Spring 5-15-2010

Cell Culture Models of Genetic Variation

Joshua T. Witten
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Molecular Cell Biology

Dissertation Examination Committee:
Barak Cohen, Chair
James Cheverud
Justin Fay
Jeffrey Gordon
Tim Schedl
James Skeath

CELL CULTURE MODELS OF GENETIC VARIATION

by

Joshua T. Witten

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2010

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Cell Cultures of Genetic Variation

by

Joshua Turner Witten

Doctor of Philosophy in Biology and Biomedical Sciences (Molecular Cell Biology)

Washington University in St. Louis, 2010

Professor Barak Cohen, Chairperson

Studying genetic variation presents a dilemma. While the genetic variation of greatest interest is that causing variation in traits and disease risk in natural populations, natural populations have characteristics that make them challenging to study. In this work, I have assessed the use of cell culture methods as a solution to some of these challenges. In particular, I studied genetic variation in the budding yeast *Saccharomyces cerevisiae* that was generated by selection in the lab as a model for natural genetic variation. I have found that even simplistic selection programs in the laboratory, including the use of chemical mutagenesis to introduce genetic variation, can be used to rapidly generate genetic variation with the same characteristics as that observed in natural populations of budding yeast.

I also explored the use of human-derived lymphoblastoid cell lines as source of genetic variation that eliminates some of the most challenging problems that arise from the use of humans as research subjects. In addition to the ethical limitations, there are also severe technical limitations to the study of human subjects, not least of which is the difficulty of direct experimentation to confirm hypotheses.

I found that lymphoblastoid cell lines are a reliable experimental system in which phenotypic variation, at the cellular level, primarily represents differences between lines, a significant portion of which is due to additive genetic variation. Due to the growth of publicly available genotype data, these lines can be used to locate genetic variants with phenotypic effects by linkage-association mapping. In addition to the shared database resources, cell lines are amenable to distribution from central repositories, suggesting that cell culture could form the basis of a community resource for the study of human genetic variation.

While cell culture methods have share weaknesses with traditional genetic model systems, the use of a variety of cell culture approaches, including microorganisms and human-derived cell lines, represents an important, complementary approach to the investigation of genetic variation both for basic, mechanistic questions and for understanding the genetic causes of diversity in human phenotypes.

ACKNOWLEDGMENTS

The work with the budding yeast *Saccharomyces cerevisiae* described in Chapter Two was only possible through to the generous donation of strains by Petek Çakar and Uwe Sauer. I am also grateful to Mark Johnston and John McCusker for providing plasmids, and Justin Fay and members of the Cohen Lab for advice and discussion. This work was supported by grants from the American Cancer Society (RSG-06-039-01-GMC) and the National Science Foundation (0543156).

The work with human lymphoblastoid cell lines described in Chapter Three was only possible through the generous donation of the lines by Rose Veile and Michael Lovett. I would particularly like to thank Bill Nolan for the use of equipment and Kim Lorenz for helping maintain flow cytometer. Aldi Kraja and Michael Province performed narrow-sense heritability estimates. Victoria Brown-Kennerly, Ilaria Mogno, and Michael White provided helpful suggestions and discussions, as did the other members of the Cohen Lab. Funding for this project was provided by the Children's Discovery Institute (MC-II-2008-102).

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

CHAPTER ONE: GENERAL INTRODUCTION

INTRODUCTION

Gregor Mendel's discovery of the Laws of Inheritance was, to a degree, serendipitous. The choice of model system (peas) and phenotype (round/wrinkled seeds, among others) was essential. That system presented Mendel with a situation ideally suited for the discovery of his Laws of Inheritance – a phenotype controlled by a single gene with two alleles, one of which was completely dominant (the round allele) over the other. It would, however, been very easy for Mendel to pick a system that did not have the necessary characteristics for the discovery of the Laws of Inheritance, because most of the traits he could have chosen (e.g., height, yield, drought resistance, etc.) are complex. This discussion serves not to downplay Mendel's brilliance, but to highlight the fact that no amount of genius would have given the world his Laws of Inheritance without the right phenotype in the right model system. The right system is integral to address any question in science. The main goal of the work in this thesis was to investigate cell culture methods as systems to address questions about genetic variation.

Genetics is the study of how DNA polymorphism determines phenotypic variation. In classical genetics phenotypes are connected to single large effect mutations, which often identify key genes in biological pathways. In quantitative genetics phenotypes are modeled as being the output of alleles at multiple segregating loci. Population genetics is the study of how these alleles are distributed within and between populations.

