136
Yoruba from Ibadan, Nigeria (HapMap)	YRI	120
Yoruba from Ibadan, Nigeria, from Human Genome Diversity Project	HGDP-YRI	50
African-Americans from Los Angeles, California, USA, from Multiethnic Coho	MEC-AA	138
African-Americans from Chicago, Illinois, USA	MAY	96
Native Hawaiians from Hawaii, USA, from Multiethnic Cohort	MEC-H	138
Latinos from Los Angeles, California, USA, from Multiethnic Cohort	MEC-L	138

Table 2 - Coverage of common variation in the non-HapMap population by tags picked from the HapMap samples. Coverage is expressed as the mean maximum r2 for all SNPs and "untyped" SNPs with frequency ≥5% in the non-HapMap population samples.

Reference panel (HapMap samples)	Population sample	Mean maximum r^2 for all common SNPs	Mean maximum r^2 for "untyped" common SNPs
CEU	CEPH-EXT	0.95	0.91
CEU	MEC-W	0.93	0.88
CEU	BOT	0.94	0.90
CHB	HGDP-CHB	0.93	0.90
JPT	HGDP-JPT	0.93	0.89
JPT	MEC-J	0.94	0.91
YRI	HGDP-YRI	0.94	0.86
YRI	MEC-AA	0.92	0.81
YRI	MAY	0.92	0.80

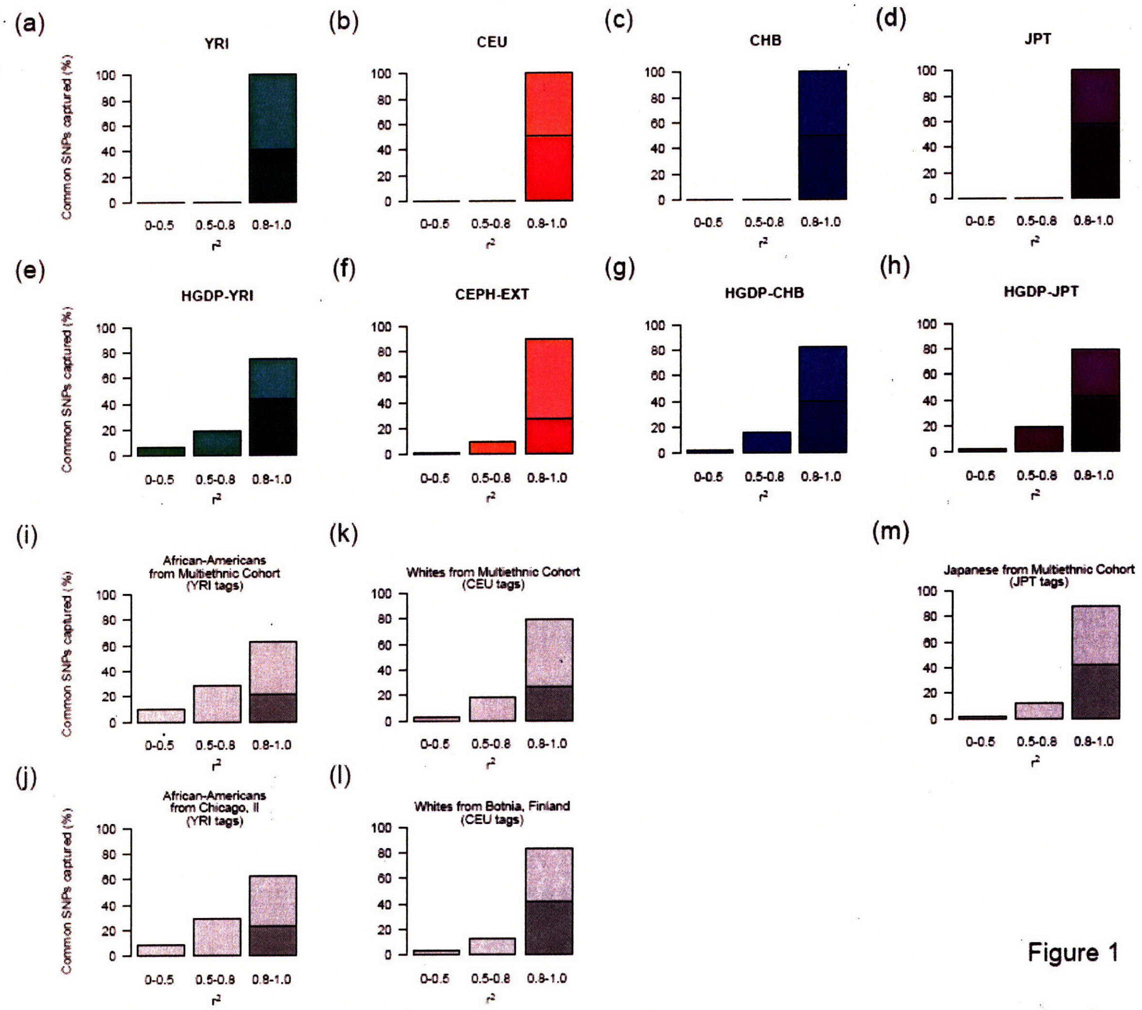


Figure 1

Figure 1. Performance of tags evaluated in multiple population samples, expressed as the percentages of common SNPs (excluding the tags) captured with the given maximum r2 (in three bins: 0 < r2 < 0.5, 0.5 < r2 < 0.8; 0.8 < r2 < 1.0). Tags are picked from the HapMap DNA samples so that every SNP with >5% allele frequency is captured by a tag with pairwise r2 > 0.8. We show performance in each HapMap sample: (a) YRI, (b) CEU, (c) CHB, (d) JPT; in additional samples from the same populations as the HapMap samples: (e) HGDP-YRI (using tags from YRI), (f) CEPH-EXT (using tags from CEU), (g) HGDP-CHB (using tags from CHB), (h) HGDP-JPT (using tags from JPT), respectively; and in other samples: (i) MEC-AA (using tags from YRI), (j) MAY (using tags from YRI), (k) MEC-W (using tags from CEU), (l) BOT (using tags from CEU), (m) MEC-J (using tags from JPT). Darker shade is used for SNPs captured by a perfect proxy (maximum r2 = 1).

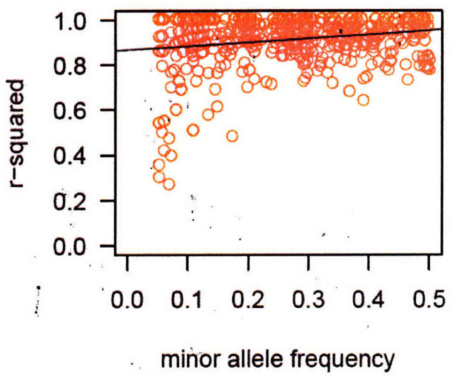


Figure 2. The effect of allele frequency on the maximum r2 between the tag SNPs (picked from the HapMap CEU samples such that SNPs ≥5% are captured with pairwise r2 ≥ 0.8) and the "untyped" common SNPs in the additional CEPH samples (CEPH-EXT). SNPs that are captured with lower maximum r2 tend to have a lower allele frequency. Plotted is the linear regression line to fit this relationship (P < 10-7).

Table 3: Relative power in simulated case-control association studies in non-HapMap populations. Tags were picked from the reference panel so as to capture all observed with ≥5% frequency in that panel with a pairwise r2 ≥ 0.8. The relative power is the power to detect causal alleles (SNPs with ≥5% frequency in the non-HapMap population sample) in comparison to the observed power when all causal al are tested directly (i.e. no tagging). Power is given for all common SNPs as well as the subset of common SNPs that were not picked as tags ("untyped"). Cosmopolitan tagging refers to picking tags to capture common variation in all four HapMap populations simultaneously.

Reference panel (HapMap samples)	Number of picked tags	Case-control panel	Relative power for all common causal alleles (%)	Relative power for "untyped" common causal alleles (%)
CEU	470	CEPH-EXT	97	95
CEU	470	MEC-W	96	91
CEU	470	BOT	96	89
CHB	388	HGDP-CHB	96	92
JPT	415	HGDP-JPT	96	92
JPT	415	MEC-J	96	92
YRI	724	HGDP-YRI	95	87
YRI	724	MEC-AA	92	81
YRI	724	MAY	92	81
Cosmopolitan	885	MEC-AA	96	90
Cosmopolitan	885	MAY	96	89
CEU	470	MEC-H	94	89
Cosmopolitan	885	MEC-H	99	96
CEU	470	MEC-L	92	83
Cosmopolitan	885	MEC-L	97	94

Supplementary Table 1: Summary of SNP discovery through resequencing

Locus	Chr	# coding exons	# coding exons passing QC	# Bases sequenced	# Bases passing QC	# SNPs discovered[1]	# SNPs in dbSNP[2]	# Novel SNPs attempted for validation[3]	# Novel validated SNPs [4]	# Novel missense SNPs [5]	# Novel missense SNPs above 1% (in resequencing panel)[6]	# Novel missense SNPs above 1% (Multi-Ethnic Cohort panel)[7]
ACVR1	2	9	8	1530	1266	25	8	16	15	0	0	0
FSHR	2	10	10	2088	1872	21	3	18	9	0	0	0
INHA	2	2	1	1101	833	9	2	7	1	1	1	0
INHBB	2	2	1	1224	776	7	1	6	3	0	0	0
LHCGR	2	11	10	2100	1907	42	12	27	9	1	0	0
SRD5A2	2	5	5	764	764	14	4	10	8	1	1	1
GNRHR	4	3	2	987	767	7	2	5	3	1	0	0
FST[8]	5	--	--	--	--	--	--	--	--	--	--	--
PRLR	5	8	8	1869	1869	26	0	26	15	0	0	0
SRD5A1	5	5	3	780	426	20	12	7	5	1	0	0
CGA	6	3	3	351	351	4	2	2	0	0	0	0
ESR1	6	8	8	1788	1780	42	15	25	21	4	0	0
IGF2R	6	48	46	7476	7068	100	28	67	60	18	5	3
PRL	6	5	5	684	660	20	4	15	11	0	0	0
INHBA	7	2	1	1281	1078	2	0	2	1	1	1	1
GNRH1	8	3	3	279	279	6	4	2	2	0	0	0
CYP17	10	8	7	1527	1297	31	8	17	12	0	0	0
FSHB	11	2	2	390	390	7	4	3	2	0	0	0
PGR	11	8	7	2802	2275	40	19	18	12	1	1	0
IGF1	12	4	3	462	280	9	4	5	4	0	0	0
ESR2	14	8	8	1593	1593	19	5	13	10	2	1	0
CYP11A1	15	9	9	1566	1566	23	3	20	10	2	0	0
CYP19	15	9	8	1512	1463	41	12	23	12	1	1	1
HSD17B2	16	5	4	1164	951	11	0	11	10	3	3	1
SHBG	17	8	8	1209	1209	16	5	10	5	0	0	0
Total		185	170	36527	32720	542	157	355	240	37	14	7

validation rate=
67.61%

[1] Number of SNPs identified by resequencing
[2] Number of SNPs already described in dbSNP (build 117)
[3] Number of novel SNPs attempted for validation (successful genotyping assay designed)
[4] Number of novel validated SNPs that passed the QC thresholds described in the methods
[5] Number of novel missense SNPs validated in the resequencing panel (n=190)
[6] Number of novel missense SNPs above 1% in the resequencing panel (n=190)
[7] Number of novel missense SNPs above 1% in the Multi-Ethnic Cohort panel (n=349)
[8] The FST gene was not resequenced as part of this study

Supplementary table 2 is 40 pages long and appears only in the online supplement to the original publication.

Supplementary Table 3: Genotyping summary of the final data set

				QC passing SNPs[1]		Polymorphic SNPs[2]	
Locus	Chr	Position (hg16)	Size (kb)	Number	Average spacing (kb)	Number	Average spacing (kb)
ACVR1	2	158784754-158916981	132	84	1.6	75	1.8
FSHR	2	49157637-49375736	218	141	1.5	131	1.7
INHA	2	220611238-220652419	41	13	3.2	9	4.6
INHBB	2	121171479-121214066	43	13	3.3	11	3.9
LHCGR	2	48878904-48977278	98	116	0.8	108	0.9
SRD5A2	2	31715642-31793965	78	30	2.6	27	2.9
GNRHR	4	68594632-68648684	54	46	1.2	44	1.2
FST	5	52776937-52806789	30	26	1.1	24	1.2
PRLR	5	35084938-35301367	216	183	1.2	166	1.3
SRD5A1	5	6666565-6730425	64	52	1.2	50	1.3
CGA	6	87780870-87819426	39	39	1.0	37	1.0
ESR1	6	152136584-152472230	336	246	1.4	238	1.4
IGF2R	6	160276511-160448209	172	163	1.1	148	1.2
PRL	6	22375954-22438440	62	67	0.9	60	1.0
INHBA	7	41449576-41512218	63	44	1.4	42	1.5
GNRH1	8	25287836-25325323	37	12	3.1	12	3.1
CYP17	10	104234154-104360966	127	50	2.5	41	3.1
FSHB	11	30198703-30228677	30	10	3.0	9	3.3
PGR	11	100437852-100555849	118	106	1.1	96	1.2
IGF1	12	101244363-101400231	156	93	1.7	75	2.1
ESR2	14	62680052-62769843	90	36	2.5	33	2.7
CYP11A1	15	72335354-72402356	67	21	3.2	18	3.7
CYP19	15	49188301-49381873	194	95	2.0	91	2.1
HSD17B2	16	81813649-81929978	116	137	0.8	118	1.0
SHBG	17	7711853-7754309	42	19	2.2	16	2.7
Overall			2,623	1,842	1.4	1,679	1.6

[1] These SNPs pass QC in all 15 population samples, including monomorphic SNPs
[2] These SNPs are QC passing and polymorphic in at least 1 of the population samples

Supplementary Table 4: Genotyping summary by population

	HapMap samples				additional samples as HapMap				Multiethnic Cohort						
	YRI	CEU	CHB	JPT	HGDP-YRI	CEPH-EXT	HGDP-CHB	HGDP-JPT	MEC-AA	MEC-H	MEC-J	MEC-L	MEC-W	MAY	BOT
Total SNPs passing QC [1]	2811	2843	2827	2815	2625	2901	2744	2712	2656	2661	2656	2658	2665	3005	3029
Monomorphic SNP	783	841	1051	1050	724	799	996	1004	530	708	904	615	657	632	972
Polymorphic SNP	2028	2002	1776	1765	1901	2102	1748	1708	2126	1953	1752	2043	2008	2373	2057
Total SNPs failing QC [2]	480	452	475	487	677	372	558	590	646	641	646	644	637	297	273
Genotyping <90%	469	440	462	475	674	353	555	586	615	615	615	615	615	272	271
HWE P<0.001	11	11	12	11	3	19	3	4	23	19	23	22	14	25	2
>1 Mendel error	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
>1 concordance error	0	1	1	1	0	0	0	0	8	7	8	7	8	0	0

[1] Total number of SNPs that passed quality control (QC) parameters defined in methods

[2] Total number of SNPs that failed quality control (QC) parameters defined in methods

References

1. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-96 (2003).
2. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
3. de Bakker, P.I.W. et al. Efficiency and power in genetic association studies. *Nat Genet* **37**, 1217-1223 (2005).
4. Kolonel, L.N. et al. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol* **151**, 346-57 (2000).
5. Kolonel, L.N., Altshuler, D. & Henderson, B.E. The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nat Rev Cancer* **4**, 519-27 (2004).
6. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **34**, D173-80 (2006).
7. Shifman, S., Kuypers, J., Kokoris, M., Yakir, B. & Darvasi, A. Linkage disequilibrium patterns of the human genome across populations. *Hum Mol Genet* **12**, 771-6 (2003).
8. Sawyer, S.L. et al. Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* **13**, 677-686 (2005).
9. Evans, D.M. & Cardon, L.R. A Comparison of Linkage Disequilibrium Patterns and Estimated Population Recombination Rates across Multiple Populations. *Am J Hum Genet* **76**, 681-7 (2005).
10. Beaty, T.H. et al. Haplotype diversity in 11 candidate genes across four populations. *Genetics* **171**, 259-67 (2005).
11. Willer, C.J. et al. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet Epidemiol* **30**, 180-90 (2006).
12. Parra, E.J. et al. Estimating African American admixture proportions by use of population-specific alleles. *Am J Hum Genet* **63**, 1839-51 (1998).
13. Pe'er, I. et al. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* **38**, 663-667 (2006).
14. Weale, M.E. et al. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* **73**, 551-65 (2003).
15. Nejentsev, S. et al. Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. *Hum Mol Genet* **13**, 1633-9 (2004).
16. Mueller, J.C. et al. Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet* **76**, 387-98 (2005).
17. Ahmadi, K.R. et al. A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat Genet* **37**, 84-9 (2005).
18. Ramirez-Soriano, A. et al. Haplotype tagging efficiency in worldwide populations in CTLA4 gene. *Genes Immun* **6**, 646-57 (2005).
19. Ribas, G. et al. Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. *Hum Genet* **118**, 669-79 (2006).
20. Stankovich, J. et al. On the utility of data from the International HapMap Project for Australian association studies. *Hum Genet*, 1-3 (2006).

21. Huang, W. et al. Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proc Natl Acad Sci U S A* **103**, 1418-21 (2006).
22. Gonzalez-Neira, A. et al. The portability of tagSNPs across populations: A worldwide survey. *Genome Res* (2006).
23. Montpetit, A. et al. An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet* **2**, e27 (2006).
24. Smith, E.M. et al. Comparison of linkage disequilibrium patterns between the HapMap CEPH samples and a family-based cohort of Northern European descent. *Genomics* (2006).
25. Cann, H.M. et al. A human genome diversity cell line panel. *Science* **296**, 261-2 (2002).
26. Rosenberg, N.A. et al. Genetic structure of human populations. *Science* **298**, 2381-5 (2002).
27. Stephens, M. & Donnelly, P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* **73**, 1162-9 (2003).
28. Marchini, J. et al. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* **78**, 437-50 (2006).
29. Carlson, C.S. et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **74**, 106-20 (2004).
30. Pe'er, I. et al. Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am J Hum Genet* **78**, 588-603 (2006).
31. Howie, B.N., Carlson, C.S., Rieder, M.J. & Nickerson, D.A. Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. *Hum Genet* (2006).

Chapter 4: Evaluating and Improving Power in Whole Genome Association Studies using Fixed Marker Sets

Material in this chapter is reprinted by permission from Macmillan Publishers Ltd from:

Pe'er I, de Bakker PI, Maller J, ***Yelensky R***, Altshuler D, Daly MJ.
Evaluating and improving power in whole-genome association studies using fixed marker sets.
Nat Genet. 2006 Jun;38(6):663-7. Epub 2006 May 21.
PMID: 16715096 [PubMed - indexed for MEDLINE], Copyright 2006

in accordance with the following license (and Nature Group author rights):
Licensee: Roman Yelensky, License Date: Jun 03, 2008, License Number: 1961471199328
Publication: Nature Genetics, Title: Evaluating and improving power in whole-genome association studies using fixed marker sets, Type Of Use: Thesis / Dissertation

Introduction

Whole genome association studies are a comprehensive approach to testing the hypothesis that common alleles contribute to heritable phenotype variation[1-3]. While neither resequencing every base nor typing all 11 million currently known polymorphic sites in the human genome[4] is yet technically feasible, a practical path to genomewide association studies has been opened by the introduction of genomewide SNP arrays [5,6] that type 100,000 to 500,000 SNPs per sample.

Such association studies benefit greatly from linkage disequilibrium [1,2,7], the correlation between the SNPs on each array and other nearby (untyped) putatively causal alleles[8]. With the completion of the Phase II of HapMap[9], it becomes possible to address two important questions with respect to the use of these arrays. First, to what extent do the fixed set of SNPs on these arrays capture the information about common variation in the human genome?[10] Second, is it possible to devise analytical strategies that make use of HapMap data to increase the chance to discover a true association?

Results

We evaluate three whole-genome products: the 100K and 500K GeneChip Mapping Sets of Affymetrix[6], and the Sentrix HumanHap300 BeadChip by Illumina[5] (products which contain 116,204, 504,152 and 317,503 SNPs, respectively). Figures for the GeneChip 500K and HumanHap300 products are based on lists of SNPs included on the product (rather than established genotyping performance in laboratories around the world), and thus should be considered preliminary, best-case scenarios. Updated information about evaluations of these and subsequent products are and will be available online.

SNPs included on the Affymetrix products have been pre-selected primarily on the basis of technical quality and thus represent a quasi-random set of SNPs. In contrast, SNPs on the Illumina product were selected using a pairwise-correlation-based algorithm applied to genotype data of HapMap Phase I SNPs in the CEU panel (CEPH-collected samples of Utah residents with European ancestry)[11].

Evaluation of each marker set would ideally involve measuring the extent to which they are correlated to every putative causal common allele along the genome. While complete polymorphism data does not yet exist to support such an analysis, all three array SNP sets have been typed in the HapMap reference samples of 270 individuals from four population samples[9]. These panels therefore allow, in principle, evaluation of correlation in two datasets: the ENCODE data of 10 regions spanning 5Mb, with essentially complete ascertainment for alleles with frequency ≥5% [12,13], and the genomewide Phase II HapMap, which includes roughly 3.9 million SNPs successfully typed to date. We therefore evaluate the GeneChip 100K and 500K arrays vis-à-vis ENCODE, while evaluating all three arrays on the Phase II HapMap data.

A full exploration of the utility of a SNP set involves estimating the power to detect association under many study-design and disease scenarios[14]. Yet, a simpler, study-independent measure of utility is the square of the correlation coefficient (r^2) between any observed marker and a putative causal allele[15]. This metric is interpretable as the expected drop in non-centrality of an association test statistic under specified conditions[16], and has become one standard for evaluating performance of marker sets[17-20].

Figure 1 shows the correlation between common SNPs in the Phase II data (i.e., SNPs with minor allele frequency (MAF) ≥5%) and markers on the whole-genome arrays (see Figure 1 legend for details). The fraction of SNPs captured is a function of the threshold correlation coefficient required for tag SNP selection. For example, in the CEU panel, 45% of all common Phase II SNPs are captured by the GeneChip 500K array at r^2 of 1 (i.e. no loss of power compared to testing the putative causal SNP directly); while 62% of common SNPs are captured at r^2 of 0.8 and 80% with $r^2 \geq 0.5$ (i.e., highly significant correlations to untyped alleles but with modest loss of power in association settings). As expected, SNPs on the array capture a smaller proportion of variants in the most genetically diverse panel, YRI when compared to CEU, CHB+JPT panels, in which the fractions of SNPs captured are higher and similar to one another.

Figure 2 examines correlations of SNPs in the more fully ascertained ENCODE regions for the GeneChip arrays. This cross-validates the results of common Phase II SNPs and allows examination of a substantial, yet incomplete set of SNPs with frequency 1-5%. The representation of the latter set of SNPs is limited and biased by the scope of SNP discovery efforts, which tend to miss the rarer alleles. The examined set of SNPs therefore reveals only an upper bound on the ability of the arrays to capture low frequency alleles, which is nevertheless much poorer than corresponding ability for common ones[21]. This highlights the focus of the array content at common variants, where association studies are most powerful to detect (subtle) genetic effects[22]. Comprehensive scans for rare causal alleles will require other sets of markers, more involved analysis methods[23,24], and where possible, complete resequencing.

Even though a considerable fraction of common variants are captured by the current generation of genomewide arrays, there exists a substantial component of common variation not highly correlated to a SNP on each array. We set out to analytically improve ability to capture common variants using only the SNPs on these arrays and knowledge of LD in available HapMap data. Here we describe an approach in which HapMap data is used to detect correlations between specific combinations of alleles for SNPs on each array (called multi-marker predictors[20]) and a putatively causal allele previously uncaptured. We and others[17-20,25] have elsewhere introduced this concept in the context of tag SNP selection, avoiding the typing of certain SNPs to improve typing efficiency while maintaining study power. In the context of fixed-content SNP genotyping products we propose to use specific multi-marker predictors of untyped SNPs (inferred from the

HapMap) as tests of association, thereby increasing study power without performing additional genotyping.

We observe that multi-marker predictors based on combinations of alleles of 2 or 3 SNPs can capture (at $r^2 \geq 0.8$) an additional 9-25% SNPs in ENCODE or HapMap Phase II (Figure 3). Notably, using these specific tests (fully listed online), the HumanHap300 and GeneChip 500K arrays gain the ability to capture 80-86% of common alleles in the CEU population with this high level of correlation. These tests also facilitate pooling association results from studies that used different arrays, through combined predictions of the same SNPs. This gain in power is achieved without any additional genotyping and thus permits more comprehensive association studies with current products, at no extra cost.

A possible concern is the potential of overfitting based on HapMap relationships involving limited sample sizes (120 chromosomes for CEU and YRI, 180 chromosomes for CHB+JPT). Mathematically, however, the chance of a highly correlated ($r^2 \geq 0.8$) common variant in this sample size is much smaller ($<10^{-12}$) than the space of predictors searched for each SNP. We verified this empirically by developing multi-marker predictors to unlinked SNPs: we never observed spurious correlations of $r^2 > 0.35$ in HapMap data. While for rare alleles overfitting is indeed an issue using the HapMap sample sizes, we are confident that relationships at thresholds such as $r^2 > 0.5$ involving common SNPs are robust and reliable.

These results suggest that in studies where direct typing of a common causal SNP would be successful, use of one of these genotyping arrays would most often be successful as well[20]. Yet, focusing on scenarios where power is limited, the benefits of capturing more variants by our method needs to be appraised versus the statistical cost of performing additional hypothesis tests. This is because addition of statistical tests could, in principle, lead to a reduction in power via the requirement of increased statistical significance thresholds to maintain constant type I error rates (or, conversely, allowing substantially more false positives if statistical thresholds are unchanged).

This tradeoff is of particular relevance to multimarker predictors, as they capture on average fewer untyped SNPs than do single SNPs. That is, we observe that statistical tests based on the genotype of a SNP on the array have more proxies on average in HapMap Phase II than do statistical tests based on two and three marker haplotype predictors: 3.85 vs 1.55 putative causal alleles captured, respectively, on the GeneChip 500K array in the CEU panel. At the extreme, testing all observed allele combinations[26], rather than only the SNPs and specified multimarker predictors might not pay off, as the dramatic increase in degrees of freedom[18,25] results in only a tiny increase in the fraction (3% in CEU) of common SNPs captured [9]. Adding many tests while increasing by a small amount information capture can result in a loss of power for association to common alleles[18,20].

We consider a Bayesian strategy to tests all alleles without suffering from an increased burden of multiple testing. The standard, frequentist strategy for genomewide association studies[8] assigns a 1- or 2-degree of freedom score to each variant tested and searches for *p*-values deemed significant. While *p*-values speak to the extent of surprise by observed data under the null (i.e., no association) hypothesis, external information may be quite relevant to the alternative hypothesis (that the tested, or a nearby correlated, variant is truly causal). Intuitively, not all tests are created equal – those hypothesis tests which capture the genotypic variance at many SNP sites, or those which correspond to known functional alterations, may rightly be considered more likely *a priori* to be true positives than those hypothesis tests that capture only a single variant site (of unknown functional significance). This highly relevant information is not customarily considered *a priori* in a formal fashion (although often discussed in a post-hoc manner). Analysis of the HapMap data makes it possible to incorporate such information up front in association analysis. Specifically, we define prior probabilities based on the identities and number of putative causal alleles captured by each allelic hypothesis test. Having assigned a prior probability for each of being causal, we can evaluate the *a posteriori* likelihood of association given the data (see Methods).

We demonstrate this framework using one objective, simple, and universal hypothesis used in simulation studies[20,26] - namely that each common SNP in the genome is equally likely to be causal. The *a priori* likelihood of association to each marker on the array is therefore proportional to the number of SNPs it captures. The number of variant sites captured by each hypothesis test is highly variable, as even very large clusters of correlated SNPs may be represented by a single SNP, while other SNPs capture only themselves. We show by simulated association studies that incorporation of such Bayesian priors (see Methods) modestly but consistently (and statistically significantly) improves power to detect association as compared to a frequentist framework. For example, association testing to 100 SNPs, chosen either randomly or by LD tagging, is improved by 4% by this approach (Supplemental Figure 1). Moreover, the value of this approach will only increase as genomic annotation improves the estimate of the prior probability of each variant site in the genome being causal. Moreover, individual investigators can tailor analysis based on their own views of how to weight SNPs that are coding[27], associated with variation in gene expression, under a compelling linkage peak[28], or in genes whose function is tied to a particular pathway.

Discussion

The simultaneous emergence of genomewide genotyping arrays and comprehensive, deeply ascertained SNP data from HapMap provides for the first time a toolkit to evaluate association between common genetic variation and disease throughout the genome. We find that current products capture a sizeable portion of genomic variation, and describe methods to utilize the HapMap data for testing additional non-array SNPs *in silico* without further genotyping. Finally, we have developed a framework to prioritize the tested SNPs based on external information provided by HapMap and, potentially, additional genomic annotation. Such methods should help enable systematic and more powerful evaluation of the contribution of common alleles to complex phenotypes.

Methods and Materials

Data sets

We used the phased ENCODE data from HapMap (release 16c.1. We also used genotype data from Phase II HapMap (release 19), and merged these with the genotype data generated by the GeneChip 500K array, ported the data to NCBI build 35 (UCSC hg17), and subsequently phased the final data using the EM algorithm [29].

Choosing multi-marker predictors

For every array product, we have specified a set of haplotype tests based on HapMap using Tagger[20]. For every SNP that is not typed on the array, we aim to find the allelic test (predictor) with the highest r^2 to it, exploiting the knowledge which SNPs are present on the array. The predictors are identified by performing an aggressive search among combinations of 2 or 3 SNPs (on the array), evaluating the r^2 between the generated haplotypes and the allele we want to capture. While many of the untyped SNPs are captured by high pairwise correlation to a SNP on the array, a substantial fraction of the (common) SNPs is not. We have made the multi-marker predictors for all three arrays evaluated here available on our website.

Simulating case-control panels

Our simulation framework follows a recently published protocol[20]. Briefly, the phased ENCODE chromosomes (n=120 from unrelated individuals in CEU) were resampled to create 1000 cases and 1000 controls (4000 chromosomes in total). For controls, resampling was uniform. For cases, we designated one SNP to be causal. For this causal SNP, we calculated an effect size (and corresponding allele frequency in the cases) such that if it were to be the only SNP tested, power would be 95% to detect it at a nominal P value of 0.01. In terms of relative risk, the simulated effect size was therefore larger for rare alleles (see Supplemental Fig. 2). We created 250 case-control panels for each causal SNP, where we allowed, at random, either allele of a given SNP to be causal. We repeated this for all common SNPs in a region, and for all ten ENCODE regions separately (a total of nearly 10,000 SNPs). We also generated 250,000 null panels (without a causal SNP) for evaluation of the null distribution.

Power calculations

Power is defined as the fraction of the simulated case-control panels in which the test statistic exceeds the significance threshold (when an association can be declared), averaged over all ten ENCODE regions. We use the maximum of the 2x2 chi-square comparison over all allelic tests (single-marker tests and optionally the specified multi-marker tests) as the region-wide test statistic. The significance threshold is derived by performing the same allelic tests from the null panels (to achieve a region-wide corrected P value of 0.01). The absolute power to detect association at $P < 0.01$ after multiple testing correction is 68%, if all common SNPs are evaluated. Power remains >90% of this figure when the best tags (with most proxies) are selected at a density of 1 per 5 kb, if these tests are given uniform weights[20]. Introducing weights based on LD improves power.

Derivation of weights for allelic tests

Suppose the set of m putative causal alleles is $A = \{a_1, \ldots, a_m\}$. Denote by $C[a_i, I]$ the count of the allele a_i in a set I of individuals. Let I_1, I_0 be sets of cases and controls, of sizes N_1, N_0, respectively. Define the normalized difference statistic

$$Z(a_i) = \frac{\frac{C[a_i, I_1]}{N_1} - \frac{C[a_i, I_0]}{N_0}}{\sqrt{\left(C[a_i, I_1] + C[a_i, I_0]\right)\left(1 - \frac{C[a_i, I_1] + C[a_i, I_0]}{N_1 + N_0}\right)}}$$

Suppose further that the set of n tests (single- or multi-marker predictors) used to capture these alleles is $T = \{t_1, \ldots, t_n\}$, and extend the definition of the count operator [] and the statistic Z to these tests.

The null hypothesis is simple: $Z(t_1), \ldots, Z(t_n)$ are all standard normal variables (a.k.a. z-scores). In contrast, the alternative hypothesis is complex: it states that a causal allele is chosen out of A according to some prior distribution $D : A \rightarrow [0,1]$ (where $D(a_i)$ denotes the probability of a_i to be chosen as causal) and given that choice, all tests that are correlated with a_c are normally distributed with means greater than zero. More specifically, let μ_c be the effect size for the causal allele a_c, represented in terms of mean offset of $Z(a_c)$ from the origin. For each test t_j, let $r_{c,j}$ denote its correlation coefficient to a_c. Hence, if a_c is causal, $Z(t_j)$ is normally distributed with mean $\mu_c \cdot r_{c,j}$.

In this study we denote the normal p.d.f. and c.d.f. by ϕ and Φ, respectively. We use the simulation assumption[20] that $\mu_c = \Phi(0.95) + \Phi(0.99) \approx 3.97$. We use Haploview [30] to compute the matrix $R = \left[r_{c,j} \right]_{m \times n}$ of correlation coefficients between all tests and all alleles, and transform it into a matrix $W = \left[w_{i,j} \right]_{m \times n}$ where $w_{i,j}$ is the probability that given a_i is causal, it will be detected by t_j, i.e., the top scoring test for a_i is t_j, and it is above the null signal. Formally, if $r_{i,j} = 0$, we set $w_{i,j}$ to zero as well. Otherwise, to compute $w_{i,j}$ we integrate over the real signal, $Z(a_i)$, given which we can write down the score distributions of the current test t_j and the scores in needs to exceed: the null signal, as well as any true signal by some other test $t_{j'}$ correlated to a_i. We approximate such tests as being dependant through a_i only. We can thus express all relevant probabilities as functions of $Z(a_i)$, as follows:

$$w_{i,j} \approx \int_{z=-\infty}^{\infty} P\left(Z(t_j) > null \mid Z(a_i) = z\right) \cdot \left(\prod_{j' \mid r_{i,j'} \neq 0,\, j \neq j'} P\left(Z(t_j) > Z(t_{j'}) \mid Z(a_i) = z\right) \right) \cdot \phi(z - \mu_i)dz \approx$$

$$= \int_{z=-\infty}^{\infty} P_{null}\left(r_{i,j} \cdot z\right) \cdot \left(\prod_{j' \mid r_{i,j'} \neq 0,\, j \neq j'} \Phi\left(\frac{z(r_j - r_{j'})}{\sqrt{2 - r_j^2 - r_{j'}^2}} \right) \right) \cdot \phi(z - \mu_i)dz$$

where *null* represents the region maximum score in a null panel and $P_{null}(z)$ is the empirically derived c.d.f. of this maximum. We can now write the likelihood ratio test for a dataset with the maximum-scoring test t_j achieving an observed score z_j :

$$\frac{\Pr\left[Z(t_j) = z_j \mid H_1\right]}{\Pr\left[Z(t_j) = z_j \mid H_0\right]} = \left[\sum_{i=1}^{m} \left(D(a_i) \cdot w_{i,j} \right) \right] \cdot p - value(z_j)$$

We thus compute a weight factor $W_j = \sum_{i=1}^{m} \left(D(a_i) \cdot w_{i,j} \right)$ for each test t_j employed, and use that to prioritize all p-values.

Figures

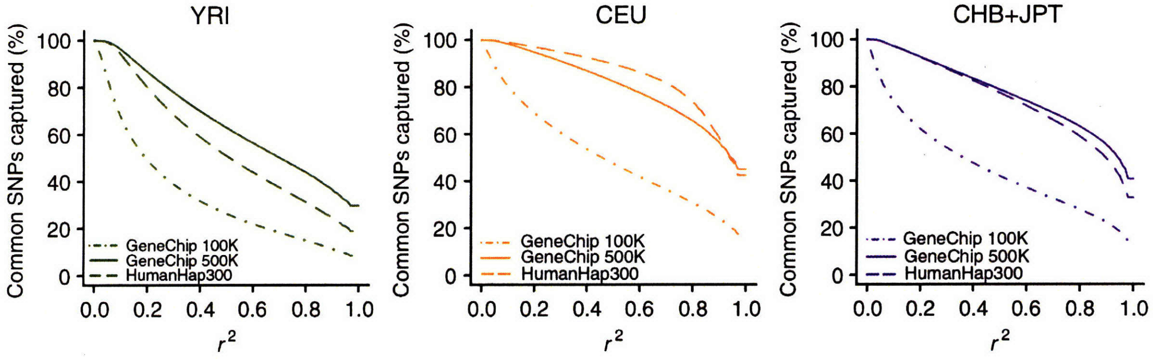


Figure 1: Fraction of common (MAF ≥5%) Phase II HapMap SNPs (y axis) captured by array SNPs as a function of the r2 cutoff (x-axis). Data is presented for the GeneChip 100K (dash-dot line), GeneChip 500K (solid line) and HumanHap300 (dashed line) arrays, for each of the three HapMap analysis panels: Yoruba people ascertained in Ibadan, Nigeria (YRI; left; green); The CEPH collected samples of European ancestry, ascertained in Utah (CEU; middle; orange); and Han Chinese samples from Beijing with Japanese samples from Tokyo (CHB+JPT; right; purple).

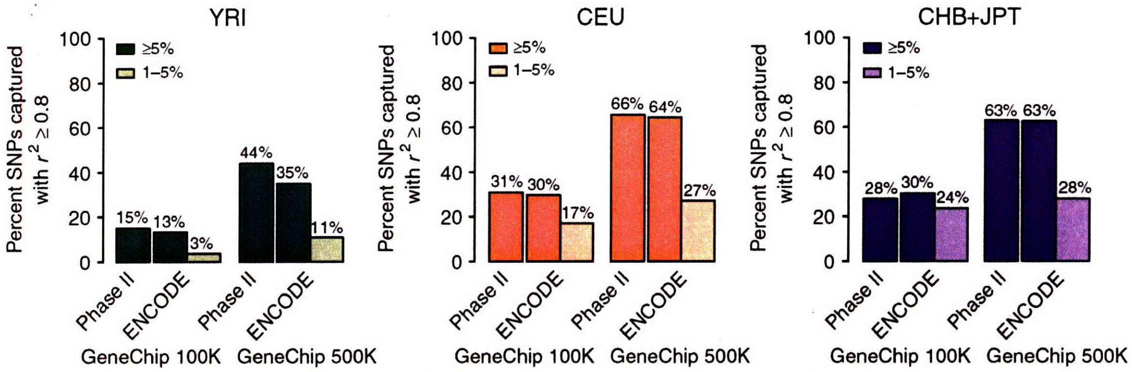


Figure 2: Fraction of SNPs (y axis) captured by SNPs on GeneChip 100K and 500K arrays as at r2 ≥ 0.8 in the three HapMap panels: YRI (left) CEU (middle) and CHB+JPT (right). Data is presented for common SNPs (dark bars) as observed in HapMap Phase II and ENCODE and less common (MAF 1-5%) SNPs (light bars) as observed in ENCODE. As ENCODE data do not fully represent SNPs from the latter category, but rather includes only a partial set of such SNPs that happened to have been discovered (and tend to be more common), results presented here should be considered as upper bounds for the ability to capture the complete set of alleles of frequencies 1-5%.

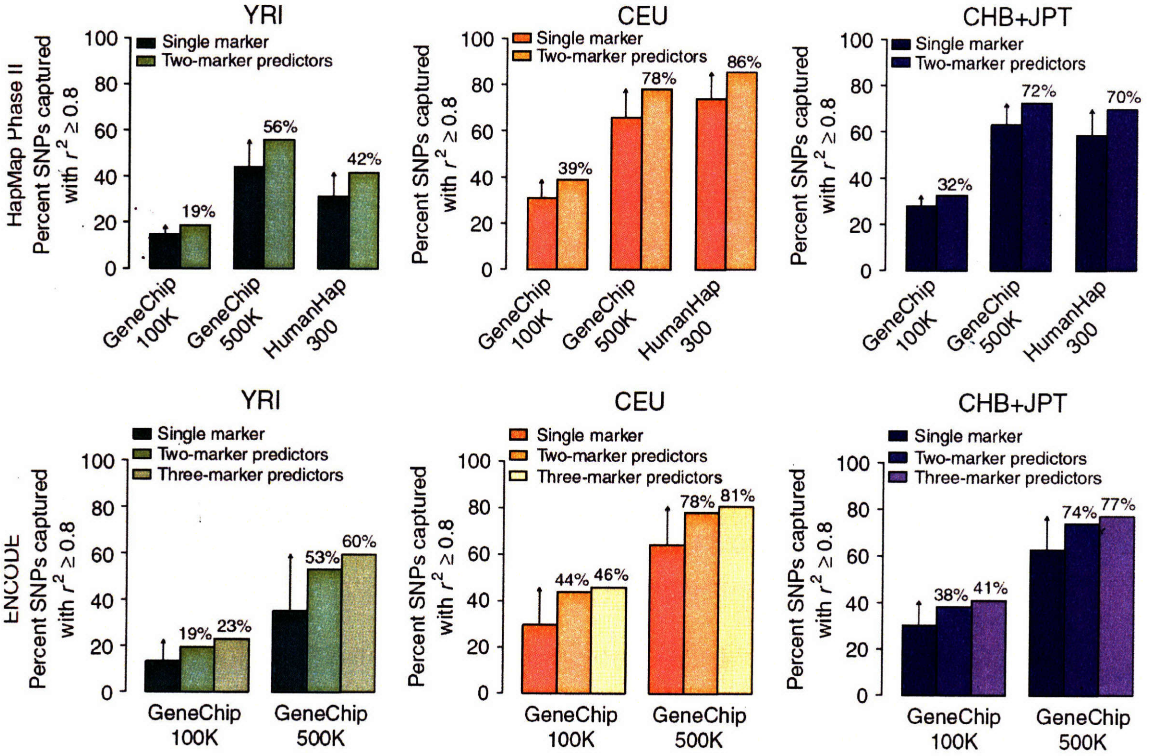
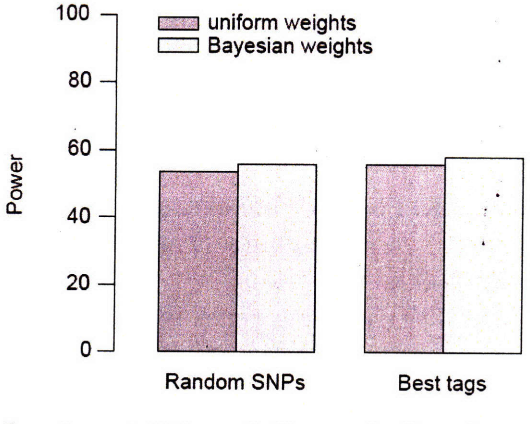


Figure 3:Fraction of common SNPs (y axis) captured by single array SNPs vs. multi marker predictors in three HapMap panels (YRI – left; CEU – middle; CHB+JPT – right). Data is presented for HapMap Phase II (top) as well as ENCODE (bottom). For Phase II data, we evaluated the GeneChip 100K, GeneChip 500K and HumanHap300 arrays with 2-marker predictors. In ENCODE, we evaluated the GeneChip arrays with 2- or 3-marker predictors. Since SNP selection for the HumanHap300 product is based on LD information from Phase I HapMap data (including ENCODE), evaluation using this dataset would be biased upwards, and is therefore omitted. We report results only for common SNPs in order to minimize risk of overfitting in the multimarker predictors and thus overestimation of ability to capture rare alleles.