As the ability to determine genotype becomes simpler, the study of genetic variation could be augmented by model systems that can represent natural genetic variation, are amenable to high-throughput approaches, and allow methods to experimental test genotype-phenotype association hypotheses. A particularly challenging area that might benefit from model systems to complement traditional practices is research of natural genetic variaton. In this work, I have assessed the utility of studying selection in the laboratory using the budding yeast *Saccharomyces cerevisiae* as a model system for natural variation. I have also examined the utility of lymphoblastoid cell lines (LCLs) as a model system for the study of human genetic variation. Background specific to each model system follows below

A description of my research on selection in *S. cerevisiae* as a model for natural genetic variation appears in Chapter Two. A description of my research on LCLs as a model for human genetic variation appears in Chapter Three. General conclusions and discussion appear in Chapter Four.

CELL CULTURE AS A MODEL FOR NATURAL GENETIC VARIATION

The fundamental distinction between studies of laboratory and naturally occurring strains of a model organism is the degree to which the environment can be controlled. The variables affecting phenotype in laboratory strains can be limited and controlled by the researcher. This allows variables (e.g., genotype, environment, etc.) to be manipulated individually, which is ideal for the study of mechanism. Naturally occurring strains, on the other hand, are the product of multivariate processes, which have not been observed,

resulting in multifactor differences between strains. The variables affecting phenotype between natural strains are complex and have not been controlled.

There are, however, reasons why laboratory researchers might wish to step away from pristine laboratory strains to examine variation in natural strains. Most phenotypic variation observed in nature is continuous, genetically complex, and dependent on the environment. The progressive goal of genetics is to build models that predict phenotype from knowledge of genotype and environment with increasing accuracy. While these models require a mechanistic basis, they also require an understanding of the consequences of small variation in their parameters that are frequently below the resolution of mechanistic studies. Variation in natural strains provides examples of this type of variation (GERKE *et al.* 2009; GERKE *et al.* 2006). The study of natural variation complements mechanistic study of laboratory strains.

In addition, naturally occurring strains are reservoirs of variation and novelty in potentially useful traits. The search for useful variation in nature is a major component of agricultural science and pharmaceutical research.

One weakness in studying natural strains is that they are inherently anecdotal. These studies tend to assume that the experimental tools and accumulated knowledge base from the laboratory strain of the species (or closely related species) are directly applicable to the natural strains. Because each example of natural variation occurs in an uncontrolled, multivariate environment and is determined by chance events, it can be very difficult to identify the environmental variables driving that variation and to re-examine the history of that variation in detail. The multivariate, contingent evolutionary

history of natural strains makes it difficult, if not impossible, to systematically explore the effects of individual variables on phenotype.

Laboratory selection as a model system: Can the systematic control of variables in laboratory strains be brought together with the continuous, complex variation of natural strains in a model system for the study of naturally occurring genetic variation? The use of selection in the laboratory to generate variation with the characteristics of natural variation is a potential solution.

Artificial selection (breeding determined by phenotype of interest) has a long and successful history in the agricultural breeding of plants and animals. William Ernst Castle, one of the earliest geneticists, used directional selection in hooded rats to demonstrate transgressive segregation and disproved the generality of the pure line genetic concept (CASTLE 1951). Since then, laboratory selection has been used to study a wide variety of areas of interest in genetics and evolutionary biology, such as the fitness effects of mutation (DE VISSER and LENSKI 2002; DE VISSER and ROZEN 2006; ESTES *et al.* 2004; HEGRENESS *et al.* 2006; KASSEN and BATAILLON 2006; LENSKI and TRAVISANO 1994; OSTROWSKI *et al.* 2005a; SILANDER *et al.* 2007), the role of genome rearrangements (DUNHAM *et al.* 2002), the effects of haploidy and diploidy (PAQUIN and ADAMS 1983a; ZEYL *et al.* 2003), the frequency of parallelism versus convergence (BUCKLING *et al.* 2003; COHAN and HOFFMANN 1986; COOPER *et al.* 2003; HERRING *et al.* 2006; LENSKI and TRAVISANO 1994), the effects of asexual and sexual lifestyles (DE VISSER and ROZEN 2006; GRIMBERG and ZEYL 2005a), and evolutionary change in gene expression (COOPER *et al.* 2003; FEREA *et al.* 1999; PELOSI *et al.* 2006; RIEHLE *et al.*

2003; RIFKIN *et al.* 2005). Selection in the laboratory has also been used extensively to investigate questions in evolutionary theory and quantitative genetics.