Supplemental Figure 1: Power of a Bayesian approach (using variable weights of the allelic tests of association) vs. the existing, frequentist approach (employing uniform weights) in simulated association studies with either SNPs selected at random or LD-based tag SNPs, at a density of 100 SNPs per 500kb ENCODE region. Power is computed by simulating case-control association studies on resampling of haplotypes in CEU. Results are averaged over all ten ENCODE regions.

Supplementary Figure 2: Same as supplementary figure 1 in Chapter 2

References

1. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311-22 (1995).
2. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-7 (1996).
3. Collins, F.S., Brooks, L.D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8**, 1229-31 (1998).
4. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **34**, D173-80 (2006).
5. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. & Chee, M.S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* **37**, 549-54 (2005).
6. Matsuzaki, H. et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* **1**, 109-11 (2004).
7. Reich, D.E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199-204 (2001).
8. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108 (2005).
9. Altshuler, D. et al. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
10. Kruglyak, L. Power tools for human genetics. *Nat Genet* (2005).
11. Carlson, C.S. et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **74**, 106-20 (2004).
12. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat Genet* **27**, 234-6 (2001).
13. Pe'er, I. et al. Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am J Hum Genet* **78**, 588-603 (2006).
14. Purcell, S., Cherny, S.S. & Sham, P.C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149-50 (2003).
15. Pritchard, J.K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**, 1-14 (2001).
16. Sham, P.C., Cherny, S.S., Purcell, S. & Hewitt, J.K. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* **66**, 1616-30 (2000).
17. Crawford, D.C. et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet* **74**, 610-22 (2004).
18. Chapman, J.M., Cooper, J.D., Todd, J.A. & Clayton, D.G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**, 18-31 (2003).
19. Weale, M.E. et al. Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* **73**, 551-65 (2003).

20. de Bakker, P.I. et al. Efficiency and power in genetic association studies. *Nat Genet* **37**, 1217-23 (2005).
21. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15**, 1496-502 (2005).
22. Pritchard, J.K. & Cox, N.J. The allelic architecture of human disease genes: common disease-common variant.or not? *Hum Mol Genet* **11**, 2417-23 (2002).
23. Cohen, J. et al. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* **37**, 161-5 (2005).
24. Cohen, J.C. et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869-72 (2004).
25. Stram, D.O. et al. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered* **55**, 27-36 (2003).
26. Lin, S., Chakravarti, A. & Cutler, D.J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* **36**, 1181-8 (2004).
27. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33 Suppl**, 228-37 (2003).
28. Roeder, K., Bacanu, S.A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet* **78**, 243-52 (2006).
29. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**, 921-7 (1995).
30. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-5 (2005).

Chapter 5: Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants

Copyright notice: Material in this chapter appeared originally in

Pe'er I, ***Yelensky R***, Altshuler D, Daly MJ.
Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol. 2008 May;32(4):381-5. PMID: 18348202 [PubMed - in process]

Copyright © 2008 John Wiley & Sons Inc
Reprinted with permission of Wiley-Liss Inc. a subsidiary of John Wiley & Sons Inc. (see appendix)

Introduction

Whole Genome Association Studies (WGASs) are examinations of a dense set of SNPs across essentially all available regions of the genome to survey much of common genetic variation for a role in heritable disease traits. WGASs (Hirschhorn and Daly 2005) offer a systematic strategy to assess the influence of common (minor allele frequency $\geq 5\%$) genetic variants on phenotypes (Risch and Merikangas 1996). Although the number of SNPs typed in such a study may vary, typically between 10^5 and 10^6 SNPs, statistical analysis often involves additional testing, so that number of added tests dominates the number of typed-SNPs tested. This additional testing may involve consideration of combinations of typed, promising SNPs that predict nearby alleles in the original samples (Klein et al. 2005) (Wellcome Trust Case Control Consortium 2007), of experimental, second stage typing of such alleles in additional samples (Arking et al. 2006) or of additional sets of SNPs typed in another study for joint analysis (Saxena et al. 2007; Scott et al. 2007; Zeggini et al. 2007) . In all these scenarios, the WGAS aspires to test association to more variants than physically typed, ideally testing all common variants in the genome. Most variants tested will not be associated to any particular phenotype, but may produce false positive association signals, masking potential true positives. Forecasting the null-distribution of these false-positives is important as a practical guideline for interpreting genomewide association scans, akin to classical work (Lander and Kruglyak 1995) directing genomewide linkage analysis of indirectly typed variants. The concrete question is, given an association signal of a certain nominal p-value, how unlikely is it in a WGAS that attempts to examine all common variants?

Results

The number of SNPs on the array may guide multiple testing correction if only these SNPs are tested for genetic association. In contrast, we focus on testing not only typed SNPs, but also most other common variants in the genome. Naïve, Bonferroni (Sidak 1967) corrections for standard testing of multiple, independent hypotheses are overconservative in this context: local correlation among these tests means that effectively there are considerably less independent tests than Single Nucleotide Polymorphisms (SNPs) examined. Theoretical (Tavare et al. 1997) and simulation studies (Lin et al. 2004) relate the number of such tests to the number of historical recombinations, estimated to be much smaller. Yet, no previous systematic evaluation of the testing burden is available on a dense dataset that can mimic fine mapping on a near complete scan of variation, such as the second stage in a multi-staged design.

Such an evaluation is particularly critical to study designs that include a second stage of additional genotyping (Thomas et al. 2004; Skol et al. 2006) or analysis (Klein et al. 2005) around putative causal SNPs that are proposed by the first stage analysis, as these designs do not trivially lend themselves to significance evaluation by permuting phenotypic labels. For 2-stage genotyping designs, common variation is first screened for association signals using cost-effective typing of hundreds of thousands of SNPs (Barrett and Cardon 2006; Pe'er et al. 2006). Next, regions of potentially positive signals are followed-up with denser, saturated SNP sets, in order to validate, refine and strengthen the associations. As well worked out in linkage analysis (Kruglyak and Daly 1998), this directed increase in marker density around positives alters the null signal distribution with the practical effect of mimicking a WGAS of all 6-7 million common SNPs. Hence, permuting 1st-stage data with only the smaller, typed set of SNPs underestimates expected false positives. Permuting the 2nd stage data is possible only for the regions that were followed up, therefore impossible to implement in a nested fashion for every permutation run of the 1st stage.

Implementation of a permutation procedure for study designs with a 2nd stage of analysis in promising regions requires rigorous, automatic criteria for such followup. Since 2ndstage analysis may be based on post-hoc review of the associated region, pinning down the desired followup criteria in an objective fashion is challenging.

The testing burden associated with examining all common alleles does lend itself to empirical evaluation from data, thanks to the Human Haplotype Map (HapMap) ENCODE regions (Altshuler et al. 2005). These regions offer near-complete description of common SNPs (Pe'er et al. 2006) across 1/600 of both the physical and the genetic length of the genome. The demonstrated ability of these regions to represent LD among common variants across the genome (Pe'er et al. 2006) (Altshuler et al. 2005) allows their use for simulating association studies with no true signal (de Bakker et al. 2005). More specifically, we generate the genetic data for a simulated (case or control) individual at an ENCODE region by randomly pairing two of the phased chromosomes available from HapMap trio parents for that region. We repeat this to obtain 2000 individuals randomly labeled cases or controls, mimicking a null study. The maximal Z-score difference in allele frequencies between "cases" and "controls" across all SNPs in such a region is evaluated for significance, and the p-value distribution is estimated by repeating the simulation $N=10^7$ times. This distribution observes more significant p-values then theoretically distributed p-values for a single test statistic due to multiple testing. We repeat this evaluation procedure for the trio-base HapMap populations (CEU and YRI), for all ENCODE regions, and for different cohort sizes. The per-region testing burden is the factor by which significance is exaggerated. As ENCODE regions represent the genomewide average recombination and mutation rates, we propose ENCODE-based extrapolation to estimate the genomewide testing burden in such an association study.

We now outline a formal procedure for estimating testing burden. Suppose the simulation considers a region that spans a fraction $g=1/600$ of the genome (all ENCODE regions totaling 5Mb). For a nominal p-value p that is computed from the theoretical distribution of the association statistic, we tally $n(p)$, the number of studies out of N simulated, at which the best regionwide nominal p-value reaches or exceeds p. $n(p)/N$ is therefore an estimator of the permutation-based p-value regionwide. The expected number H of hits - regions that have a SNP whose score exceeds p - across the genome is therefore

$H(p)=n(p)/gN$. Testing burden is defined to be the ratio between the nominal and permutation-based p-values: $n(p)/pN$ regionwide, or $n(p)/pgN$ genomewide. This can be estimated for every p. Choosing p such that $H(p)=0.05$ would be relevant for the genomewide significance threshold in the initial cohort, whereas $H(p)>1$ would be relevant to a 2-stage design that carries over $H(p)$ false-discovery loci to be typed in additional samples. We chose the middleground, focusing on the value p relevant for a single null hit genomewide. This is motivated by two potential practical outcomes of a study. If a study includes several positive findings, false discovery rate will be much lower than one even when $H(p)=1$, motivating interest in SNPs at that significance level. Alternatively, even in studies consistent with the null hypothesis of no association, this significance level is interesting, as it is approached or attained by the top SNPs that are the most suggestive candidates such a study may propose for additional investigation. We note that this threshold does not formally control familywise error rate, nor false discovery rate, and is intended to provide practical guidelines, rather than be taken literally.

We observe that when $H(p)=1$, the genomewide burden is simply $1/p$. Putting this observation to practical use, we sort the N respective top single-hits in each of the simulations from the smallest (most significant) to the largest. We choose p to be the gN-th value up this list, and report the reciprocal as the testing burden. We note that for a single ENCODE region the expected number of runs achieving such a p–value amongst $N=10^7$ simulations is $gN=\frac{500kb}{3Gb}\times10^7\approx1700$, and the standard deviation of this number is $\sigma=\sqrt{g(1-g)N}\approx40$. This provides a practical way to estimate confidence in estimating the gN-th order statistic due to the number of simulations being finite by considering $(gN-2\sigma)$-th and $(gN+2\sigma)$-th order statistics. Another source of sampling error has to do with the small fraction of the genome being analyzed. The differences in estimation across ENCODE regions can guide us with respect to this sampling error.

Figure 1a reports the extrapolated number of independent tests required to mimic the expectation of the best p-value in a WGAS, i.e. the empirical testing burden. For all

ENCODE SNPs, we find the testing burden to be around one million tests in the HapMap European (CEU) samples, and for all common SNPs, we find the testing burden to be roughly half million tests in the same population: considerably lower than available bounds' that prove the number of edges in the Ancestral Recombination Graph to exceed the number of independent tests in a dataset (Lin et al. 2004). As such edges can be attributed to either splits or recombinations, their number depends on the sample size (negligible in the context of the entire genome) and ancestral recombination events. The formula $\mathrm{Log}(k)\times N_e \times R$ in (Tavare et al. 1997) estimates 1.1 million common recombinations in Europeans, where: k is the number of coalescence branches considered the reciprocal of the minor allele frequency threshold for sites considered, i.e. $k = 20$ for common SNPs; N_e is the effective population size , ~10,000 in Europeans; R is the average number of recombination events per meiosis, 36.

A practical, first-cut guideline for correcting nominal p-values may be multiplying them by this genomewide testing burden. This means, for instance, that the probability of a WGAS in a European population that examines all common alleles to exhibit, by random chance alone (no true genetic effect), a result with p-value$<10^{-7}$ is smaller than 0.05. In the HapMap African (YRI) samples, that have more SNPs, and less linkage disequilibrium, testing burden is higher at one million. Since ENCODE data are still incomplete w.r.t. rare variants, they provide only a lower bound on their associated testing burden, showing it to be more than 2-fold higher than for common alleles.

Testing burden varies across the different ENCODE regions, which may be expected given that ENCODE regions deliberately represent a variety of genomic characteristics (The International HapMap Consortium 2003). Empirical standard deviation across the 10 regions amounts to 19.6% of the testing burden, in both YRI and CEU populations (Fig 1a), suggesting a standard error of 6.2% in estimating average testing burden from 10 regions. We have evaluated the sampling error due to finite number of simulation by considering different order statistics as described above, and showed it to be smaller than 0.2%.We therefore ascribe most of the observed variation in estimates to sampling different regions. Yet, the process of selecting of ENCODE regions made sure their average GC content, gene content, recombination rate, etc. were similar to the genomewide average (The International HapMap Consortium 2003). While we offer no genomewide evidence that ENCODE is representative of the genome in terms of other measures such as testing burden examined here, this premise had been adopted by others using ENCODE data is used as a standard benchmark for estimating frequencies of genomewide phenomenon in a wide domain of applications (Birney et al. 2007)..We note that testing burden is not strongly correlated neither with the actual number of common SNPs in the particular region ($R^2<0.03$), nor with the regionwide recombination rate ($R^2<0.01$; see Fig 1b). In retrospect, this justifies extrapolation of our measurements from ENCODE to the entire genome by physical span.

It is important to realize that testing burden is not constant across p-values: association signals with more extreme p-values involve more burden (Fig 1c). This means that accurately correcting statistical tests by a constant factor is impossible. Our simulations validate the formal analysis of modeling multiple genetic tests (Dudbridge and Koeleman 2004)(Hirschhorn and Daly 2005) in pointing out that restriction of such modeling to a constant testing burden does not sufficiently capture the full correlation structure between tests. There is no genomewide testing burden to fit all significance levels, but rather one can correct for such multiple testing by a burden function which depends on the significance level of interest. This means the best practice for correcting a nominal p-value for the entire genome is to use a lookup-table, rather than a fixed correction factor.

In order to better understand the intuition behind this variable testing burden, we recall that a constant testing burden arises in the context of independent multiple statistical tests. In contrast, dense SNPs along the genome are partially and locally correlated to varying extents. Formally, the pair (Z_a, Z_b) of Z score statistics of two correlated alleles of different, nearby SNPs, a and b, respectively, will have a bivariate normal distribution, with mean (0,0) and covariance matrix $\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$. If the allele a is significantly associated, showing a standard normal score $Z_a = z_0$, then given this association, the allele b will have a nonzero expected standard score, with the conditional distribution being $(Z_b \mid Z_a = z_0) \sim N(rz_0, 1-r^2)$. The chance of *b* to achieve *a*'s significance level is $\Pr(N(rz_0, 1-r^2) > z_0) = \phi\left(-z_0\sqrt{\frac{1-r}{1+r}}\right)$. The events X_b and X_b of *a* and *b* achieving this significance level, respectively, thus have correlation

$$\rho(X_a, X_b) = \frac{Cov(X_a, X_b)}{\sqrt{Var(X_a)Var(X_b)}} = \frac{\phi(-z_0)\phi\left(-z_0\sqrt{\frac{1-r}{1+r}}\right) - \phi(-z_0)^2}{\phi(z_0)\phi(-z_0)} = 1 - \frac{\phi\left(z_0\sqrt{\frac{1-r}{1+r}}\right)}{\phi(z_0)},$$

which is decreasing with Z_0. This means that the more significant is the p–value, the lesser the correlation coefficient is, or in other words, the lower the significance the less correction for multiple testing the correlated tests require.

Fortunately, in a 2-stage design of a WGAS, the first stage is designed for a true positive to reach only a moderate p-value, expected to be achieved by numerous sites (Skol et al. 2006). Such a stage would require less correction for multiple testing than the final stage aiming at genomewide significance.

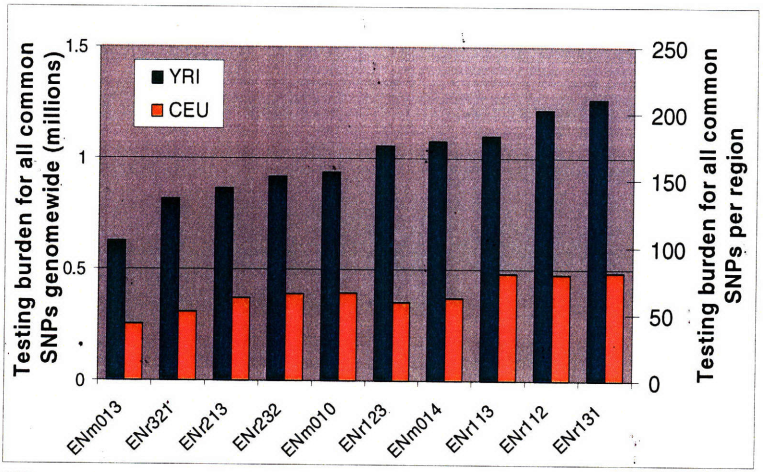
Finally, studies of larger size show more burden of multiple testing (Supplementary fig 1). We hypothesize that this effect is also related to the increased power of larger studies to distinguish highly- (but not perfectly-) correlated causal variants. An alternative explanation is that this observed effect is an artifact of our oversampling design: Rather than simulating data by a true bootstrap procedure that samples real data without replacement for each simulated dataset. We are simulating datasets of thousands of individuals based on 120 chromosomes only. We note that a similar result was not observed in a similar set of oversampling analyses (Dudbridge 2006), suggesting attribution of this increased burden to the density and redundancy of ENCODE data we use.

Discussion

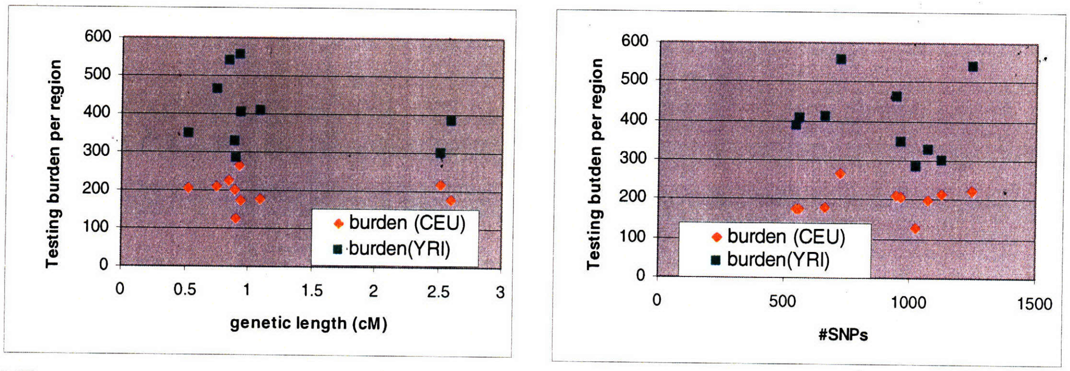
These and other results offer considerable understanding of the distribution of null signals in idealized association studies. Practical association studies may exhibit more extreme p-values then predicted by our study even without real effects due to demographical and genotyping technology differences between cases and controls that create artifactual hits. Furthermore only the accumulating experience in such studies will reveal more about the complementary parameters describing the alternative hypothesis, which speak to the number and strength of true signals. Together, the distribution of null and true signals will enable rigorous decision whether a given result indicates true association.

Figures

1A



1B



1C

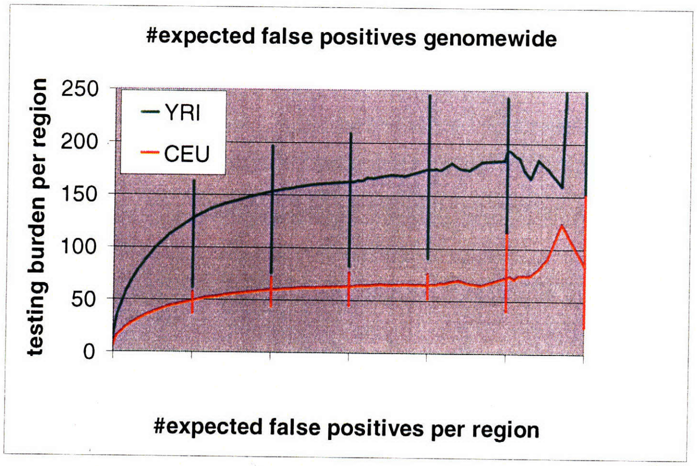
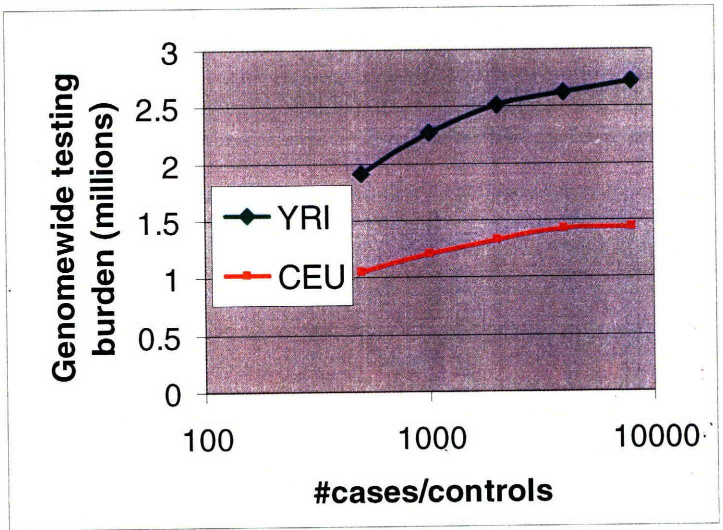


Figure 1 legend: A. The empirical testing burden (y-axis) for all common SNPs in different ENCODE regions in the HapMap panels of Yorubans from Ibadan, Nigeria (YRI; green) and CEPH individuals of European ancestry fro Utah (CEU; orange). Testing burden is estimated from simulated null studies of 1000 cases, 1000 controls extrapolate to the entire genome, as extrapolated from ENCODE. **B.** The testing burden (y-axis) of each region as a function of the region's length in centiMorgans (x-axis, left) or of the number of SNPs tested (x-axis, right) **C.** The testing burden (y-axis) of all (smooth) or common (tick-marked) SNPs in a typical ENCODE region (ENr213), as a function of the empirically evaluated p-value (x-axis).



Supplementary Figure 1 legend:
Testing burden (y-axis) extrapolated to the entire genome from simulated studies different numbers of cases/controls (x-axis) in YRI (green) and CEU (orange) data from ENCODE, averaged across all regions.

References

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ et al. (2005) A haplotype map of the human genome. Nature 437(7063): 1299-1320.

Arking DE, Pfeufer A, Post W, Kao WH, Newton-Cheh C et al. (2006) A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat Genet 38(6): 644-651.

Barrett J, Cardon L (2006) Study Design Issues in Whole Genome Association Studies. Nat Genet Submitted.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146): 799-816.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ et al. (2005) Efficiency and power in genetic association studies. Nat Genet 37(11): 1217-1223.

Dudbridge F (2006) A note on permutation tests in multistage association scans. Am J Hum Genet 78(6): 1094-1095; author reply 1096.

Dudbridge F, Koeleman BP (2004) Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet 75(3): 424-435.

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nature reviews 6(2): 95-108.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308(5720): 385-389.

Kruglyak L, Daly MJ (1998) Linkage thresholds for two-stage genome scans. Am J Hum Genet 62(4): 994-997.

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11(3): 241-247.

Lin S, Chakravarti A, Cutler DJ (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat Genet 36(11): 1181-1188.

Pe'er I, Chretien Y, PIW PdB, Barrett J, Daly M et al. (2006) Biases and reconciliation in estimations of linkage disequilibriumin the human genome. American journal of human genetics 73(4).

Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D et al. (2006) Evaluating and Improving Power in Whole Genome Association Studies using Fixed Marker Sets. Nat Genet 38(6).

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273(5281): 1516-1517.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316(5829): 1331-1336.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316(5829): 1341-1345.

Sidak Z (1967) Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association 62: 626-633.

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38(2): 209-213.

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145(2): 505-518.

The International HapMap Consortium (2003) The International HapMap Project. Nature 426(6968): 789-796.

Thomas D, Xie R, Gebregziabher M (2004) Two-Stage sampling designs for gene association studies. Genet Epidemiol 27(4): 401-414.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145): 661-678.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS et al. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316(5829): 1336-1341.

Chapter 6: Genetic Analysis of Human Traits *In-Vitro*: Drug Response and Gene Expression in Lymphoblastoid Cell Lines

Copyright notice: Material in this chapter is currently in review/revision for publication at the journal PLOS Genetics and is reproduced here under the Creative Commons Attribution License:

This work is licensed by the authors Edwin Choy and Roman Yelensky under the Creative Commons Attribution 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

Introduction

Genetic mapping by linkage and association is an unbiased approach to discover genes and pathways influencing disease traits and responses to drugs and environmental exposures [1]. Unlike model organisms that can be exhaustively phenotyped and readily exposed to drugs and toxins in the laboratory, there are substantial limits to the phenotypes that can be safely elicited or measured in human subjects. Thus, there would be great value in a human *in-vitro* model that faithfully reflects both *in-vivo* genetics and physiology while allowing for systematic perturbations and characterizations in high throughput. Such a model would be particularly useful to study the function of sequence variants mapped by whole genome association studies of common human diseases that do not fall in obvious coding sequences [2-6], many of which are presumed to influence disease traits by regulating gene expression.

One such model system has been proposed and extensively studied: EBV-transformed lymphoblastoid cell lines (LCLs) derived from human B-lymphocytes [7-13]. Lymphoblastoid cell lines have long been produced as renewable sources of DNA as part of normal and diseased cohorts. Initially, LCLs derived from genotyped CEPH pedigrees [14] and HapMap participants [15] were used to identify genomic regions linked to and associated with inter-individual variation in RNA transcript levels (these "expression" QTLs are referred in the text below as "eQTLs") [16-19]. A small number of such eQTLs have been found to also be associated with human disease [20-22]. Moving beyond studies of gene expression, LCLs have also been used to attempt to identify genetic variants that predict for response to radiation and drugs *in vitro* [23-26]. The mapping of eQTLs and drug response QTLs have been combined by some investigators, seeking non-random relationships between genotypes at single nucleotide polymorphisms (SNPs), baseline RNA levels, and response to chemotherapeutic agents [27,28]. One study claimed identification of eQTLs that explain up to 45% of the variation seen between individuals in cell sensitivity to chemotherapy [28].

The generality and utility of findings in LCLs is ultimately a function of how well the biology and genotype-phenotype relationships of LCLs reflect reality *in-vivo*. While the DNA sequence of an LCL is typically a stable representation of the human donor [29], little is known about the stability and biological meaning of cellular traits studied *in vitro*. The degree to which potentially confounding variables can influence LCL-based studies has not been studied. There are many opportunities for non-genetic variability to be introduced in the path from the whole human to LCL's *in-vivo* (Figure 1): the random choice of which subpopulation of B-cells (i.e. naïve B cells vs memory B cells vs activated B cells with or without terminal differentiation) are selected in the process of immortalization, the amount of and individual response to EBV, the history of passage in cell culture and culture conditions, the laboratory protocols and reagents with which assays are performed, and the measurements used to assess drug response and RNA phenotypes.

We set out to map genetic contributors predicting *in vitro* drug response in LCLs with a series of experiments designed to also address concerns about the potential for noise and unmeasured confounders in such a complex model [30] . While our data is broadly consistent with results of previous studies, we find that measurable confounders are a stronger influence on drug responses and RNA levels in LCLs than are DNA variants. Even after incorporating both SNP and RNA data in our model, few compelling associations of SNPs to drug response were observed. Our study identifies and addresses several potential issues in the design and interpretation of experiments that aim to find relationships between DNA variants, expression levels, and drug response. As with *in vivo* genotype-phenotype studies, larger samples sizes are required to elucidate true relationships. Notably, this is not a comprehensive survey of all traits in LCLs, but rather an in-depth analysis of several important traits that are readily measurable and increasingly studied. Thus, there may well exist other traits that are more robust and less influenced by confounders and noise. However, in so far as the traits we approached are typical, the findings are germane to any *in-vitro* system used to study genotype-phenotype relationships.

Results

Data Collected

We studied 269 cell lines densely genotyped by the International HapMap Project [31]. Cell lines were cultured and characterized at baseline for a variety of cellular phenotypes including growth rate, ATP levels, mitochondrial DNA copy number, EBV copy number, and measures of B-cell relevant cell surface receptors and cytokine levels. Each cell line was exposed in 384-well plates to a range of doses for each of seven drugs selected based on their divergent mechanisms of action and importance in clinical use for treatment of B-cell diseases, focusing on anti-cancer agents: 5-fluorouracil (5FU), methotrexate (MTX), simvastatin, SAHA, 6-mercaptopurine (6MP), rapamycin, and bortezomib. Drug response was measured using Celltiter Glo, an ATP-activated intracellular luminescent marker that, when compared to mock-treated control wells, can represent relative levels of cellular viability and metabolic activity. RNA was collected at baseline and RNA transcript levels were measured genome-wide on the Affymetrix platform.

Baseline characterization and plating for drug response experiments was performed using batches of 90 cell lines from each HapMap analysis panel (CEU, JPT / CHB, and YRI) on each of three experiment days. The order of cell lines within each panel was randomized to avoid inducing artificial intra-familial correlation. Each drug was tested at a range of doses around the expected IC50 as reported for the drug by the NCI DTP; each dose of drug was tested in two wells per plate and on two separate plates. These replicate measurements for each cell line allowed assessment of intra-experimental variation.

To evaluate day-to-day (i.e. inter-experimental) variation in all traits, a subset of 90 cell lines (30 from each of the three HapMap panels) was grown from a fresh aliquot and the entire experiment was repeated. To evaluate the effect of technical error on measured RNA levels, a set of 22 RNAs previously expression profiled (using Illumina HumanChip) at Wellcome Trust Sanger Institute (WTSI) (generously provided by Emmanouil T. Dermatsakis) was included in expression profiling at the Broad on Affymetrix arrays.

Data can be downloaded from the Broad Institute web site: (http://www.broad.mit.edu/~yelensky/cell_lines_paper/). Please see Materials and Methods for details of QC, normalization, etc.

Cell line sensitivity to chemotherapeutic drugs

Gene mapping of drug response (or any cellular phenotype) in LCLs requires that the phenotype be: (1) precisely-measured technically, (2) biologically reproducible across independent experiments, and (3) remain relatively free from confounding factors. We assessed each of these characteristics in turn before performing genome-wide association scans.

To evaluate variability in drug response across replicate plates assayed on a given experiment day (technical reproducibility), we calculated the "relative" response of a cell line to a drug by measuring the (signed) distance of that cell line's dose-response curve for the drug on a given plate to the dose-response curve for the drug averaged across all cell lines assayed that day, in that replicate plate set. (The two replicate plates for each cell line performed on an experiment day were arbitrarily placed into set A or B.) This non-parametric approach allowed all drugs to be treated uniformly (see Methods) and generated two data points per cell line, per drug, per day. We ranked the cell lines based on their relative response in plate set A and separately based on values from plate set B. The rank-correlation (Spearman's rho) for relative response across sets A and B was high (rho=0.86 to rho=0.99, Supplemental Table 1), indicating that drug response on a given day is highly reproducible.

To evaluate variability across independent experiments assayed on separate days (biological reproducibility), we repeated the assay on a subset of ~90 cell lines (30 from each of the three HapMap ethnic panels). We calculated the relative response of a cell line to a drug on each day by measuring the (signed) distance of that cell line's dose-response curve for the drug (estimated using all values on both plates) to the dose-response curve for the drug averaged across all cell lines on that day, producing two values per cell line per drug (day 1 and day 2). (At this point, we noted that our assays for rapamycin and bortezomib suffered from weak response and strong dependence on drug batch respectively, and these drugs were not studied further, see Methods for a full account). For the remaining five drugs, cell lines were ranked based on relative response on day 1 and again on day 2, and the rank-correlation (Spearman's rho) was calculated. In comparison to the high technical reproducibility on a given day, drug response was more variable across independent experiments (rho=0.39-0.82, Supplemental Table 2)

We next noted that the rank order of cell lines based on relative response was strikingly similar between three drugs (5FU, 6MP, and MTX). In fact, the rankings of cell lines for these three drugs were as similar as the cell line rankings for biological replicates of the same drug on different days (Figure 2A and Supplement table 3). Furthermore, we found a similar correlation of relative response to a distinct pair of drugs, 5FU and docetaxel, in the publicly available data of Watters et al.[25] (Figure 2B). (This correlation likely explains why these investigators found linkage for both drugs to the same genomic locus). While correlation in relative response to multiple drugs could, in theory, indicate a shared genetic mechanism common to many drugs, it could also suggest the influence of an experimental confounder that more strongly influences drug response than does genetic variation.

Indeed, we found that just such a confounder, as the baseline growth rates of the individual cell lines contributed to the above correlation between drugs (Figure 2C; Supplemental Table 3). Growth-rate was modestly reproducible across days (rho=0.37), though not significantly heritable (h2=0.4, pval=0.1). The dependence of drug response on growth rate in LCLs, though never previously reported, is unsurprising as all three agents impact the cell cycle. Using a differential equation model of drug response accounting for the kinetics of exponential growth under exposure to drug (see Methods), we estimated a growth rate adjusted EC50 for each cell line for each of the three affected drugs. This approach removed the bulk of the correlation between drug responses and between drug response and growth rate (Supplement table 4), though some correlation of responses persisted. Standard EC50s were fit for Simvastatin and SAHA.

Baseline ATP concentrations (i.e. the average of Celltiter glo values for all mock-treated wells across both assay plates for each cell line in our drug-response experiments, see Methods) were then found to be correlated to growth rate adjusted EC50s for MTX and 5FU (Figure 2D). We interpret such ATP levels as a relative measure of the gross sum of metabolic activity combined across all viable cells within an assay well. Like growth rate, ATP levels were reproducible across biological replicates (rho=0.6) but not significantly heritable (h2=0.19, pval=0.38). After further adjusting the growth rate adjusted EC50s for MTX and 5FU for ATP levels using linear regression, the correlation across drugs was nearly abrogated (Supplemental Table 5).

Having adjusted for confounding due to growth rate and ATP levels where necessary, we performed genome-wide association studies examining EC50s for each drug and SNPs from HapMap Phase 2 with Minor Allele Frequency (MAF) >10% [32]. We did not observe any associations that surpassed genome wide significance (p-val < 5e-8). In addition, the distributions of statistical association between SNPs and EC50s did not significantly exceed expectation under the null. Our lack of evidence for association between SNPs and drug responses is consistent with prior publications [24-28], none of which identified specific SNPs with genome wide significance.

Variability in RNA expression

Previous studies observed baseline levels of RNA expression correlated with response to cisplatin and etoposide [24,27,28]. While associations with RNA levels need not imply a causal relationship as a common experimental or confounding process could simultaneously affect both RNA levels and drug response (Figure 6A), the subset of causal associations that are influenced by genetic variation may be highlighted by integrating information on SNP genotype, SNP associations to RNA levels (eQTLs), and RNA correlations to EC50s (adjusted for growth rate and ATP levels where appropriate) [27,28]. In other words, by examining a two-step model SNP -> RNA -> drug response, power may be increased to detect SNPs that influence drug response through their effect on RNA. We thus turned our attention to RNA measurements in LCLs.

As with drug response, measurements of RNA expression, to be biologically meaningful, need to be reproducible between replicates on a given day, between experiments performed on different days, and largely unconfounded by experimental artifacts. One common metric for estimating reproducibility in expression data is to rank the level of expression of all genes in a single hybridization, and to compare the rank of all genes in a separate hybridization of either another aliquot of the same RNA (technical replicates) or newly derived RNA from the same cell line on a different experiment day (biological replicates). When we assessed the reproducibility of ranked RNA levels using this common metric, we observed a high correlation between biological replicates, i.e. samples that were independently thawed and profiled twice in our experiments (Figure 3A – black curve). Yet, we also saw a similarly high correlation between profiles from any pair of unrelated individuals (Figure 3A – red curve) and even across human-derived cell lines paired with chimpanzee-derived cell lines (Figure 3A – blue curve). These observations simply reflect the unsurprising reality that inter-individual variation in the level of a single gene is small compared to the overall dynamic range in expression levels across all genes. Thus, reproducibility of the rank order of transcript levels across an individual is uninformative with regard to mapping of DNA variants influencing specific genes.

A more useful metric for gene mapping is the reproducibility in rank order of individuals based on the level of expression of a given gene. If the level of a single RNA transcript in one individual is reproducibly higher than the same RNA transcript in another individual, then it may be possible to identify genetic variants contributing to inter-individual variation of this RNA transcript (i.e. an eQTL). If variation of an RNA transcript level among individuals is low relative to technical and biological noise, however, then it will not be possible to map genetic influences for expression of the gene.

We performed such an analysis for each of the 3,538 genes expressed in the cell lines (using standard criteria for Affymetrix expression arrays). The analysis used samples from 49 unrelated individuals that were independently thawed, cultured and profiled on two different days (Figure 3B). In contrast to the results in Figure 3A, we see that the rank-correlation of individuals based on individual genes is typically modest (rho = 0.25-0.3). That is, only a fraction of the 3,538 RNA transcripts examined in LCLs vary enough in levels among individuals (relative to technical and biological noise) to be reliably measured for association to genetic variants.

To parse out the contributions of technical vs. biological noise to this variation, we examined the reproducibility of rank orders of cell lines when the same RNA sample was profiled on two different platforms (thereby eliminating variability due to cell culture and RNA isolation). Specifically, RNA for 14 unrelated individuals (from YRI HapMap subset) and their expression profiles (generated using the Illumina system) were generously provided by M. Dermitzakis (WTSI). These same RNA samples were profiled at Broad Institute on Affymetrix microarrays. When we calculate the reproducibility in the rank order of RNA samples based on these two cross-lab, cross-platform technical replicates, we observe a median rank-correlation of rho = 0.55 (Figure

3C – gold curve). This is much higher than the corresponding biological reproducibility (independently repeating cell culture and RNA extraction) for the cross-lab, cross-platform comparison recomputed for the same 14 individuals (rho~0.2, Figure 3C), demonstrating the substantial biological variation of RNA expression in LCLs, in addition to the expected measurement error inherent in such experiments.

To limit the impact of technical noise, we restricted analysis to one thousand genes that supported the greatest technical reproducibility in rank order of individuals (rho>~0.7, median rho ~ 0.85). Genes excluded by this threshold include both those that are technically well measured but invariant across individuals and those for which there exists inter-individual variation but technical noise on one platform or the other overcomes the signal. (As the WTSI performed four technical replicates while we performed only a single technical replicate, the WTSI data had lower overall variance.) Genes excluded by this filter typically varied less across individuals, particularly in the better-measured WTSI dataset. (median standard deviation of 1000 best-measured genes = 0.27 vs 0.17 for the other ~2500 expressed genes; p-val<1e-15).

When analysis of reproducibility was limited to the one thousand genes with the highest reproducibility in technical replicates, the correlation across biological replicates was improved but still modest (rho=0.55, Figure 3D – cyan). That is, despite excellent technical reproducibility (panel 3A) even relative to inter-individual variation (Panel 3C), the rank order of individuals based on most genes was only partially reproducible.

We reasoned that some of the biological noise might be due to other measured factors, as had been the case for drug response. Above a threshold of 5% variance explained, growth rate was associated with levels of expression for relatively few genes (<5%), but ~15% of genes showed association to EBV copy number (Figure 4A), some of which encoded for proteins known to participate in transduction pathways downstream of EBV signaling. Moreover, >25% of genes were associated with ATP levels (Figure 4B). In total, over

40% of genes have at least 5% of their variation in RNA levels associated with one of the three measures above (Figure 4E).

The association of RNA levels to such factors could, in principle, represent intrinsic characteristics of each LCL (which could potentially be due to inherited DNA sequence variation, acting indirectly through susceptibility to EBV infection or inducing a metabolic state). Alternatively, growth rate, EBV infection, and metabolic state could simply represent experimental noise that obscures genetic contributions to gene expression variation. Interestingly, measurements of EBV copy number, ATP, and growth rate at Broad correlate with levels of RNA expression generated independently at WTSI [18,19] (Figure 4F), albeit more weakly than for the expression profiles generated on the same samples at the Broad. Thus, these confounders display a component intrinsic to each cell line, not only noise.