Laboratory selection of microorganisms as a model system: Microorganism cell culture is particularly advantageous for selection experiments in the laboratory. Cell culture systems using microorganisms generally have the following advantages: large population size, short generation time, amenability to high-throughput techniques, long-term storage, environmental control, experimental tractability, and small genome size.

Microorganisms routinely achieve extremely large population sizes under standard cell culture conditions ($n > 10^9$ individuals/mL of culture for *E. coli* and $n > 10^7$ individuals/mL of culture for *S. cerevisiae*) than those achievable for multicellular organisms. Large population sizes increase the efficiency of selection relative to drift (FALCONER 1989) making it tractable to detect small changes in fitness.

The short generation time of microorganisms increases the number of generations that can be observed. The observation of evolutionary dynamics over long periods of the relevant time scale (i.e., generations) allows the detection of small fitness changes and a thorough exploration of the the genotype-phenotype landscape.

Large population size and short generation time makes microorganisms suitable for culturing and assaying using high-throughput methods, including growth in 96-, 384-, and 1024-well plates. Because each well is an isolated environment, dense cell culture on plates increases sample sizes and the number of variables that may be tested.

In addition, standard, laboratory microorganism model systems can be maintained in long-term frozen storage, allowing samples from different generational time points during the evolution experiment to be directly compared.

The environment of microorganisms - the variables determining the nature and intensity of selection - can be strictly controlled in the laboratory. The use of defined media allows the researcher to control resource availability and the presence of chemical challenges. The availability of both liquid and solid phase media allows the researcher to control the spatial structure of the environment. The use of incubators allows the researcher to control temperature, humidity, and gas mixture. Strict, well documented cell culture protocols have the additional advantage of controlling variables that have not been formally considered by the researcher. Environmental control removes uncertainty about the variables driving selection and explicit testing of evolutionary questions.

Microorganims are also experimentally tractable. Depending on the species of microorganism, extensive experimental options for both phenotyping assays and direct hypothesis testing exist. For example, allelic conversion via homologous recombination in yeast can be used to test the phenotypic effect of a single nucleotide polymorphism (SNP) (GERKE *et al.* 2009). The experimental tools available for data collection and hypothesis testing in microorganisms are more extensive and more easily applied to large sample sets than those available in multicellular organisms. Among those experimental tools is the expansion of sequencing capacity with "next-generation" sequencing technology, which makes sequencing entire genomes to identify genetic variation between lines possible, especially for microorganisms due their small genome sizes. The

ability to experimentally confirm hypotheses, such as identifying the phenotypic effects of a SNP, is critical for understanding the molecular basis of genetic variation.

The paradigmatic laboratory selection experiment using microorganisms is Richard Lenski's long-term experimental evolution in *Escherichia coli*. At last report, twelve lines of *E. coli* split from a single clone (i.e., all lines had the same genotype at the beginning of the experiment with variation introduced by mutation accumulation) were passaged through over 40,000 generations of asexual growth in glucose-limited medium for nearly twenty years (BARRICK *et al.* 2009).

By way of comparison, a similar long-term selection in a multicelluar organism (*Zea mays*) has been operated at the University of Illinois since 1896. In that time, the divergent selection for kernel oil concentration on two lines has progressed through approximately 100 generations (LAURIE 2004). The Lenski selection experiment using *E. coli* has passed six times as many lines through 400 times as many generations in one-fifth the time of a long-term selection in multicellular organisms (BARRICK *et al.* 2009; LAURIE 2004).

The Lenski selection experiment, which has dominated both the literature and intellectual space of experimental evolution, illustrates the advantages of microorganism cell culture as a model system for laboratory selection experiments. The large population sizes and number of generations allowed very small fitness effects to be detected (DE VISSER and LENSKI 2002; OSTROWSKI *et al.* 2008). The number of generations allowed the dynamics of the evolutionary process to be observed, such as the slowing of the response to selection due to the number and effect size of available, potentially adaptive