To examine how much of the variability in gene expression could be attributed to inherited DNA variation, and how much to other factors, we examined cis-eQTLs that influence RNA expression levels in our experiment. Using HapMap Phase 2 SNPs with MAF>10% that lie within a 0.15Mb window around each gene, we performed standard linear regression between expression values of that gene and SNP genotypes coded 0,1,2 (representing the number of minor alleles carried by the individual). In our dataset, ~9% of genes harbored a cis-eQTL that explained at least 5% of the gene's variance in expression levels in excess of the fraction of genes showing such association in permuted datasets (Figure 4C). Even more eQTLs were evident in the WTSI expression data (which employed four technical replicates), in which >20% of genes were associated with a SNP that explains 5% or more of the variance (Figure 4D). Consistent with previous analyses [16-18], in both data sets RNA transcript levels for a small fraction of genes were associated with a cis eQTL that explained a large proportion of the variance. Moreover, the overall proportion of genes that showed association to growth rate, EBV, and ATP far exceeded the proportion associated to a cis-eQTL of the same strength (compare figure 4E to 4C).

Inter- and Intra- individual variance component analysis

To parse the association of SNPs and other measures to variation in gene expression further, we decomposed the total variance in expression of each gene into components contributing towards inter-individual versus intra-individual (experimental) variation. When we examine the effect of each cis-eQTL, growth rate, EBV copy number and ATP levels on each variance component (see Methods), as expected, eQTLs contribute only to inter-individual variation (Figure 5A). EBV and ATP, on the other hand, influence either inter-individual or intra-individual variation, depending on the gene (Figure 5B and 5C).

Taken together, these observations have a number of implications: First, more genes are associated with the measured non-genetic cellular factors than are associated with individual cis-eQTLs. Second, these non-genetic factors influence gene expression not only by varying across cell lines in a reproducible manner (like SNPs), but also by varying across experiments for the same cell line. Third, for some genes, a given non-genetic factor is associated only with inter-individual variation (genes arrayed along the x-axis in Figure 5), and yet for other genes that same factor is associated only to intra-individual variation (genes arrayed along the y-axis). Factors that are associated with inter-individual variation could, in principle, represent causal processes related to the action of a genetic variant, whereas those that only vary across experiments represent noise with respect to genotype-phenotype correlation.

Having characterized non-genetic influences on levels of RNA expression, we proceeded by looking at variation in RNA expression levels to understand its contribution to the variation observed in drug response. Consistent with prior reports, we observed a large number of genes whose level of RNA expression at baseline was correlated to drug response. Levels of RNA transcripts for 20% of genes in the Broad Institute dataset and 18% in the WTSI dataset were associated (at a rho^2>0.05) to growth-rate and ATP adjusted EC50 of at least one of the drugs assayed. EC50s for SAHA and 5FU appeared to have the strongest relationship to RNA levels, correlating to 8.7% of genes measured at the Broad and to 7.7% of genes measured at WTSI.

Applying the variance components analysis to see how inter- and intra- individual variation in growth-rate and ATP adjusted EC50s are influenced by RNA levels (and "assigning" to a given gene its strongest correlated drug), we observed that RNA levels are predominantly associated with inter-individual differences in EC50s (Figure 5D). Much less of the correlation between RNA expression and EC50s reflects intra-individual variation.

This observation that variation in RNA levels mainly influences inter-individual variation in EC50s allow us to hypothesize that eQTLs for such RNAs could be used to identify genetic contributors to drug response.

Integrating data from eQTLs and drug response in LCLs

Given that we observe significant correlation of both SNPs to RNA expression levels (eQTLs) and RNA levels to inter-individual variation in drug response,

we sought to increase confidence regarding SNPs modestly associated with drug response by combining evidence from the two relationships. The hypothesis we hope to prove is that the RNA-drug response relationships are causal (i.e. $eQTL_A$ and RNA_A in figure 6A), and thus eQTLs for these RNAs are causal influences on drug response themselves, and can be detected by following the path from genetic variant to RNA level to drug response phenotype. The "null" hypothesis is that the association of SNPs to RNA levels and of RNAs to drug response are orthogonal and independent. (i.e. $eQTL_B$ and RNA_B in figure 6A). Under this scenario, the apparent correlation between RNA level and drug response is induced by another factor (a non-genetic confounder), and there is no actual causal link between RNA and drug response directly.

To test these hypotheses, we set out to examine the fraction of genes whose expression is influenced by an eQTL and correlated to drug response. As seen in Figure 4, ~14% and 4.5% of genes have cis-eQTLs (r^2>0.08, FDR<10%) in the WTSI and Broad Institute datasets respectively. At the same time, 18% (WTSI) and 20% (Broad) of genes are correlated to drug response (rho^2>0.05, FDR<10%). When we consider the intersection of eQTL-bearing genes and drug-response correlated genes in each dataset independently, we see that 1.4% (WTSI) and 0.9% (Broad) of genes are both correlated to drug response and bear a cis-eQTL. Neither intersection contains more genes than would be expected by chance alone (and only a small fraction of genes are involved), suggesting that the effect of genetics of RNA expression on drug response is likely small.

Nevertheless, it is possible that some SNP-RNA-Drug response "tuples" contained in this intersection are of type $eQTL_A$ and RNA_A (from Figure 6A); if these can be distinguished from others of type $eQTL_B$ and RNA_B, it may be possible to discover true causal relationships. To evaluate this possibility we focused on genes that both correlated to drug response and bore a cis-eQTL among the set of 1000 "best-measured" genes in each RNA dataset independently, yielding a total of 23 unique SNP-RNA-Drug response tuples for examination.

If a SNP genotype induces differences in RNA levels of a gene between individuals, and these RNA differences induce inter-individual differences in response to drug, we should be able to observe significant association between SNP and drug-response directly. Consistent with this idea, when we regress the drug EC50 against SNP genotype for each of the 23 pairs above, we do see a modest degree of excess association (Figure 6B). Moreover, the association appears to be in the expected direction, based on the directions of the pair-wise relationships SNP-RNA and RNA-Drug response (Figure S1). Notably, a simulated dataset with the same SNP/RNA/Drug variances and independent SNP-RNA/RNA-Drug pairwise covariances (i.e. the $eQTL_B/RNA_B$ scenario in Figure 6A) as the real 23 tuples fails to demonstrate the excess association between SNPs and drug-response we see in the real data (Figure 6B – gray lines). Though no highly significant examples were documented, these observations would appear to support the existence of causal SNP-RNA-Drug response relationships (i.e. the $\mathrm{eQTL_A/RNA_A}$ scenario in Figure 6A).

In (Figure 6B) most p-values deviate (however modestly) from null, suggesting that many may be causal. However, we recognize one oft-overlooked bias in this type of analysis which introduces inflation to the SNP-Drug p-values in absence of (or in addition to) real signal: the "winner's curse" [33] overestimate of effect size incurred during discovery of eQTLs. Specifically, because we set a threshold (r^2>0.08) for the discovery of SNP-RNA associations and because the bulk of the effect size distribution is at or below this threshold, many of the identified associations (particularly those just crossing the threshold) will have been identified because the true effect was overestimated by chance (Figure S2). While the true portion of the SNP-RNA association does not cause correlation between SNP and drug response, the chance correlation over and above the true connection may also be correlated to drug response; that is, the SNP is by chance correlated to orthogonal variance in the RNA, which could in turn be correlated to drug response by another factor (the $\mathrm{eQTL_B/RNA_B}$ scenario). To bear this intuition out, we replaced all real eQTL effects modeled in the simulation in Figure 6B with an eQTL whose true effect is r^2=0.05, but only look at the subset of simulated datasets where the observed effect is r^2>0.08. In this simulation, likely more representative of the thresholded discovery overestimate of effect, we recreate an inflation of p-values similar to that observed in the real data (Figure 6C).

Clearly, we would like to distinguish any true SNP-Drug response associations from those that are spuriously inflated, but null. Unfortunately, because we don't know the real effect sizes of discovered eQTLs, it is not possible to do so analytically. However, it is helpful to consider what a real, causal SNP-RNA-Drug response tuple might look like. In particular, we would expect the causal tuples with strong SNP-RNA associations to have, on average, stronger SNP-Drug associations. Tuples where no causal relationship between RNA and Drug response (and thus SNP and Drug response) exists would exhibit either strong SNP-RNA association or increased SNP-Drug association, but not both. If we plot the strength of each eQTL against the strength of the SNP-Drug response association, a striking picture emerges: most of increased association between SNP and Drug response comes from the weaker eQTLs, while most of the stronger eQTLs have no association to drug response (Figure 6D), consistent with the bias exposed above. Additionally, an interesting group of 3 tuples emerges that are both relatively strong SNP-RNA and SNP-Drug response associations (Figure 6D blue arrow). These are: rs1384804-C8orf70 (Ensembl:ENSG00000104427)-MTX, rs3733041-GLT8D1 (Ensembl:ENSG00000016864)-5FU, and rs2279195-SH3TC1 (Ensembl:ENSG00000125089)-Simvastatin with SNP-Drug p-values of 0.03, 0.05, and 0.02 respectively. We propose these are interesting candidates for follow-up and suggest that similar analyses be carried out in future studies integrating genetic variation, RNA expression, and other traits.

Discussion

Recent studies have shown that a substantial fraction of genes contain cis-eQTLs that explain a modest fraction of inter-individual variation in RNA levels. Other studies used LCLs to perform linkage and association scans for drug response traits [26-28]. However, few reports characterize the biological reproducibility of these phenotypes, and none to our knowledge document their correlation to non-genetic factors such as growth rates, EBV copy number, and metabolic activity. We document that most traits, whether drug responses or RNA transcript levels, are only partially reproducible across experiments, and that more genes are correlated to cellular growth rate, ATP levels, and EBV copy numbers than to genetic factors (at comparable fractions of variance explained). Thus, in addition to issues of statistical power relative to genetic size of effect, day to day variability in a trait and confounding factors are major influences on gene mapping experiments in LCLs. Though we looked at only a limited number of traits in LCLs (response to seven drugs and RNA transcript levels), our findings are applicable to any experimental system where genotype to phenotype correlations are studied.

Consistent with prior reports, our genome-wide association studies of drug response did not reveal any SNPs directly associated to drug response with genome-wide significance. The inability to detect such SNPs is likely due to lack of power to detect small sizes of effect with limited sample size and in the presence of significant confounding and noise.

Several studies attempted to improve power to discover SNPs associated with drug response [15,16] by integrating data on RNA levels and eQTL mapping [18,19]. Whether these eQTLs are incidentally or causally associated with drug response depends on whether the cognate RNAs influence drug response or are merely correlated to drug response by a non-genetic factor that simultaneously affects both phenotypes. Our results show modest enrichment for association to drug response (EC50s adjusted for growth rate and ATP levels) among eQTLs in genes whose RNA levels correlate to drug response; most of this enrichment can be attributed to unavoidable biases in the analysis. Accounting for these biases, our three most promising associations are rs1384804 near C8orf70 to MTX, rs3733041 near GLT8D1 to 5FU, and rs2279195 near SH3TC1 to Simvastatin.

This result is similar to that recently published by Emilsson et al. [34] in which eQTLs for genes whose RNAs were correlated to obesity were not convincingly associated to obesity. Thus, while it is attractive to envision a general overlap between SNPs associated to drug response or disease traits and those associated to RNA levels in accessible tissues, empirical evidence for truly causal relationships remains anecdotal [22]. Even in cases such as IRF5, where there is both strong association to disease and to RNA levels in LCLs, the actual patterns of association to RNA levels and disease are quite disjointed for the different mutations at the locus [21].

A major limitation of the HapMap samples is their relatively small sample size for performing a genome-wide association study. While better powered studies (such as those proposed to study cell lines from eight thousand and one-hundred thousand individuals by the Framingham Heart Study [29] and the National Children's Study [30]), can certainly address the power issue, these larger studies face even greater challenges in managing LCL culture conditions and experimental confounders in high throughput to minimize bias and noise [30]. By highlighting these aspects of the LCL model, as well as tackling how some of them can be addressed, we hope to build a stronger foundation on which these important experiments can be planned and carried out.

Methods and Materials

Cell Culture

EBV-transformed lymphoblastoid cell lines were acquired from the NHGRI Sample Repository for Human Genetic Research in frozen aliquots. Cells were thawed in 5mL culture medium (RPMI medium 1640 (Invitrogen) supplemented with 10% FetalPlex (Gemini), 2mM L-Glutamine (Invitrogen), and 1x penicillin/streptomycin (Invitrogen)). Cell lines were counted daily using Z2 Coulter Counter (Beckman Coulter) and passaged as needed to maintain a concentration of 2-5 x 1e5 cells/ml at 37C in a 95% humidified 5% CO2 atmosphere.

Initially, cells were grown until 5 x 1e5 cells/ml were reached in 50 mL total volume. Then, ten identical aliquots were frozen in 1 mL freezing media containing 50% FetalPlex, 40% RPMI 1640 medium, and 10% DMSO (Sigma) at -80C for 24 hrs and transferred to liquid nitrogen. These aliquots were used to provide biologic replicates for the experiments described below.

Aliquots were thawed on experiment day #1 as described above. Cell lines were counted daily and passaged as need to maintain a concentration of 4-8 x 1e5 cells/ml in 10 mL culture medium. On experiment day #7, cells were counted and distributed for use in the various experiments described below. One cc of culture was used for immediate immunophenotyping via FACS and Luminex beads. One cc of culture was used for RNA and DNA extraction using Trizol (Invitrogen) following the manufacturer's protocol. The remaining eight cc of culture were used for drug response assays described below.

Drug Response Assay

The drugs that we studied are bortezomib (courtesy of T. Hideshima), rapamycin (Biomol), 5-fluorouracil (Sigma), methotrexate (Sigma), 6-mercaptopurine (MP Biomedicals), SAHA (Biovision), and simvastatin (Calbiochem). These drugs were arrayed in a source plate in the concentrations according to supplemental figure. The source plate was pinned into each cell line in duplicate, resulting in each drug concentration being assayed in each cell lines 4 times.

For drug response assays, LCLs for each cell line were diluted to 1 x 1e5 cells/ml, and 25 uL of cell culture were plated into each well of two white solid flat bottom 384 well plates (Corning cat# 3704) using a microplate dispenser (Multidrop Combi, Thermo Scientific). Next, 100 nL was pin-transferred from the source plates into the plates containing cells using an automated 384 channel simultaneous pippettor (CyBi-Well, CyBio). Plates were incubated at 37C in a 95% humidified 5% CO2 incubator.

After 48hrs, plates were removed from the incubator to room temperature for 10 minutes prior to being vortexed for 30 seconds. 25uL of Celltiter Glo (Promega Cat No. G7573) diluted 1:3 in PBS was added to each well with the Mutlidrop microplate dispenser and shaken for two minutes. Luciferase luminescence was then immediately measured for each well using a multiplate illuminometer (Envision, Perkin Elmer).

Raw data is publicly available online:

http://chembank.broad.harvard.edu/assays/view-project.htm?id=1000477

The experiment was monitored for cell-culture handling, plating, pinning, and assay errors and failed cell lines/plates/drug-rows were excluded from down-stream analysis. (Most cell lines were successfully assayed on two plates for all drugs, however; specific counts are below.)

Luminescence values in drug-exposed wells were divided by the median control-well luminescence in the same plate row (after excluding plate edge wells) to obtain 4 viability fractions per cell line, per drug, per dose, in each experiment. For evaluation of technical reproducibility, the median of the 2 fractions on each plate was taken as the cell line's response to that dose on that plate. For evaluation of biological reproducibility and all other analyses, the median of the 4 fractions was taken as that cell line's response to that dose in the experiment. Drug responses were examined, and it was noted that the experiment failed to achieve meaningful cytotoxic response to rapamycin, with most cell lines reaching a maximum fractional viability of only ~0.6-0.7, even at highest concentration of drug assayed. It was concluded that the viability assay was not a relevant read-out for rapamycin response, and the drug was not considered in further analyses.

Overall cell line response to a given drug was then calculated by taking the average response to a dose across all cell lines in the experimental batch (cell lines were assayed in batches of ~90), subtracting the average from the value for each cell line, and then averaging the result for each cell line across all doses. (The 4-5 low-concentration doses where all cell lines had a fractional viability of ~1 were excluded from the calculation.) In this way, the (single value) relative response of a given cell line to a drug was calculated, representing the non-parametric distance of that cell line's dose-response curve to the average dose-response curve for that drug in the experiment. (For the analysis of technical reproducibility, the calculation was done using only replicate plate A for all cell lines, and then using only replicate plate B, and the two values were compared). Quality control then proceeded by examining the dependence of response on the compound stock plate from which the drugs were pinned. (Compound stock plates were prepared with enough drug to run ~20 cell lines and drug response should be independent of the drug stock.) Indeed, it was noted that for 5FU, 6MP, Simvastatin, SAHA, and MTX, dependence on drug stock was weak, while for bortezomib, the dependence was profound, with large differences in response between different plates, significantly in excess of the differences between cell lines on a given plate. Thus, bortezomib was excluded from further analysis. Though dependence on compound plate for the other 5 drugs was weak, average response for each compound stock plate was subtracted from each cell line using that plate (for each drug independently) and this normalized response was carried forward.

In summary, after the processing steps above in the main batch of experiments, 254 cell lines were successfully assayed for response to 6MP, 256 for MTX, 260 for Saha, 262 for Simva, and 259 for 5FU. 84 cell lines were then again successfully measured for all 5 drugs as biological replicates. (For ease of comparison, technical reproducibility is also reported using only the two plates from these biological replicate samples.) These values are available as "relative responses" in the online supplement. Analyses in Fig 2 use this data for the ~200 successfully measured unrelated individuals, after again centering within each HapMap panel. Also, the median (non-boundary) control well luminescence over the two plates for each cell line was taken as the "ATP content" of the cell line. The value was divided by 100,000 and centered within each HapMap panel.

Modeling Drug response

To account for the effect of growth-rate on response to MTX, 5FU, and 6MP, we reasoned as follows: Assume a simple ODE model of cell line population growth: $\frac{dP}{dt} = rP$, where P(t) is the # of cells in the population at a given time, and r is the (unobserved in the specific drug-exposure experiment) growth rate parameter. This ODE has the solution: $P(t) = P_0 e^{rt}$. When the cell line is exposed to drug, its growth-rate is impaired in a concentration-dependent manner. Taking inspiration from first-order Michaelis-Menten kinetics, we can model this as:

$$\frac{dP_{drug}}{dt} = r\left(1 - \frac{\text{Max reduction} * [ConcDrug]}{\text{Concentration for half maximal reduction aka "EC50"} + [ConcDrug]}\right)P,$$

which is solved by $P_{drug}(t) = P_0 e^{r\left(1-\frac{\text{MaxRed}*[ConcDrug]}{\text{EC50}+[ConcDrug]}\right)t}$. As our observed luminescences are ratios between drug wells and control wells at given concentrations, we can write

$$\frac{P_{drug}(t)}{P(t)} = \frac{P_0 e^{r\left(1-\frac{\text{MaxRed}*[ConcDrug]}{\text{EC50}+[ConcDrug]}\right)t}}{P_0 e^{rt}},$$ which can simplified as $$\frac{P_{drug}(t)}{P(t)} = e^{\frac{-r*\text{MaxRed}*[ConcDrug]}{\text{EC50}+[ConcDrug]}t}.$$

There are two identifiable parameters in this model: the concentration necessary for half-maximal reduction in growth-rate (EC50) which is independent of growth rate r itself, and r * maximal reduction of r, a product term dependent on growth rate whose components cannot be independently estimated. The model was fit for each cell line, for each drug independently, using median measurements at all doses. QC was performed by excluding all models with RSS>0.08. The –r*MaxRed term was discarded, and the EC50 was carried into further analysis after centering the values within each HapMap panel. (257 cell lines were successfully fit for 5FU, 251 for 6MP, and 255 for MTX.) Models were also successfully fit to all 84 biological replicates of 6MP and 5FU, and 82 replicates of MTX. ATP correction for 5FU and MTX was then carried out by taking the residuals of the linear regression DRUG~ATP.

SAHA and Simvastatin were modeled by a standard sigmoid[35], with response (fractional viability) at a given dose $= \text{Max Inhibition} + \frac{1\text{ - Max Inhibition}}{1 + e^{slope*(\log(dose)-\log(\text{EC50}))}}$. Notably, max inhibition and EC50 are *not* the same as above, here representing a minimal viability and the concentration at which that minimal viability is achieved, respectively. Maximum inhibition (aka minimum viability) were <0.05 for most cell lines for simvastatin and varied between ~0.1-0.3 for SAHA. The EC50 was carried into further analysis after centering the values within each HapMap panel. Again, QC was performed by excluding all models with RSS>0.08. (257 cell lines were thus successfully fit for Saha and 261 for Simvastatin.) Models were also successfully fit to all 84 biological replicates of Saha and 5FU, and 83 replicates of Simvastatin. The GWAS for drug response was performed with all successfully measured individuals, while analyses presented in Figures 2,5,6 were performed with unrelated individuals only.

Growth Rate Measurements

Each cell line was seeded at a concentration of 2 x 1e5 cell/mL in 2 mL. LCLs were counted daily for five consecutive days with an automated particle counter (Z2 Coulter Counter, Beckman Coulter). A regression of the form $\log(\text{conc day i}) = r * i + \log(\text{conc day 0})$ was fit for each cell line to obtain the estimate of growth rate *r*. QC was performed by evaluating the 95% confidence interval of the *r* estimate and rejecting estimates whose interval width exceeded 1.1. Thus, estimates of growth-rate for 237 cell lines were obtained. An abbreviated second replicate of the experiment was repeated on a subset (155) of the cell lines with only the 3rd day counts collected to evaluate growth rate reproducibility.

FACS Analysis

From each LCL, ~25,000 cells were incubated with R-Phycoerythrin–conjugated mouse anti-human antibody to cell surface markers (CD19, CD20, CD21, CD40, CD58, CD80, CD86, CD95, CD227, IgD, IgG, IgM, HLA-DQ, HLA-DR, and IL6R) at 4°C for 30 min. Cells were washed once with PBS and 1% fetal bovine serum and were fixed with 1% paraformaldehyde. Data on cell-surface expression in each cell line were acquired using a fluorescence-activated cell sorter (BD Biosciences FACSCalibur system). To quantify expression for each LCL, we used flow cytometry, requiring at least 500 cells per LCL for it to be included in our analysis. Fluorescence intensity was measured for the anti-cell surface protein antibody and a control isotype antibody for each LCL. A marker (and, separately, a control) histogram was created by placing individual cell measurements into 1,024 equally spaced intensity bins. Counts in the control histogram were subtracted from the marker histogram to obtained a "normalized" histogram of cell-counts in each of the 1,024 intensity bins. The average intensity was then calculated from this normalized histogram and the log of this value was carried forward into QC as the average normalized marker expression for that LCL.

QC then proceeded by regressing this marker expression on the total cell count obtained for that marker within a given experimental batch of LCLs. (samples were batched by HapMap panel) We reasoned that if the experiment was successful, there should be no dependence of cell-surface marker expression on the quantity of viable cells obtained in the experiment; if there was such a dependence, the marker expression was likely reading out handling differences between LCLs, not true, intrinsic differences in expression. Indeed, by this metric, we found that during the first batch of experiments that was attempted (for the CEU panel), only 4 markers were successfully measured, while subsequent batches (YRI + CHB/JPT samples) succeeded for 14 and 9 markers respectively. In most markers that passed this filter, it was further noted that a few cell lines showed very low expression, far from the overall distribution of the values for each batch. While it is conceivable that these represent true differences, we interpreted these values as individual LCL measurement failures, and further truncated the lowest 5% of values within each marker in each batch. Thus, the final dataset contains measurements of: 85 cell lines for CD19 and CD20, 169 for CD21, 166 for CD227, 248 for CD40, 164 for CD58, 166 for CD80 and CD86, 248 for CD95, 80 for HLADQ, 85 for HLADR and IgM, and 165 for IgD, IgG, and IL6R. These values were centered within each panel and carried into further analysis.

Luminex Assay

30 HapMap cell lines were screened with a multiplex antibody bead kit from Biosource (Cytokine 25-Plex for Luminex (Catalog #LHC0009)). Of the 25 cytokines originally selected for this assay, 8 were reliably detectable (lower concentration: IL8, IL10, IL12p40, TNFa, IP10; moderate concentration: MIP1a, MIP1b, RANTES). Of these, it was found that measurements for MIP1a and MIP1b were strongly correlated; thus we decided to include only MIP1b in further experiments. These 7 cytokines were assayed in the remainder of the cell lines according to the following protocol:

One cc for each LCL was placed into a single well of 96-deep well plate. The samples were centrifuged at 500 rpm for 5 minutes at room temperature. The supernatant was placed into a new 96-well plate, and placed dry ice to be stored at -80 degrees All assays were performed on a single thaw.

The cytokines were measured following the manufacturer's protocol. In order to ensure that the measured cytokine concentration fell in the linear part of the standard curve, the lower concentration cytokines were multiplexed together (final dilution 1:2); and MIP1b and RANTES were multiplexed together (final dilution 1:6).

The concentration of each cytokine was calculated based on the standard curve generated by the same plate, after subtracting out the "blank" background. A 3-parameter model was used to convert median fluorescent intensity (MFI) to protein concentration (ng / ml). A subsequent correction was applied to account for the dilution factor at the time of the assay. All final concentrations are expressed as pg / ml and log-transformed. 262 cell lines were successfully measured for IL10, IL12, IL8, IP10, and TNFa, and 266 measurements were obtained for MIP1b and RANTES. (79 and 87 biological replicate measurements were also obtained for the above two sets of cytokines respectively).

RNA preparation and Affymetrix expression profiling

All LCLs were cultured in the fashion described above. Prior to the plating of cells for the Drug Response Assay, 5 x 10^5 cells were set aside for RNA extraction. Cells were immediately lysed with Trizol Reagent (Invitrogen). RNA was collected according to the manufacturer instructions. 1.25 ug total RNA (OD>1.8) was diluted to a total volume of 10uL. RNA was processed and hybridized onto Affymetrix Human U133A whole genome RNA expression genechip® arrays according to the manufacturer's protocol. Gene expression summary values for the whole dataset were computed by RMA[36,37] and log-transformed. Measurements were successfully obtained for 257 HapMap cell lines in the main experiment, for 64 biological replicates, for 24 cell lines originally thawed at the WTSI, as well as multiple replicates of 5 cell lines derived from chimpanzees. (raw data in "all broad cell line expression data.zip" online)

For analysis, the dataset was further processed as follows: 1) The ~22K total probe sets on the Affy U133A were restricted to the 9084 judged expressed (p-value<0.06) by the Affymetrix software in at least 2/3 of 50 randomly selected scans. 2) These 9084 expressed probes were matched by Genbank transcript accession number (NM_#) to the 13,300 targets judged expressed by the same criterion in the WTSI Illumina HapMap experiments. (using the probability of detection p-value output by the Illumina software.) This yielded a reduced set of 3600 Affy probes (3592 Illumina targets) whose transcripts were reliably detectable in both experiments. 3) To obtain a comparable dataset from the WTSI Illumina data, we took the median over their 4 technical replicates for each target and quantile normalized across all samples. 4) We averaged within each gene symbol, in each dataset, for each sample, to get the set of 3538 genes expressed in both experiments and measured on both platforms. 5) To prevent family structure from introducing bias, the dataset was restricted to unrelated individuals only for the analyses in Figures 3-6: 198 each in the main Broad and WTSI experiments, 49 biological replicates at the Broad, and 16 samples for whom RNA was extracted at the WTSI and measured in both locations. Both centered (for each gene within each panel) and uncentered data is available in "cleaned expression data used in analyses.zip" and were each used as appropriate.

Relative EBV and mtDNA Copy Number

All previously collected DNA was diluted to PCR concentration of 2.5 ng/uL and arrayed in 384 well storage plates (AbGene Cat No. AB-0564). Custom TaqMan assays were designed using Primer 3 (http://frodo.wi.mit.edu/) and ordered from Applied Biosystems. The EBV copy number assay interrogated a 66_bp fragment at the DNA polymerase locus (EBV forward primer 5'GACGA TCTTGGCAATCTCT3', EBV reverse primer 5'TGGTCATGGATCTGCTAAACC3', EBV probe 5'6FAM-CCACCTCCACGTGGATCACGA-MGBNFQ3'). The mtDNA copy number assay examined a 72 bp fragment at the ND2 locus (mtDNA forward primer TGTTGGTTATACCCTTCCCGTACTA, mtDNA reverse primer

CCTGCAAAGATGGTAGAGTAGATGA, mtDNA probe sequence 5'6FAM-CCCTGGCCCAACCC-MGBNFQ3').

As an internal reference, a 90bp assay from the NRF1 locus on chromosome 7 was multiplexed with EBV or mtDNA (NRF1 forward primer 5'CTCGGTGTAAGTAGCCACAT 3', NRF1 reverse primer 5'GAGTGACCCAAACCGAACAT 3', NRF1 probe 5'VIC-CACTGCATGTGCTTCTATGGTAGCCA-MGBNFQ 3'). Equal efficiency of amplification was observed for each assay in the multiplex reaction. Final Concentrations for EBV primers, mtDNA primers, EBV probe, mtDNA probe, NRF1 primers and NRF1 probe were .25 uM, .25 uM, 10uM, 10uM, 1uM and 10uM respectively. 5ng of DNA template was used for each TaqMan reaction performed according to the manufacturer's protocol. Relative EBV and mtDNA copy number was determined by the difference of CT method[38]. Log-transformed. EBV measurements were obtained when cell lines were first received from Coriell (257), during the main batch of experiments (257), and for the biological replicate set (86). Mitochondrial DNA measurements were obtained only for 252 cell lines in the main experiments.

Fraction of RNA variance explained by cellular phenotype or eQTL (Figure 4)

We are interested in the fraction of gene-trait (or gene-eQTL) relationships that are real (i.e. would reach statistical significance given enough samples) and above a given r^2 (rho^2) thresh-hold in the current sample. So, we want $P(real, r^2 >= c)$ in joint distribution notation, i.e. a relationship can be real (non-null) or spurious (null) and can exceed a certain threshold or not. By regressing a trait on multiple genes, we observe: $P(r^2 >= c)$. It is the fraction of relationships exceeding any given threshold, the green curve. By permutation, we also have: $P(r^2 >= c \mid not\_real)$, the blue (average of black) curve. So, we write, by conditioning on whether a relationship is real or not:

$$\begin{aligned} &P(r^2 >= c) = \\ &= P(r^2 >= c \mid real)P(real) + P(r^2 >= c \mid not\_real)(1 - P(real)) \\ &= P(real, r^2 >= c) + P(r^2 >= c \mid not\_real)(1 - P(real)) \end{aligned}$$

Or, rewriting, we have:

$$P(real, r^2 >= c) = P(r^2 >= c) - P(r^2 >= c \mid not\_real)(1 - P(real))$$

Everything on the right hand side is known, except P(real), the true proportion of gene-trait relationships in the data. This can theoretically be estimated ala Storey et al. 2003 [39] but the estimate can be unreliable in the setting of dependencies, as is the case in our data since genes are largely in clusters. So, we take the worst case scenario, setting P(real)=0. Thus, we have:

$$P(real, r^2 >= c) \geq P(r^2 >= c) - P(r^2 >= c \mid not\_real)$$

So, $P(r^2 >= c) - P(r^2 >= c \mid not\_real)$ is then a lower bound for $P(real, r^2 >= c)$, the black curve. It is important to note that the interpretation of this lower bound is limited to the sample size used in the analysis. Given more samples, the estimate will change to even more genes being affected by traits or eQTLs, albeit at lower r^2s.

Decomposing gene expression into inter- and intra- components (Figure 5)

To estimate the amount of inter- and intra- individual variation present for each gene in the ~50 unrelated individuals thawed and measured twice at the Broad Institute, we fit a random effects model of the form $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, where i indexes the individuals, and j is 1 or 2 for the biological replicate being considered. The estimated variance component σ_α^2 is then the inter-individual variation in gene expression for the gene, while the residual variance σ_ε^2 is the intra-individual variation. To evaluate the effect of a cis-eQTL or cellular phenotype on an RNA, a fixed effect x corresponding to trait was then added to the model to get: $y_{ij} = \mu + \beta x_{ij} + \alpha_i + \varepsilon_{ij}$. The resultant change in variance components σ_α^2 and σ_ε^2 can then be interpreted as the "effect" of that trait or snp on RNA expression. The directionality of the effect is clearly only known for SNPs, but the nature of relationship (inter-, intra-, or both) can be examined for any trait. It's worth pausing to reflect on what these "effects" mean: If including a QTL SNP genotype in the model reduces inter-individual variance (as the overwhelming majority of SNPs do, Fig 5a), it implies that fixed differences in genotypes (QTLs) between individuals correlate to fixed differences in expression between individuals in the corresponding gene. (as one would expect) If, on the other hand, the intra-individual variance component is reduced when accounting for a given trait, the implication is that day-to-day variations in the trait correspond to day to day variations in the RNA. As would be expected, some genes also show a combination of the two effects. Finally, these estimates are quite noisy, suffering from random fluctuations in RNA levels, measurement error, and the relatively small sample size available for the analysis; estimation is likely even less reliable for weaker effects. Nevertheless, the analysis is instructive for the stronger signals and overall patterns and would improve given more samples and technical replicates.

GWAS for drug response

1,045,141 autosomal SNPs with MAF>10% in each of the 3 (CEU, YRI, CHB/JPT) HapMap panels were selected from the Phase 2 HapMap build 21 for association testing to drug response phenotypes. The between/within family model of association was tested for each SNP against each drug, in each panel independently, using PLINK[32] v1.02 with options "--qfam-total --geno 1 --aperm 100 100000000 0.00000005 0.0001 5 0.001". For each drug, p-values for each SNP were then combined across panels using Fisher's method. 25,735 X-chromosome SNPs were tested analogously, but using an additive model on unrelated individuals only with PLINK command line "--assoc --geno 1"; none exceeded 5e-8. QQ plots for the autosomal SNPs for each drug are available at: http://www.broad.mit.edu/~yelensky/cell_lines_paper/snps_vs_drug_response_pvalues/.

R – Aside from GWAS scans performed using PLINK, all other analyses were performed using R version 2.5.0[40].

Figures and Tables

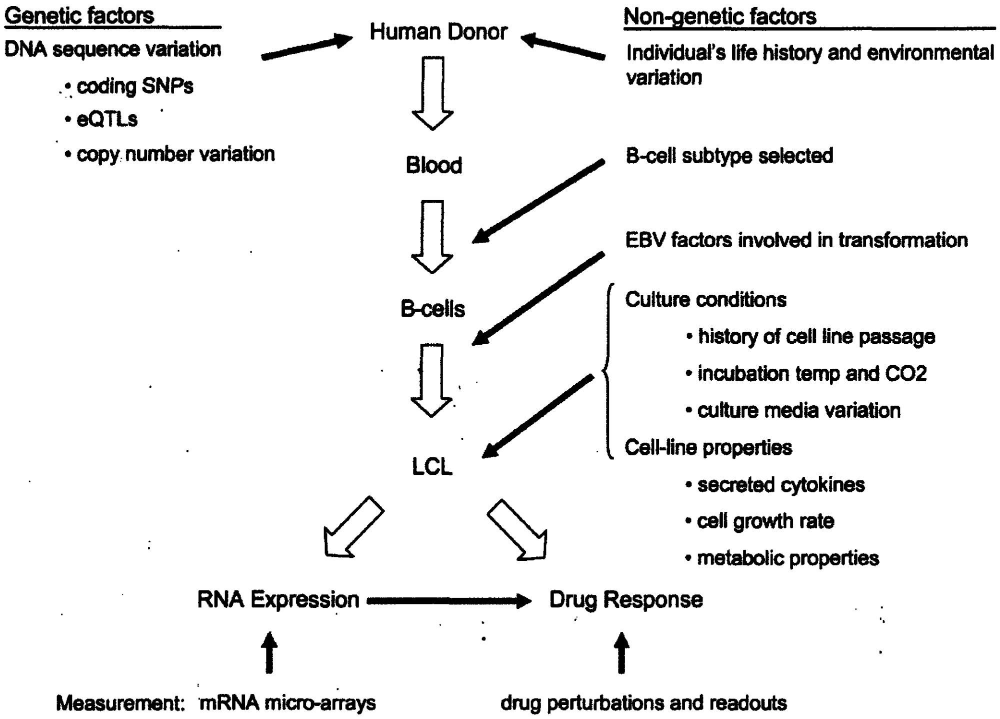


Figure 1: Genetic and Non-genetic Factors influencing lymphoblastoid cell lines as a model system to understand human physiology.

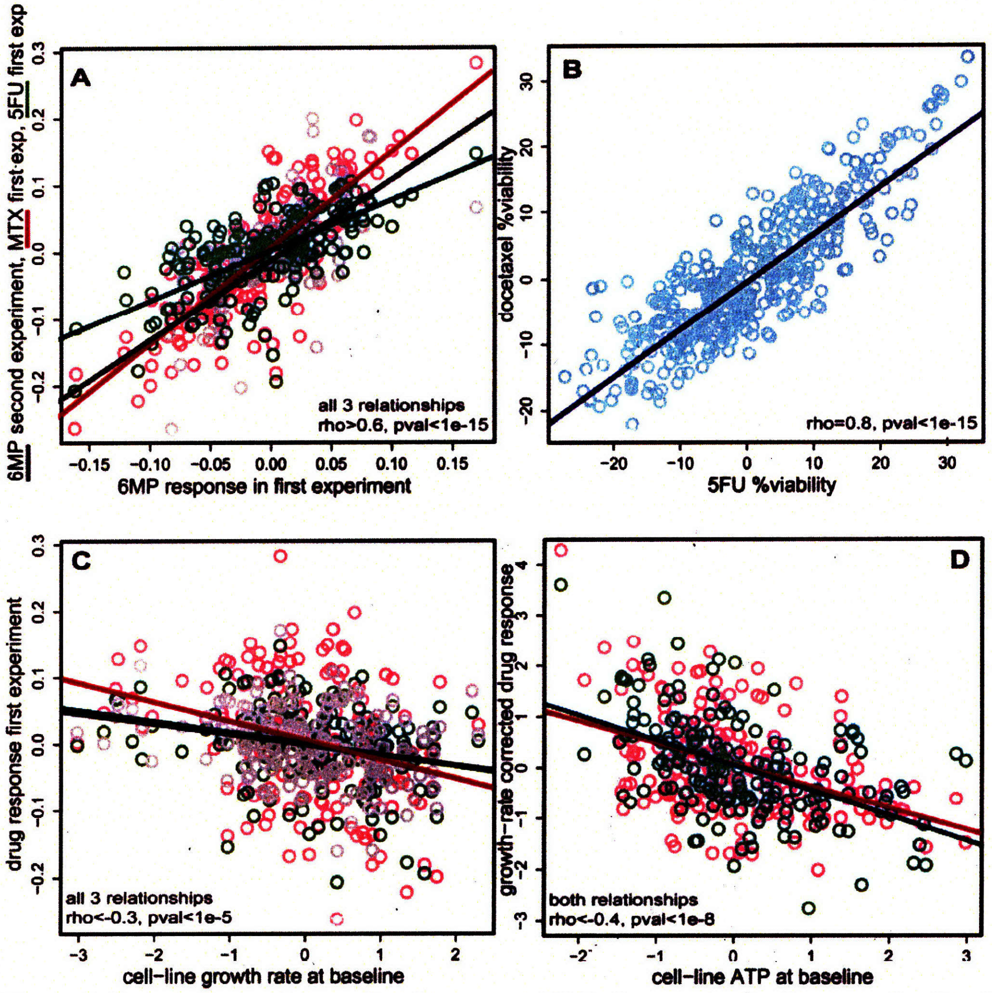


Figure 2: Drug response is correlated across multiple drugs, to growth rate and to baseline ATP levels of the cell line.

(A) Relative drug responses were calculated for each individual as described in Methods to obtain a single number summary of the cell line response to each drug on each day. The black circles represent an individual cell line's relative response to 6MP assayed on day one plotted against 6MP relative response assayed on day two. The red circles similarly represent relative response to 6MP plotted against relative response to MTX, both assayed on day one. The green circles represent relative response to 6MP plotted against relative response to 5FU, again both assayed on day one. Lines represent regressions for each of the three comparisons and show that not only is relative drug response a reproducible trait, but also can be correlated across multiple drugs.

(B) Using online data made publicly available by Watters et al. [25], relative drug response to docetaxel and 5FU was calculated using the 427 individuals with no missing data to obtain a single number for each drug, in each individual, as in (A). Response to docetaxel was plotted against 5FU for each individual. The line represents the regression for the comparison and indicates that the effect observed in (A) is neither limited to our experiments, nor to the particular drugs we attempted.

(C) The baseline growth-rate of each individual's cell line was estimated as described in the Methods. This growth rate is plotted against relative response for 6MP (black), MTX (red), and 5FU (green). Lines represent regressions for the respective comparisons and all correspond to significant correlations.

(D) For each individual, baseline ATP levels were measured using Celltiter glo in the mock-treated wells in drug response assays. EC50 response was calculated correcting for growth rate (see Methods). Relative ATP levels were plotted against the growth-rate corrected EC50 for MTX (red), and 5FU (green). Lines represent regression for the comparisons and indicate significant correlations.