mutations decreasing as the fitness landscape becomes more completely explored over evolutionary time (ARJAN *et al.* 1999; DE VISSER and LENSKI 2002). Samples from each line were taken and stored periodically, allowing samples from different generational time points in the evolutionary timeline to be directly compared (BLOUNT *et al.* 2008; DE VISSER and LENSKI 2002) and questions of historical contingency to be investigated (BLOUNT *et al.* 2008). This experiment has used the ability to control the *E. coli* environment to maintain constant cell culture conditions for nearly 20 years (BARRICK and LENSKI 2009) and to make defined changes in individual nutrients to examine the pleiotropic effects of adaptive mutations (OSTROWSKI *et al.* 2005b; OSTROWSKI *et al.* 2008). The experimental tractability of *E. coli* has permitted the identification of molecular difference between the lines(WOODS *et al.* 2006), and the testing of those differences for their effects on fitness (OSTROWSKI *et al.* 2008). Most recently, the newest generation of high-throughput sequencing technology has begun to be applied to these lines to detect sequence changes during the evolutionary process (BARRICK and LENSKI 2009; BARRICK *et al.* 2009; JEONG *et al.* 2009; OSTROWSKI *et al.* 2008; STUDIER *et al.* 2009). The Lenski selection experiment has exploited many of the advantages of microorganisms for selection experiments to make wide-ranging contributions to the study of evolution.

Beyond the contributions of the literature devoted to this resource, the Lenski selection experiment stands as a rigorous, documented, empirical example of microevolution (i.e., changes in gene frequency due to selection, drift, mutation, and migration) in action, which alone has tremendous epistemological and rhetorical value.

The choice of *E. coli* as the model system and the design of the Lenski selection experiment impose certain constraints on the question that it can address. *E. coli* is an asexual, haploid making *E. coli* an unsuitable model system for the investigation of some of the most difficult issues in evolutionary biology. As an asexual organism, *E. coli* is not accommodated within the traditional textbook definition of speciation (FRASER *et al.* 2009) proposed by Ernst Mayr, "groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups" (MAYR 1942), although others have argued that the definitional species concept problem is not significant (DE QUEIROZ 2007). *E. coli* is not a suitable model system to investigate macroevolution (i.e., speciation) under the most widely accepted species concepts.

Asexuality also makes *E. coli* an unsuitable model system for the investigation of the relative advantages of asexual and sexual reproduction. Similarly, the fact that *E. coli* is a facultative haploid throughout its life cycle (except in the interval between completion of DNA replication and cytokinesis in mitosis) makes it an unsuitable model system to compare haploidy to diploidy – a distinct question from asexual/sexual comparisons as reproductive strategy does not necessarily constrain ploidy. The simplicity of *E. coli* is a strength, but it also prohibits the study these issues.

An interesting feature of the Lenski selection experiment is that no variables are manipulated in the selection. All twelve lines had the same ancestral genotype and were cultured in the same environment. Random effects, like mutation and experimental error, have introduced all variation between the lines. In this context, the number of lines under study is a limitation. In order to rigorously study the effects of such random processes,

one would prefer far more than twelve samples. Because this long-term evolution experiment did not use high-throughput cell culture methods, the number of samples that could be studied was limited by practical considerations. The decision to passage these lines through as many generations as possible dedicated experimental resources, which might have been used to explore environmental variables or expand the number of lines, to maximizing the number of generations that might have been used to explore environmental variables or expand the number of lines. In addition, using mutation accumulation to introduce genetic variation increases the time required for beneficial mutations to occur. The use of a chemical mutagen to introduce variation at the beginning of the experiment could reduce the time required for beneficial variant to arise in the population as well as permit the study of the evolutionary dynamics of highly variable populations. The Lenski selection experiment has been extremely productive, but has also been limited by the choice of model organism and the study design.

Saccharomyces cerevisiae as a model system for natural genetic variation: As an alternative to the *E. coli* model system used in the Lenski selection experiment, the budding yeast *S. cerevisiae* is a potential model system for natural genetic variation that has the standard advantages of microorganism cell culture as well as a number of additional characteristics that are useful for the study of genetic variation.

S. cerevisiae is a eukaryote making it a relevant model system for natural genetic variation in eukaryotes, due to the presence of eukaryotic specific features (e.g., mitochondria, etc.) that might be affected by genetic variation. *S. cerevisiae* can reproduce asexually or sexually, allowing the effects of reproductive strategy to be

investigated. *S. cerevisiae*, in the laboratory, can be maintained in either a haploid or a diploid state, allowing the effects of ploidy to be investigated independent of reproductive strategy. While questions regarding reproductive strategy (GRIMBERG and ZEYL 2005b; ZEYL *et al.* 2005a; ZEYL and BELL 1997), ploidy (PAQUIN and ADAMS 1983a; PAQUIN and ADAMS 1983b; ZEYL 2004; ZEYL 2005; ZEYL *et al.* 2003), and eukaryotic features (TAYLOR *et al.* 2002; ZEYL *et al.* 2005a) were inaccessible in *E. coli*, they are active areas of research in *S. cerevisiae*.