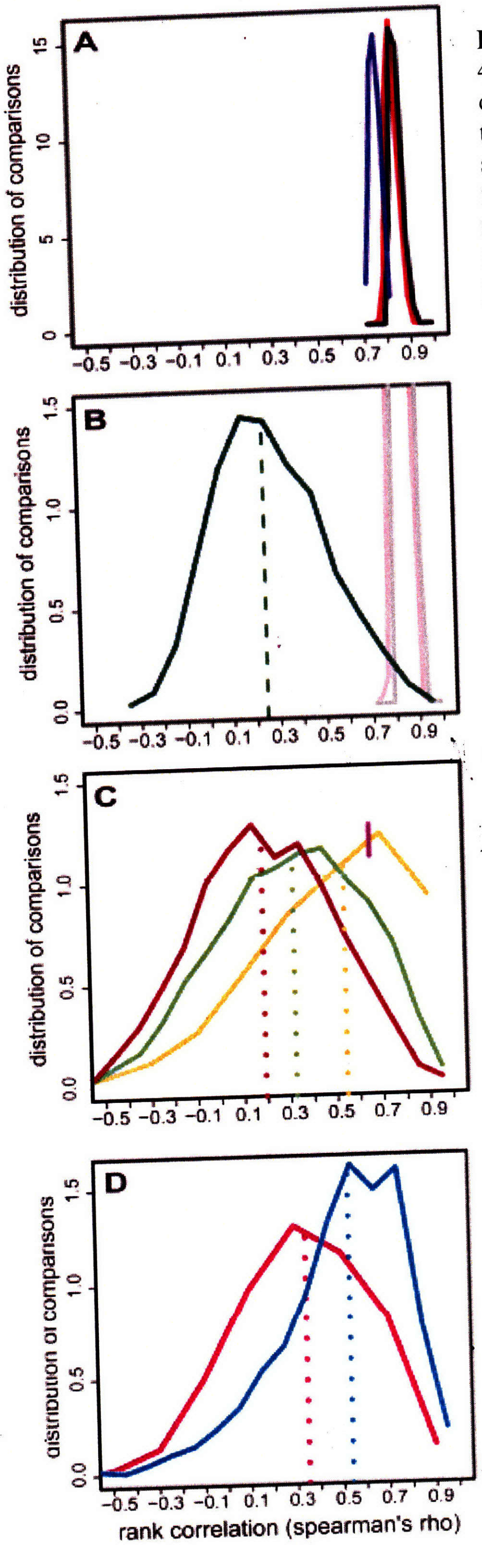


Figure 3: Biological variation in RNA expression

49 unrelated individuals were whole-genome RNA profiled on the Affymetrix platform in two independent experiments at the Broad Institute. (same-platform biological replicates) A subset of 14 (of the 49) were also profiled independently at the WTSI on the Illumina platform (cross-platform biological replicates) and an aliquot of that RNA ("WTSI RNA") was again profiled at the Broad Institute on the Affymetrix platform. (cross-platform technical replicates)

(A) Expression values of all 3538 expressed genes were ranked in each of the 14 unrelated individuals in the two Broad Institute biological replicate experiments and ranks were compared between: the same individuals in two separate experiments (black); all pairs of unrelated individuals across two experiments (red); 5 chimpanzees assayed in the first experiment and all individuals assayed in the second experiment (blue). Plot shows that *overall* expression profiles in LCLs are highly similar across biological replicates, between unrelated individuals, and even across species.

(B) The 49 individuals were ranked according to their relative levels of each gene in the first Broad experiment. The ranking was then independently repeated for the second Broad experiment. Ranks were compared across the two experiments for each gene and the results plotted in (green), with the median of the distribution in (dotted green). Plot shows that when any given gene is examined, there is substantial variation in the relative order of *individuals* between two independent experiments, despite the relative order of *genes* being highly stable as shown in (A). Light black and red lines are same as (A) for comparison.

(C) On the set of 14 individuals, per-gene rank comparisons as in (B) are computed for: WTSI RNA assayed on the Illumina platform vs. WTSI RNA assayed on the Affymetrix platform (gold solid and dotted); WTSI RNA assayed on the Illumina platform vs. RNA extracted at the Broad Institute during the first experiment and assayed on the Affymetrix platform (brown solid and dotted); the two independent Broad experiments as in (B), (green solid and dotted). Plot shows substantial *biological* variation in the relative levels of any given gene when profiling experiments are repeated, far in excess of that might be expected from measurement error alone. Magenta dash indicates the cut-off for the 1000 "technically best-measured" genes to use in (D).

(D) The analysis for the brown and green curve in (C) is repeated only for the 1000 "best-measured" genes and plotted in magenta and cyan respectively. Plot shows that even if measurement noise is limited, a substantial portion of the variance in gene expression represents biological noise.

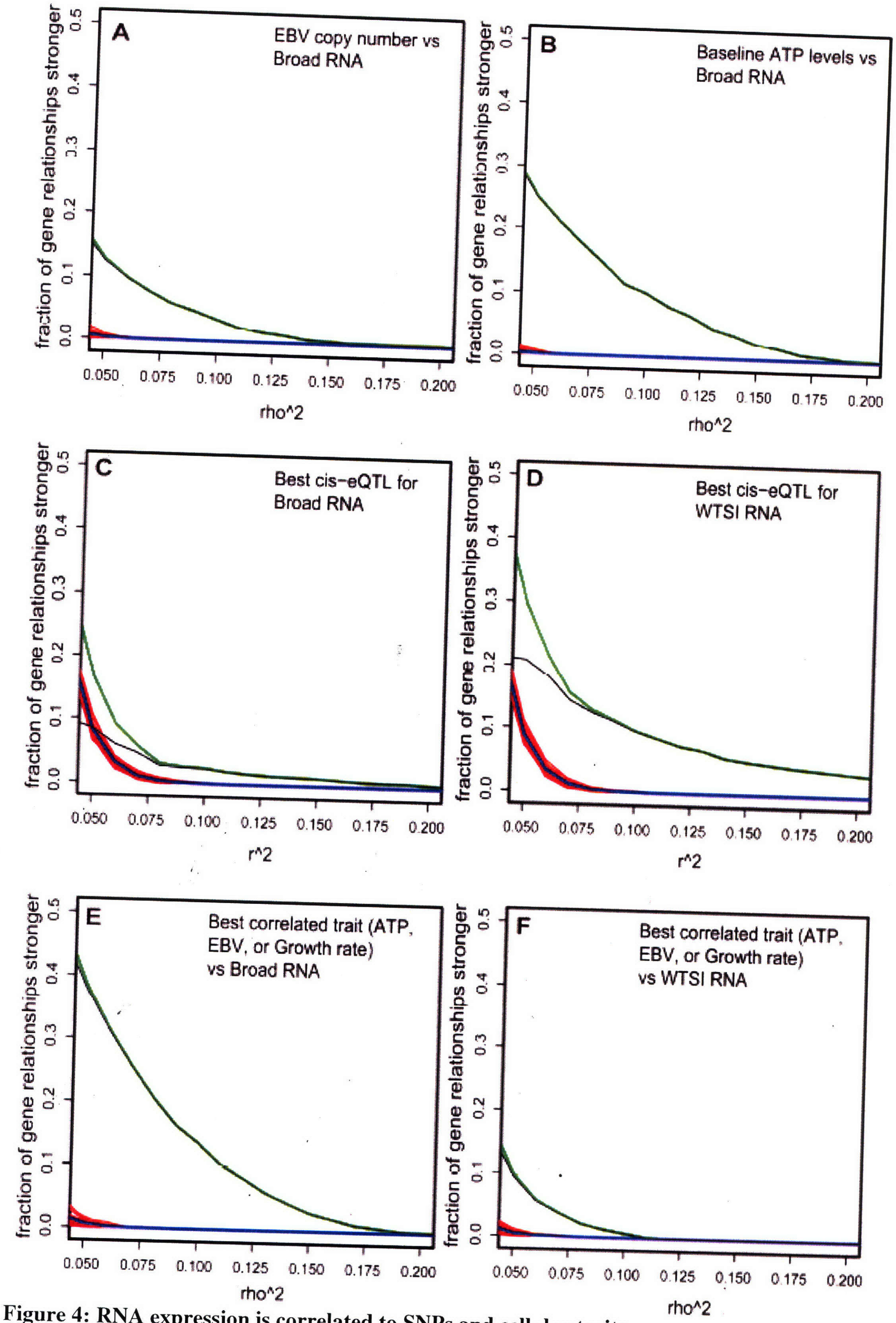


Figure 4: RNA expression is correlated to SNPs and cellular traits.
198 unrelated individuals were whole-genome RNA profiled on the Affymetrix platform at the Broad Institute ("Broad RNA") and independently on the Illumina platform at WTSI ("WTSI RNA"). The 1000 "best-measured" genes identified in Figure 3 were tested for correlation to SNPs and cellular traits.

(A) For each tested gene, Broad RNA expression levels were rank-correlated to copy numbers of EBV, as determined by quantitative PCR. The correlation was expressed as rho^2 and curves representing distributions of the rho^2 values are plotted. The green curve is the observed distribution of EBV-RNA correlations. The red curves represent 20 permuted distributions. The blue curve is the average of permuted distributions. The black curve is the difference between observed and permuted values and thus a lower bound (see Methods) of the fraction of genes correlated to EBV at a given rho^2. Plot shows that ~15% of expressed genes have >5% of their (rank) variance in expression explained by EBV levels.

(B) For each tested gene, Broad RNA expression levels were correlated to baseline ATP levels determined by measuring Celltiter glo in mock-treated wells in the drug response assays. Curves representing the distribution of rho^2 values were plotted for the tested genes as in (A). Plot shows that >25% of expressed genes have >5% of their variance in expression explained by ATP levels.

(C) For each tested gene, Broad RNA expression levels were correlated to all SNPs with MAF>10% within a 0.15Mb window around the gene, using the HapMap phase II data. Curves representing the distribution of the largest r^2 value was plotted for each tested genes as in (A). Plot shows that >9% of genes have >5% of their variance in expression explained by SNPs in the Broad RNA dataset.

(D) For each tested gene, Sanger RNA expression levels were correlated to all SNPs with MAF>10% within a 0.15Mb window around the gene, using the HapMap phase II data. Curves representing the distribution of the strongest r^2 value was plotted for each tested genes as in (C). Plot shows that >20% of genes have >5% of their variance in expression explained by SNPs in the WTSI RNA dataset.

(E) For each tested gene, Broad RNA expression levels were correlated to EBV, growth rate, and relative ATP, and the strongest observed correlation among the 3 phenotypes was plotted. Strikingly, plot shows that >40% of genes have >5% of their variance in expression explained by one of these covariates.

(F) For each tested gene, WTSI RNA expression levels were correlated to EBV, growth rate, and relative ATP, and the strongest observed correlation among the 3 phenotypes was plotted. Strikingly, plot shows that the effect of covariates in (E) is observable even when looking at a completely separate expression experiment, performed independently of covariate collection.

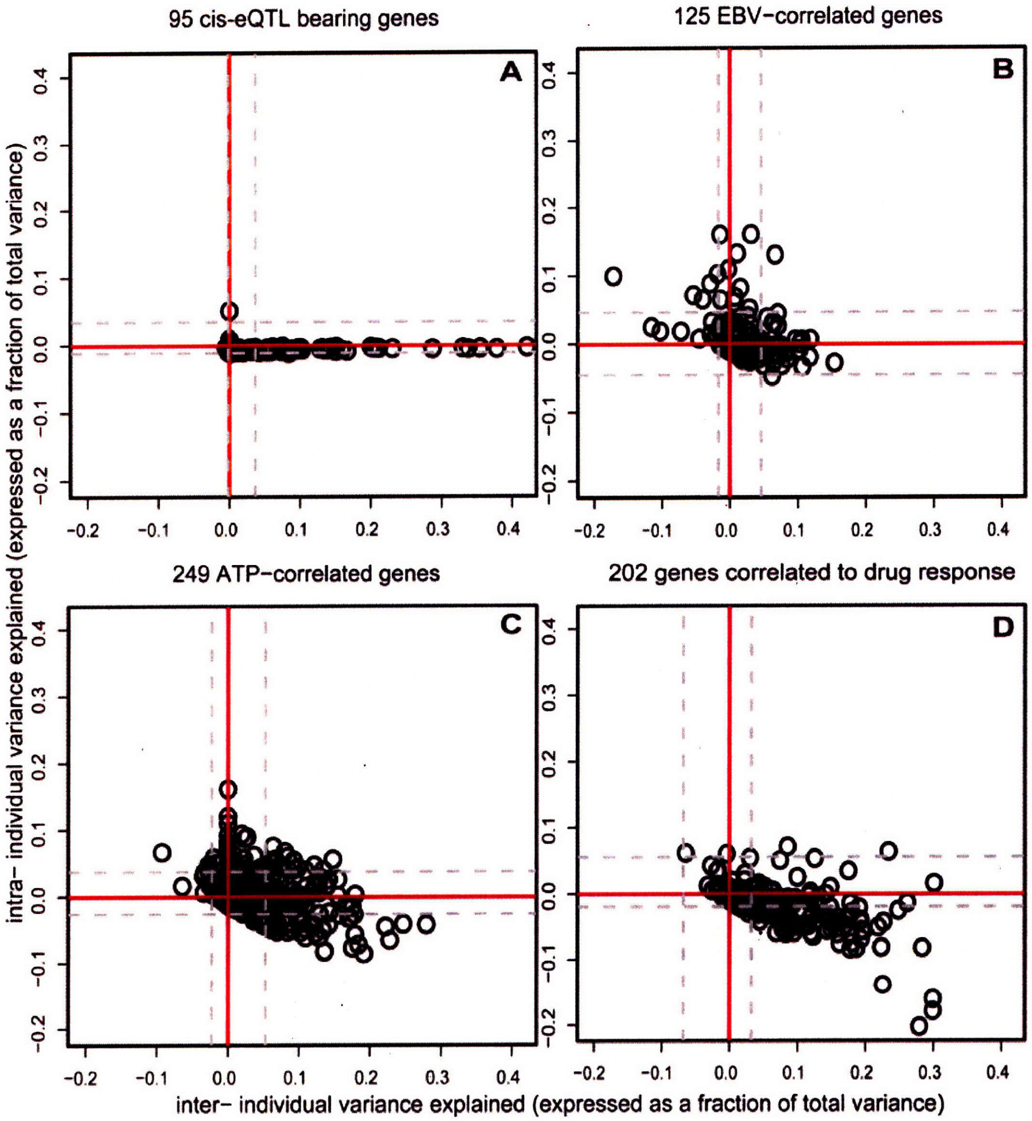


Figure 5: Contributions by eQTLs, EBV, and ATP to inter- and intra -individual variation in RNA expression levels, and contributions by RNA expression levels to inter- and intra- individual variation in drug response.
Total variance for each of the 1000 "best-measured" genes was separated into inter- and intra- individual variance components (see Methods) using expression data from the 49 unrelated individuals measured twice at the Broad Institute on the Affymetrix platform.

(A) 95 genes with eQTLs that explained >10% of expression variance (FDR<10%) in the WTSI dataset were selected (to maximize eQTL detection power) and the SNP genotype was included in the variance components model of the gene to "account" for its effect. -1 times the change in each variance component is plotted for each gene. As expected, the plot shows that that SNPs (which remain fixed across experiments) only explain inter-individual variation in expression. Grey dashed lines indicate the inter- and intra- 2.5% and 97.5%-tiles of the distribution of variance component change estimates when the entire analysis is repeated on a permuted dataset.

(B) 125 genes correlated to EBV at rho^2>.05 (FDR<10%) were selected and the EBV measurement was included in the variance components model of the gene to "account" for its effect. -1 times the change in each variance component is plotted for each gene. The plot shows that EBV can contribute to inter-individual differences in gene expression that persist across experiments, intra-individual fluctuation in gene expression between experiments, or both, depending on the gene in question. Grey dashed lines are as in (A).

(C) 249 genes correlated to ATP at rho^2>.05 (FDR<10%) were selected and the ATP measurement was included in the variance components model of the gene to "account" for its effect. -1 times the change in each variance component is plotted for each gene. The plot shows that ATP can contribute to inter-individual differences in gene expression that persist across experiments, intra-individual fluctuation in gene expression between experiments, or both, depending on the gene in question. Grey dashed lines are as in (A).

(D) 202 "drug-response correlated" genes were defined as in Figure 6. The expression of each gene was incorporated in a variance components model of the assigned drug response EC50 to examine the contribution of the gene to its strongest correlated drug. -1 times the change in the variance components of drug response is plotted for each gene, showing that it is mostly the *inter*-individual differences in gene expression that are correlated to cell line drug response. Grey dashed lines are as in (A).

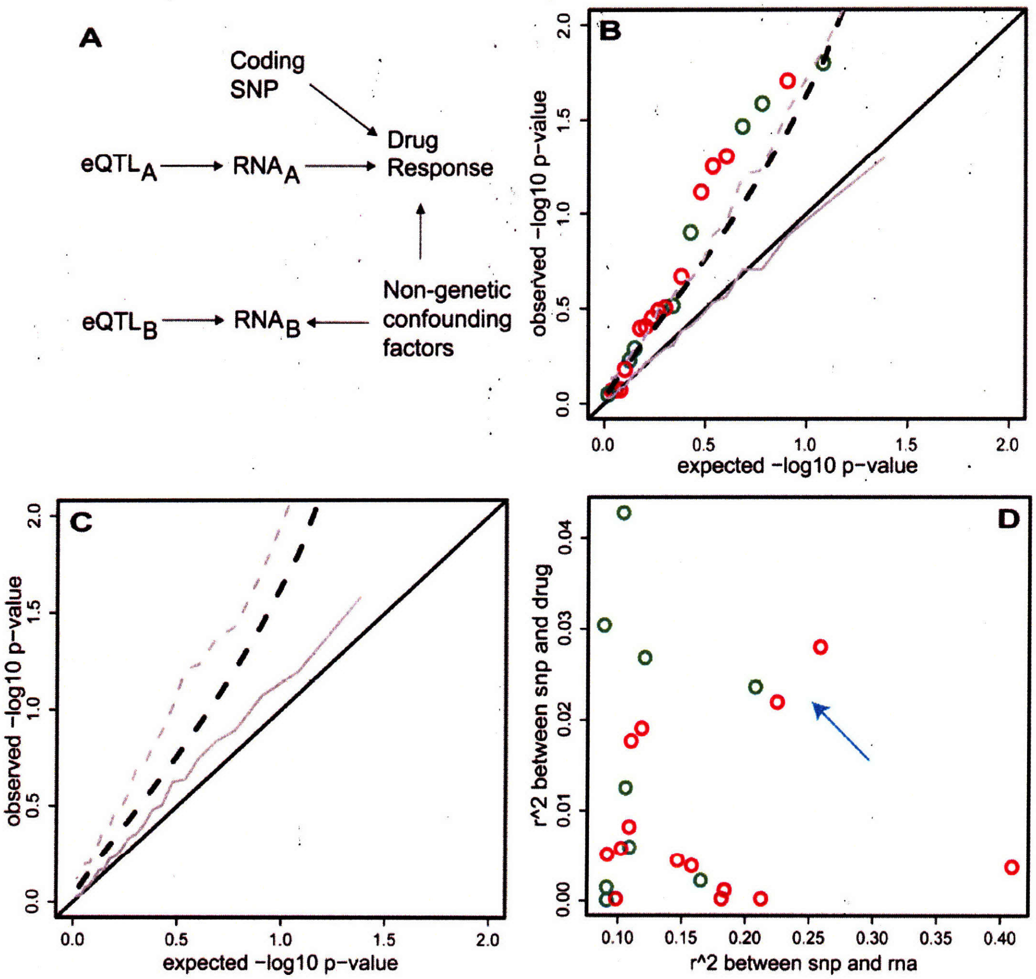


Figure 6: Effect of cis-eQTLs in drug-response correlated genes on drug-response
The 198 unrelated individuals were ranked by RNA expression value for each of the 1000 "best-measured" genes. These individuals were then ranked by response (growth/ATP- corrected EC50) to each of the 5 assayed drugs. Rank-correlations (spearman's rho) were computed for each gene-X-drug pair (1000x5) and the drug with the strongest correlation to a given gene was "assigned" to that gene. The 202 genes

whose strongest drug correlations exceeded rho^2=.05 (FDR<10%) were taken as "drug-response correlated" genes. If such a gene also had a cis-eQTL that explained at least 8% (FDR<10%) of its variance, the tuple SNP-RNA-Drug was considered in the foregoing panels. We considered 23 tuples (14 derived using WTSI RNA dataset + 9 derived using the Broad Institute RNA dataset)

(A) Diagram of different classes of influences on drug response and RNA levels. Solid arrows represent correlation due to a causal effect.

Coding SNPs have direct (non-RNA mediated) effects on drug response by altering protein function. No SNPs of this class were found at genome-wide significance in our GWAS scan.

Changes in RNA_A directly cause changes in drug response. An
eQTL for one of these RNAs (i.e. $eQTL_A$) is thereby causal for drug response.

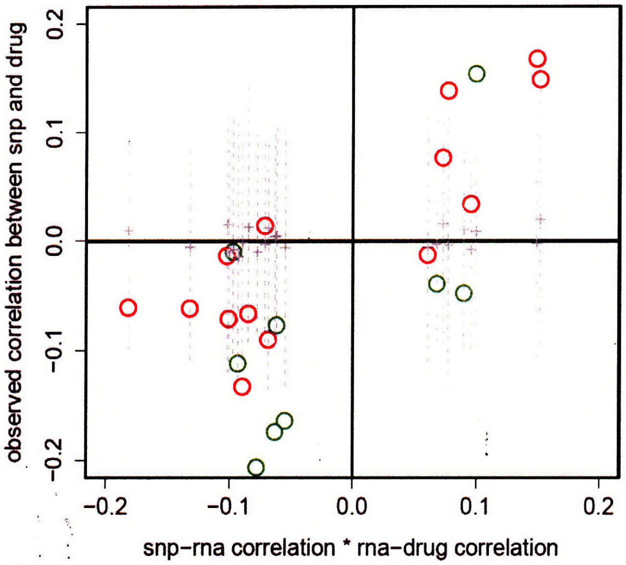
Non-genetic confounding factors simultaneously influence RNA_B levels and drug response and changes in RNA_B do not cause changes in drug response (this is the expected scenario for most RNAs). If levels of these RNAs are affected by eQTLs, these eQTLs cannot be causal for drug response (the $eQTL_B$ effect on RNA_B and the RNA_B relationship to drug response are orthogonal).