In addition, *S. cerevisiae* have many advantages that are relatively specific to yeast. Homologous recombination is efficient, which allows the phenotypic effects of SNPs to be tested by direct experimentation (DEUTSCHBAUER and DAVIS 2005; GERKE *et al.* 2009; GERKE *et al.* 2006). *S. cerevisiae* also enjoys extensive resource sharing in the research community. In addition to the reference genome (GOFFEAU *et al.* 1996), genomic sequences are available for several closely related species (CLIFTEN *et al.* 2003) and the phylogenetic relationships between species and strains have been described (LITI *et al.* 2009). Furthermore, resources like the yeast deletion collection, in which all non-essential genes have been knocked-out (GIAEVER *et al.* 2002), has simplified discovery and genetic manipulation. The ability to identify causal molecular variation is combined with the community resources to allow that molecular variation to be understood in the context of the organism.

Like *E. coli*, S. *cerevisiae* has been proposed as model system to investigate genetic variation and evolutionary dynamics (ZEYL 2000; ZEYL 2006). It has been used to study the effects of reproductive strategy (GRIMBERG and ZEYL 2005b; ZEYL and BELL

1997; ZEYL *et al.* 2005b), mutation effect size and frequency (ADAMS *et al.* 1985; GRIMBERG and ZEYL 2005b; TAYLOR *et al.* 2002; ZEYL and DEVISSER 2001; ZEYL *et al.* 2001), ploidy(PAQUIN and ADAMS 1983a; PAQUIN and ADAMS 1983b; ZEYL 2004; ZEYL 2005; ZEYL *et al.* 2003), mitochondrial defects (TAYLOR *et al.* 2002; ZEYL *et al.* 2005b), environment (GRIMBERG and ZEYL 2005b), and epistasis (PAQUIN and ADAMS 1983b; ZEYL *et al.* 2005b).

In order to determine if laboratory selection of *S. cerevisiae* can be used as a model system for natural genetic variation, the characteristics of natural genetic variation in *S. cerevisiae* require description. The genetic basis of variation in sporulation efficiency in *S. cerevisiae* has been determined at the resolution of single nucleotides for several strains (DEUTSCHBAUER and DAVIS 2005; GERKE *et al.* 2009; GERKE *et al.* 2006), including a natural isolates (GERKE *et al.* 2009; GERKE *et al.* 2006). These studies have concluded that variation in sporulation efficiency is due to a small number of variants (three (DEUTSCHBAUER and DAVIS 2005) and four (GERKE *et al.* 2009)) at a small number of loci (five in total (DEUTSCHBAUER and DAVIS 2005; GERKE *et al.* 2009; GERKE *et al.* 2006)) with large additive effects (DEUTSCHBAUER and DAVIS 2005; GERKE *et al.* 2009; GERKE *et al.* 2006) and a smaller, but significant, contribution from epistatic interactions between loci (GERKE *et al.* 2009). Based on this single, well researched trait, natural genetic variation in *S. cerevisiae* can be described as being controlled by a few loci with large additive effects and some epistatic interactions between loci.

One study estimated that a small number (n=2) of *S. cerevisiae* strains produced by 2000 generations of selection for growth in low-glucose media produced strains with

similar characteristics to the genetic variation seen in natural strains. They estimated that growth rate in the derived strains was controlled by only a few loci of large effect (ZEYL 2005); supporting the use of *S. cerevisiae* strains derived from laboratory selection as a model system for the study of natural genetic variation. The small number of strains tested in this study (ZEYL 2005) underscores the need for cell culture and phenotyping techniques with increased throughput.

Laboratory selection experiments with microorganisms generally use the spontaneous accumulation of mutations to introduce genetic variation, which is generally viewed as more natural than using mutagens. As mentioned above, this increases the number of generations needed for beneficial mutations to arise in the experimental population. An alternative method is to use chemical or radiation mutagenesis to introduce genetic variation, which would reduce the time needed to observe the phenotypic effects of introduced genetic variation in the experimental population, as well as allow the study of both isogenic and genetically variable populations. Although there may be concern that the known biases of different mutagens – it should be noted that naturally occurring mechanisms of mutation are also biased – may affect the distribution of mutation effect sizes, there is no theoretical reason to expect that mutagens have a different distribution of mutation effect sizes compared to normal sources of variation, only a greatly increased rate of mutation.

Is the approach of artificially produced variation a viable method for the study of how genetic variation controls the complex phenotypic variation observed outside the laboratory? Does variation produced in the laboratory have similar genetic characteristics

to that observed outside the laboratory? Is the phenotype is controlled by a few genes of relatively large, not completely additive effects with evidence of many, small modifying variants? In experiments described in Chapter Two, I have used lines of the budding yeast *S. cerevisiae* derived by selection in the laboratory to address these questions and develop a quantitative, high-throughput phenotyping assay that should be applicable to most yeast cell culture environmental conditions. A quantitative, high-throughput cell culture model of genetic variation would be an important tool in the addressing the complex relationship between genotype and phenotype in complex traits.

MODEL SYSTEMS FOR THE STUDY OF HUMAN GENETIC VARIATION

Studying basic genetic mechanisms allows the researcher the freedom to select a model system ideally suited to the question. When the question is not a basic research one – how does genetic variation control phenotypic variation, but is directed toward understanding the causes of specific phenotypic variation in a specific organism – which genetic variants control the phenotypic variation of a specific trait, the researcher is generally restricted to pursuing the question in that organism. Understanding variation in maize crop qualities requires research in maize (LAURIE *et al.* 2004). Understanding variation in pig carcass quality requires research in pigs (HEUVEN *et al.* 2009). Understanding the genetic variation that affects both normal human variation and disease risk requires research using human subjects.

While humans are the species of greatest practical interest to human researchers, they are a difficult experimental subject. Complementary approaches to the use of human subjects could be a useful in moving the field forward. Below, I will first discuss some of

the challenges in human genetics and how cell culture, and LCLs in particular, addresses these challenges, and then I will discuss considerations in the use of LCLs as a model system in the context of the existing literature.

Challenges in human genetics: *Homo sapiens* is, at the same time, an extremely challenging species to research and the species in which we are most interested in studying. The difficulties with human studies have come under renewed focus recently as the results of genome-wide association studies continue to accumulate, but have so far failed to explain the majority of the genetic variation in common human diseases and traits.

Efforts to connect genetic variation with quantitative variation in phenotype via associations date back to at least 1953 (AIRD *et al.* 1953) – less than a decade after the discovery that DNA is the genetic molecule (AVERY *et al.* 1944) and the same year as the discovery of the structure of DNA (WATSON and CRICK 1953). The goal is to associate molecular variation in DNA with quantitative phenotypic variation. With the completion of the Human Genome Project in combination with high-density microarray technology, it became possible to genotype millions of SNPs simultaneously allowing the search for phenotype associated genetic variants across the entire genome without making any prior assumptions (BODMER and BONILLA 2008). The identification of molecular variation has become increasingly accessible, but associating it with phenotypic variation remains a significant challenge.

The technology allows an individual to be genotyped for millions of common SNPs (2007) (usually SNPs with a minor allele frequency > 5% (BODMER and BONILLA

2008)) rapidly and for relatively little cost. It is an ideal system for testing the common disease-common variant (CDCV) hypothesis. The CDCV hypothesis, as originally described by Reich and Lander, postulates that the genetic variation controlling the common versions of diseases, like diabetes, consists of common genetic variants at multiple loci, whose combined effects determine disease risk (REICH and LANDER 2001), and can be extended to other, non-disease quantitative traits. The Genome Wide Association Study (GWAS) design is an approach to genome-phenotype association studies based on this concept.

While genome-phenotype association studies have been enormously successful in identifying quantitative trait loci (QTL), they have not produced significant ability to predict phenotype from genotype for disease risk or any other human trait. At the level of the individual, the associated variants appear to have very small effects on phenotype (i.e., low penetrance). The odds ratios for most genome-phenotype association identified QTL are between 1.2 and 1.5 (BODMER and BONILLA 2008). As a result, the number of variants identified has, to some degree, supplanted predictive power as the metric by which success is measured. The largest genotype-phenotype association study group, the Wellcome Trust Case-Control Consortium (WTCCC) reports:

> The WTCCC has substantially increased the number of genes known to play a role in the development of some of our most common diseases and has to date identified approximately 90 new variants across all of the diseases analyzed. As well as confirming many of the known associations,

> some 28 in total, the WTCCC has also identified many novel variants that affect susceptibility to disease. (CONSORTIUM 2010)

It is this low predictive power that has precipitated the controversy over private genotyping services and slowed the progress toward personalized medicine.