Our goal in mapping true RNAs and eQTLs causal for drug response is to distinguish $eQTL_A$ and RNA_A from $eQTL_B$ and RNA_B.

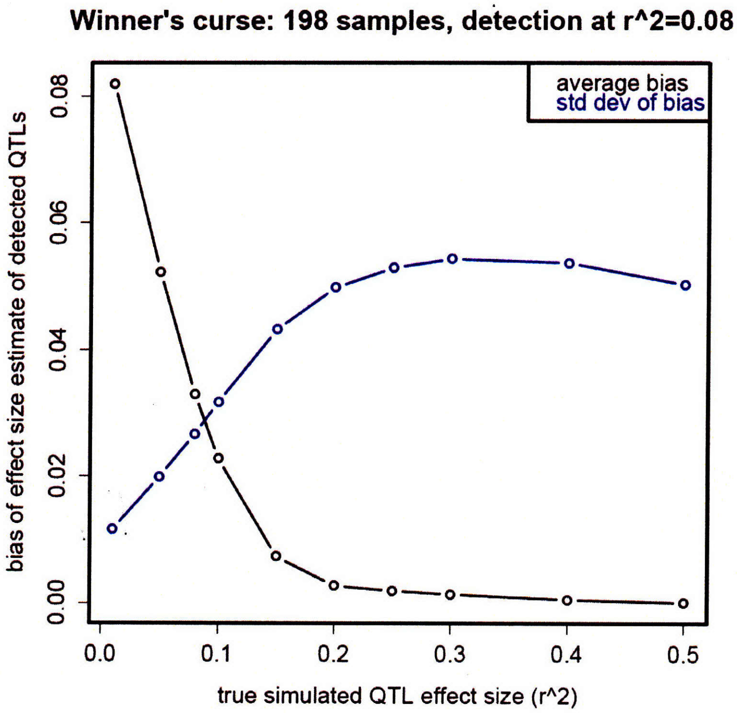
(B) For each tuple (WTSI – red, Broad – green) the drug response was regressed against the eQTL SNP genotype. P-values are plotted as open circles against their expectation under the null. Black solid line indicates the theoretical flat uniform distribution expected under the null and black dashed line is the p=.05 one-sided significance threshold for deviation from the null. Grey lines show equivalent null parameters, but derived from a simulated dataset with the same SNP/RNA/Drug variances and independent SNP-RNA/RNA-Drug pairwise covariances as the real 23 tuples. Plot shows that the observed p-value distribution for drug-response regressed against RNA eQTL SNPs deviates significantly from null, suggesting that some RNA relationships to drug response and corresponding SNP relationships to RNA are not orthogonal. (and thus possibly causal)

(C) For each tuple, simulated datasets were created with the same SNP/RNA/Drug variances and RNA-Drug pairwise covariance as the real 23 tuples, but with the real SNP-RNA covariances replaced by r^2=0.05. Then, only those simulations where the observed SNP-RNA association exceeded r^2=0.08 were used to plot the median and p=.05 SNP-Drug p-value distributions as in (B) (again, grey solid and grey dashed lines, respectively). Black lines also as in (B). Plot shows that "winner's curse" in eQTL discovery leads to an inflation of SNP-Drug associations, in the absence of any real causal relationship between RNA and Drug response (and thus, SNP and Drug).

(D) For each tuple (WTSI – red, Broad – green), the correlation between SNP and RNA is plotted against the correlation between SNP and Drug. Most increased association between SNP and Drug response comes from the weaker eQTLs, while most of the stronger eQTLs have no association to drug response, consistent with the winner's curse phenomenon displayed in (C). Additionally, 3 interesting tuples emerge that are both relatively strong SNP-RNA and SNP-Drug response associations, indicated by the light blue arrow.



Supplementary Figure S1: Direction of SNP-Drug response association
For each tuple (WTSI – red, Broad – green) in Figure 6, the product of the correlation (r) between SNP and RNA and the correlation (rho) between RNA and Drug is plotted against the correlation (r) between SNP and Drug. Black lines separate the plot into the 4 quadrants. Gray dotted lines show the expected distribution of associations between SNP and Drug under the "null" model simulated in Figure 6B. Plot shows that the direction of association SNP-Drug response tends toward the direction predicted from the directions of the SNP-RNA and RNA-Drug correlations (i.e. if the major allele drives the RNA up and more RNA makes the cell-line more sensitive to drug, then the major allele should make the cell-line more sensitive to drug). This tendency would not be expected by chance alone.



Supplementary Figure S2: Winner's curse in eQTL discovery
Simulations were performed to demonstrate that effect sizes of weaker eQTLs are overestimated, on average. Specifically, for effect sizes (r^2) between 0.01 and 0.50, 100,000 datasets of 198 values each (corresponding to the sample size of the analysis in Fig 6) were simulated from a bivariate normal distribution with mean=(0,0), variances=(1,1) and covariances=sqrt(effect size). Datasets with observed correlation (r^2)>0.08 were then considered: For each simulated effect sizes, the average difference (bias) between the observed and simulated effect size is plotted, together with the standard deviation of the distribution of differences. Plot shows that weaker eQTLs are usually over-estimated, even for true effects that are above the detection threshold. On the other hand, estimates of effect sizes of stronger eQTLs are unbiased, on average.

Supplementary Table 1

Compound	rho	rho^2	pval
MTX	0.98831	0.976749	<2.20e-16
5FU	0.97217	0.945105	<2.20e-16
6MP	0.94395	0.891043	<2.20e-16
Simva	0.91629	0.839582	<2.20e-16
Saha	0.85846	0.736949	<2.20e-16
velcade	0.88812	0.788753	<2.20e-16
rapa	0.91968	0.845818	<2.20e-16

Supplementary Table 2

Compound	rho	rho^2	pval
MTX	0.82	0.67	1.2E-15
5FU	0.63	0.39	6.6E-08
6MP	0.66	0.43	9.7E-09
Simva	0.39	0.15	2.0E-03
Saha	0.71	0.50	2.0E-10

Supplementary Table 3

Relative Drug Response	**MTX**	**6MP**	**5FU**	**Simva**	**Saha**
MTX	*rank correlation*	2.20E-16	2.20E-16	1.14E-08	0.00977
6MP	0.78	*below*	2.20E-16	7.44E-05	0.00127
5FU	0.78	0.61	*diagonal*	4.68E-14	1.26E-06
Simva	0.36	0.25	0.46	*p-value*	1.97E-04
Saha	0.16	0.2	0.3	0.23	*above*

pvalues <.001 marked in red

Correlation to Growth Rate of relative drug response

Drug	**MTX**	**6MP**	**5FU**	**Simva**	**Saha**
Rank Correlation	-0.34	-0.31	-0.3	-0.15	0.04
P-value	1.86E-07	3.21E-06	7.61E-06	0.031	0.603

pvalues <.001 marked in red

Supplementary Table 4

Growth-corrected EC50 Drug Response	**MTX**	**6MP**	**5FU**	**Simva**	**Saha**
MTX	*rank correlation*	0.02	1.74E-09	2.27E-11	0.174
6MP	0.14	*below*	0.0165	0.973	0.0524
5FU	0.37	0.15	*diagonal,*	0.00863	1.54E-04
Simva	0.41	0.0021	0.17	*p-value*	0.118
Saha	0.09	0.12	0.24	0.1	*above*

pvalues <.001 marked in red

Correlation to Growth Rate of EC50/Growth-corrected drug responses

Drug	**MTX**	**6MP**	**5FU**	**Simva**	**Saha**
Rank Correlation	-0.1	0.06	-0.24	-0.14	0.06
P-value	0.14	0.4	0.00034	0.0356	0.345

pvalues <.001 marked in red

Supplementary Table 5

Growth and ATP-Corrected Drug Response	**MTX**	**6MP**	**5FU**	**Simva**	**Saha**
MTX	*rank correlation*	0.041	0.008	2.06E-08	0.703
6MP	0.15	*below*	0.003	0.973	0.052
5FU	0.19	0.22	*diagonal,*	0.291	0.006
Simva	0.40	0.00	0.08	*p-value*	0.118
Saha	-0.03	0.12	0.20	0.10	*above*

pvalues <.001 marked in red

References

1. McCarthy, M.I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-69 (2008).
2. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
3. Easton, D.F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-93 (2007).
4. Haiman, C.A. et al. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* **39**, 638-44 (2007).
5. Plenge, R.M. et al. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. *N Engl J Med* **357**, 1199-209 (2007).
6. Saxena, R. et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-6 (2007).
7. Cheung, V.G. et al. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* **33**, 422-5 (2003).
8. Cheung, V.G. et al. Genetics of quantitative variation in human gene expression. *Cold Spring Harb Symp Quant Biol* **68**, 403-7 (2003).
9. Cheung, V.G. & Spielman, R.S. The genetics of variation in gene expression. *Nat Genet* **32 Suppl**, 522-5 (2002).
10. Dermitzakis, E.T. & Stranger, B.E. Genetic variation in human gene expression. *Mamm Genome* **17**, 503-8 (2006).
11. Monks, S.A. et al. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet* **75**, 1094-105 (2004).
12. Stranger, B.E. & Dermitzakis, E.T. The genetics of regulatory variation in the human genome. *Hum Genomics* **2**, 126-31 (2005).
13. Stranger, B.E. et al. Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1**, e78 (2005).
14. Dausset, J. et al. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575-7 (1990).
15. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
16. Cheung, V.G. et al. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365-9 (2005).
17. Morley, M. et al. Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-7 (2004).
18. Stranger, B.E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).
19. Stranger, B.E. et al. Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).
20. Dixon, A.L. et al. A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202-7 (2007).
21. Graham, R.R. et al. A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nat Genet* **38**, 550-5 (2006).

22. Moffatt, M.F. et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470-3 (2007).
23. Correa, C.R. & Cheung, V.G. Genetic variation in radiation-induced expression phenotypes. *Am J Hum Genet* **75**, 885-90 (2004).
24. Dolan, M.E. et al. Heritability and linkage analysis of sensitivity to cisplatin-induced cytotoxicity. *Cancer Res* **64**, 4353-6 (2004).
25. Watters, J.W., Kraja, A., Meucci, M.A., Province, M.A. & McLeod, H.L. Genome-wide discovery of loci influencing chemotherapy cytotoxicity. *Proc Natl Acad Sci U S A* **101**, 11809-14 (2004).
26. Duan, S. et al. Mapping genes that contribute to daunorubicin-induced cytotoxicity. *Cancer Res* **67**, 5425-33 (2007).
27. Huang, R.S. et al. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc Natl Acad Sci U S A* **104**, 9758-63 (2007).
28. Huang, R.S. et al. Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am J Hum Genet* **81**, 427-37 (2007).
29. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
30. Akey, J.M., Biswas, S., Leek, J.T. & Storey, J.D. On the design and analysis of gene expression studies in human populations. *Nat Genet* **39**, 807-8; author reply 808-9 (2007).
31. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
32. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
33. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**, 177-82 (2003).
34. Emilsson, V. et al. Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).
35. Hill, A.V. The Combinations of Haemoglobin with Oxygen and with Carbon Monoxide. I. *Biochem J* 7, 471-80 (1913).
36. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).
37. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
38. Pfaffl, M.W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* **29**, e45 (2001).
39. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).
40. Team, R.D.C. *R: A Language and Environment for Statistical Computing*, (Vienna, Austria, 2008).

Chapter 7: Summary and Discussion

The “ideal world” for a human geneticist, where we have the full sequence of every individual under study, where we know the expression, post-translational modification, and localization of every gene in every cell, and where we can safely perform *in-vivo* perturbation experiments and measure all relevant changes in physiologic state is still many years away. In the mean-time we must rely on useful and practical proxy genotypes and phenotypes to enable discoveries in human biology using the genetic mapping approach. This thesis presents one important class of proxy genotypes, those that capture most common genetic variation, as well as an evaluation, refinement and application of proxy phenotypes offered by a commonly used *in-vitro* model, the lymphoblastoid cell-line.

To develop a useful proxy for the full sequence, we investigated the selection and analysis of tag SNPs for genome-wide association studies: specifically, the relationship between investment in genotyping and statistical power. Do pair-wise or multi-marker methods maximize efficiency and power? To what extent is power compromised when tags are selected from an incomplete resource such as HapMap? We examined these questions using genotype data from the HapMap-ENCODE project, association studies simulated under a realistic disease model, and empirical correction for multiple hypothesis testing. Perhaps most impressively, we found that simply picking 1 SNP every 10kb at random from a complete map (in the CEU panel) provides >80% of the power to detect common causal variants as does genotyping all common variants directly. Even simple single-marker tagging/testing approaches can provide ~90% power with 1 tag per 5kb or less. We then demonstrated a haplotype-based tagging method that uniformly outperforms single-marker tests, and methods for prioritization that dramatically increase tagging efficiency further. Examining all observed haplotypes for association — not just those that proxy for a known SNP — increases power to detect rare causal alleles, at a cost of reduced power for common causal alleles. Power is surprisingly robust to the completeness of the reference panel from which tags are selected.

Practically, tag SNPs are chosen from data in one population sample (i.e. HapMap) and then deployed in another sample (the disease study cohort). Thus, it is also important to know how well tags picked using the HapMap capture the variation in other samples. To address this, we collected dense data uniformly across the four HapMap population samples and eleven other population samples. We picked tag SNPs using genotype data we collected in the HapMap samples and then evaluated the effective coverage of these tags in comparison to the entire set of common variants observed in the other samples. We simulated case-control association studies in the non-HapMap samples under the same disease model of modest risk as above, and observed little loss in power. These results demonstrated that the HapMap DNA samples can be used to select proxy genotypes for genome-wide association studies in many samples around the world.

As many investigators will rely on commercially available "whole genome" arrays for their GWAS, we also evaluated the extent to which the sets of SNPs contained on three widely used products capture common variation across the genome. We found that the majority of common SNPs are well captured by these products either directly or through linkage disequilibrium. (this was particular true for technologies designed with tagging principles in mind) We then explored analytical strategies that utilize HapMap data to improve power of association studies conducted with these fixed sets of markers, and show that inclusion of our specific haplotype tests in association analysis can increase the fraction of common variants captured by 25% to 100%. This idea was later expanded by others into a full multi-point imputation approach that is now a mainstay of association studies world-wide. [1]

Given that most common variation is thus being captured and that much of the variation is correlated, it became important to estimate the number of independent tests implicit in a complete common-variant GWAS for purposes of simple and accurate (i.e. Bonferroni) multiple-testing correction and evaluation of nominal association results. Applying our simulation framework to the problem, we found that GWAS for common (MAF>5%) variants correspond to ~500,000 independent tests in out-of-Africa and ~1,000,000 tests in African samples, implying a reasonable genome-wide significance threshold for declaring association of roughly 5e-8. This threshold was confirmed by others [2] (as well as predicted by Risch and Merikangas in 1996!) and is now commonly accepted and widely used.

Taken together, our work shows that it is possible and practical to comprehensively determine the common variant contribution to disease by employing a well-chosen set of genotypes and analytical approaches that can proxy for all or nearly common genetic variation in the genome. Although our investigation was primarily focused on variants with MAF>5%, nothing in principle restricts the applicability of our findings to variants with lower frequency. All that is required are sufficiently sized cohorts for the discovery of variants and assessment of LD. With the commencement of the 1000 Genomes Project [3], these resources will soon be in hand, setting up an even bigger wave of medical genetic discoveries. The tools and methods we devised, together with the theoretical/empirical understanding of LD-based genome-wide association we've obtained, will no doubt facilitate this advance.

Identification of a genetic variant in a genome-wide scan is, of course, only the beginning of the process to elucidate its function, place it in context of other genetic and non-genetic variation, and understanding the relevant physiology. As it is important to elucidate the functional effects of a mutation "proximally" to the mutated site, geneticists must seek to obtain phenotypes from the relevant cells, tissues, or systems. Because it is difficult to obtain (all but routine clinical) phenotypes from human subjects, and because most informative perturbation experiments in human beings are impossible, geneticists often rely on *in-vitro* or *ex-vivo* models to proxy for *in-vivo* genetics and physiology.

Lymphoblastoid cell lines (LCLs), originally collected as renewable sources of DNA, are now being used as one such model to study genotype-phenotype relationships in human cells. These cell lines have been used to search for genetic variants that are associated with drug response as well as with more basic cellular traits such as RNA levels. In setting out to extend such studies by searching for genetic variants contributing to drug response, we observed that phenotypes in LCLs were, in our lab and others, significantly affected by experimental confounders (i.e. *in vitro* growth rate, metabolic state, and relative levels of the Epstein-Barr virus used to transform the cells). Even after correcting for these confounders, we did not find any SNPs associated with genome-wide significance to drug response. We also evaluated whether incorporating RNA expression levels (and eQTLs) in the analysis could increase power to detect such effects. As previously shown, cis-acting eQTLs were detectable for a sizeable fraction of RNAs and baseline levels of many RNAs predicted response to drugs. However, we found only limited evidence that SNPs influenced drug response through their effect on expression of RNA; nevertheless, it remains possible that integrating SNPs, RNA, and drug response can identify novel pharmacogenetic variation mediated by RNA. Efforts to use LCLs to map genes underlying cellular traits will require great care to control experimental confounders, unbiased methods for interpreting such multi-dimensional data, and much larger sample sizes than have been applied to date.

Our experience with lymphoblastoid cell-lines is likely general and applies to other *in-vitro* models, including those that don't yet exist. Indeed as it becomes possible to reprogram any adult cell-type into stem-cells[4,5] and then re-differentiate these into any tissue, the applicability of proxy models to genetic mapping will only increase. Imagine, for instance, the impact that fully functioning beta-cells with specific genotypes at diabetes risk-conferring sites would have on untangling pathophysiology and designing targeted interventions. The challenge, as our study demonstrated, will be to detect and control for extraneous sources of phenotypic variation so that the subtle genetic differences we are interested in can emerge.

In all branches of science, from astronomy, to geology, to chemistry, reality can rarely be directly observed. Progress is often determined by the characterization and application of indirect measurements that reflect (or proxy for) some useful aspect of reality. This thesis presents proxy measurement (genotypes and phenotypes) for one subfield of biology, medical genetics. At the same time, it lays out the principles and methods for the development of further proxy measurements of this class. Together, it is hoped that our findings will enable discoveries today and facilitate medical genetic research yet to come.

References

1. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
2. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* **32**, 227-34 (2008).
3. Kaiser, J. DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science* **319**, 395 (2008).
4. Aoi, T. et al. Generation of Pluripotent Stem Cells from Adult Mouse Liver and Stomach Cells. *Science* (2008).
5. Jaenisch, R. & Young, R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* **132**, 567-82 (2008).

Appendix

The following license was obtained to re-use the material in Chapter 5:

July 11, 2008
Roman Yelensky
514 Lowell Ave.
Newton, MA 02460
yelensky@mit.edu

Dear Mr. Yelensky:
RE: Your June 04, 2008 request for permission to reuse Genetic Epidemiology 2008 May; 32(4):381-5. This material will appear in your forthcoming thesis in print and/or on a password-protected websites <http://dspace.mit.edu> to be published by Massachusetts Institute of Technology.

1. Permission is granted for this use, except that if the material appears in our work with credit to another source, you must also obtain permission from the original source cited in our work.
2. Permitted use is limited to your edition described above, and does not include the right to grant others permission to photocopy or otherwise reproduce this material except for versions made for use by visually or physically handicapped persons. Up to five copies of the published thesis may be photocopied by a microfilm company.
3. Appropriate credit to our publication must appear on every copy of your thesis, either on the first page of the quoted text, in a separate acknowledgment page, or figure legend. The following components must be included: Title, author(s) and /or editor(s), journal title (if applicable), Copyright © (year and owner). Reprinted with permission of Wiley-Liss Inc. a subsidiary of John Wiley & Sons Inc.
4. This license is non-transferable. This license is for non-exclusive English print rights and microfilm storage rights by Massachusetts Institute of Technology only, throughout the world. This License does not extend to selling our content in any format. *For translation rights, please reapply for a license when you have plans to translate your work into a specific language*
5. Posting of the Material shall in no way render the Material in the public domain or in any way compromise our copyright in the Material. You agree to take reasonable steps to protect our copyright not limited to, providing credit to the Material as specified in Paragraph 3 above.

Sincerely,
Brad Johnson
Permissions Assistant
201.748.6786
201.748.6008 (fax)
bjohns@wiley.com