Does the lack of predictive power accurately reflect the relationship of genetic variation in humans? The results of genome-phenotype association studies suggest that quantitative traits in humans are controlled by many variants of small effect. This, however, does not agree with experimentally confirmed quantitative trait nucleotides (QTN) in budding yeast, where the sporulation efficiency is controlled by a few QTN of relatively large effect (GERKE *et al.* 2009; GERKE *et al.* 2006). The dramatic differences may reflect real differences in the types of genetic variants that contribute to phenotypic variation between microorganims and multicellular organisms. On the other hand, the difference may reflect technical characteristics of the system under study. The complications inherent in research using human subjects make it difficult to clearly understand these issues.

There are numerous human characteristics that make the species a difficult genetic system in which to work, such as the long developmental period (sexual maturity at 10-15 years of age depending on sex), long gestation period (38 weeks after conception), small generation size (one child per mother per generation), and inability to conduct directed crosses are either obvious or trivial and will not be discussed in detail. More significant complications for the study of human genetic variation include ethics, sample size, environmental variation, population stratification, and limited options for

direct experimentation. Below, I discuss each of these issues, before discussing the benefits a complementary cell culture approach might provide.

Ethics: In the United States of America's National Institutes of Health *Grants Policy Statement* (Part II, Subpart A), the institutional review board evaluation of a human subject research program must meet ethical standards:

> The procedures to be used will minimize risks to subjects . . . Risks to subjects are reasonable in relation to expected benefits, if any, to subjects and the importance of the knowledge that may reasonably be expected to result. (HEALTH 2003)

Researchers are required to both minimize the risk to human subjects and those risks must be relative to the expected benefit by both professional standards of ethics and funding agencies.

Ethical standards complicate experiments in human subjects. The identification of diagnostic, genetic variants could beneficially inform treatment and prophylaxis. To use the example of personalized chemotherapeutic treatment of cancers, studying genetic variation in therapy response in human subjects would necessitate giving chemotherapeutics to otherwise healthy human subjects (MEUCCI *et al.* 2005; SHUKLA and DOLAN 2005; THOMAS *et al.* 2004). The requirement to balance potential benefit with risk poses a potentially severe, but necessary restriction on the number and extent of speculative and research programs that might be pursued, due to the potential for harm, relative to the probability of benefit.

Similarly, understanding variation in susceptibility to infectious disease has the potential to impact the approaches taken to disease prevention and response. This information could inspire new approaches to infectious agents, like human immunodefiency virus (HIV), that have proven difficult targets for prophylaxis and treatment.

Cell culture methods are a potential solution to these ethical limitations. Dangerous treatments and conditions can ethically be tested in renewable cell culture (SHUKLA and DOLAN 2005), without a high expectation of beneficial outcome to balance the perceived risks. As opposed to the concern for subject safety in human studies, the major ethical consideration with cell culture is the handling of information. Individuals require protection from having the results of testing, phenotyping, and genotyping of their derivative cell lines being associated with their personal identity. Major collections of LCLs (such as the International HapMap Project collection at the Coriell Institute of Medical Research) are completely deidentified (THE INTERNATIONAL HAPMAP 2005). Thus, the ethical limitations restricting exposure of these cell lines to potentially dangerous chemicals or infectious agents, and association of their genotypes with observed phenotypic responses are not constrained to the same degree as human subjects..

Sample size: Sample size is a critical issue in quantitative genetics because it is the sample size that directly determines the power of the study to identify the locations of causal genetic variants. There are two aspects of sample size – the number of individuals sampled (N) and the number of times an individual is sampled (n) – that affect studies of

genetic variation. Repeated sampling of an individual gives a more accurate quantification of the individual's phenotype. Sampling more individuals gives a more accurate estimate of the distribution of phenotypic variation (FALCONER 1989). The increase in accuracy from more sampling allows smaller effects to be detected (BODMER and BONILLA 2008). Limited resources generally require any experimental design to balance these aspects of sample size against each other. The total number of assays available must be apportioned between the number of individuals included in the study and the number of times those individuals are assayed (the "*N vs. n Problem*") in order to best support the goals of the research. The substantial costs and difficulties associated with recruiting and assaying human subjects (e.g., specialists required for phenotyping and limited supplies of willing subjects) imposes practical limitations on the amount of sampling that can be done.

The International HapMap Project has provided a list of 3.1 million SNPs at which human subjects may easily be genotyped (2007). Microarrays are used routinely to genotype individuals at hundreds of thousands of SNPs. While these resources have fueled the enormous output of genotype-phenotype association studies, they have also fueled a need for ever-larger sample sizes. Genotype-phenotype association studies in humans have identified many QTL, but virtually all are of very small effect (odds ratios typically between 1.2 and 1.5 (BODMER and BONILLA 2008)). Novel QTL are likely to be within this range or below it, as a common variant with a significantly larger effect than these is unlikely to have escaped detection. Identifying QTL of decreasing effect size will require increasing sample sizes.

Additionally, the large number of hypotheses to be tested (each potential SNP-phenotype association is tested as an independent hypothesis) imposes a substantial multiple testing penalty if large numbers of false positives are to be avoided. The *P*-values required for genome-wide significance can approach 10^{-7} (BODMER and BONILLA 2008). Increasing numbers of SNP-phenotype associations to test require increasing sample sizes. Therefore, the use of next-generation sequencing to identify all SNPs in individuals will actually exacerbate this problem.

While no system can eliminate the fundamental trade-off between more individuals and more replication, cell culture does address both problematic aspects of sample size. Renewable cell culture allows the same individuals to be resampled repeatedly for the same phenotype. In addition, the study of new phenotypes does not require collection of a new cohort, or obtaining informed consent for a new assay. Cell lines can be maintained in long-term frozen storage and easily distributed from centralized repositories. This allows the number of individuals sampled to be cumulatively increased over time as cell lines are added, while allowing the same individual to be tested repeatedly.

Environmental variation: A major complication of almost all research using human subjects is the lack of control over environmental variables. Control over life history prior to a study is non-existent and environmental control during a study is limited by practicality and the rate of subject compliance with study protocols, making it difficult to explicitly control environmental variables or to explicitly quantify their effects.

Stated simply, the total phenotypic variance for a trait in a population is composed of genetic variance and environmental variance:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2,$$

where σ_P^2 is the total phenotypic variance, σ_G^2 is the genetic variance, and σ_E^2 is the environmental variance (FALCONER 1989). In most studies of human genetic variation, only the phenotypic variance can be measured. Control over environmental variables is necessary to make comparisons between individuals that primarily reflect genetic differences. In the absence of control over environmental variables, statistical corrections are used to remove the effects of environmental covariates. Quantification of environmental variance allows the amount of phenotypic variance that can be explained by genetic variance to be defined. Monozygotic/dizygotic twins can be used to separate the genetic and environmental components, but are not frequent enough in the population to replace the use of non-twin individuals. Lack of control of environmental variables decreases a study's power to detect genotype-phenotype associations.

In contrast to human subjects, the use of their derivative cell lines is subject to tight control of environmental variables. Different genotypes can be exposed to exactly the same conditions (e.g., media, incubator, and culturing practices). As a result, the contribution of environmental variance to the phenotypic differences between lines is expected to be small. The ability to control environmental variables and resample individuals allows the genetic and environmental components of the total phenotypic variance to be separated. Data on this topic, however, are generally not reported in the

literature, unless individual experiences do not conform to this expectation (CHOY *et al.* 2008).

Population structure: Population structure is the difference in allele frequencies between populations, due to deviation from the Hardy-Weinberg assumption of random mating, and is a potentially serious complication for any genotype-phenotype association study in natural populations. In case-control studies, unequal representation of populations in the case and control groups can cause spurious associations with the alleles at different frequencies in the two populations. The association is not genotype-phenotype, but genotype-population. Alternatively, the causal variant may be at lower frequency or be more poorly linked to a marker variant in one population, which could cause underestimation of the effect size associated with that marker and the failure to identify a true genotype-phenotype association. Population structure can cause both false positives and false negatives in these studies.

In human studies, population structure takes on additional importance beyond its impact on the ability to correctly identify genotype-phenotype associations. When population structure is minimized, there is risk that research benefits may be population specific - depriving less studied populations of potential advances in diagnosis or treatment. In addition, there are potential sociopolitical implications of any discussion of human population differences. Population structure is both a technical complication and a heated topic in the public sphere.

Because validation of genotype-phenotype associations is technically difficult (further discussion below), one practical standard for validation of a genotype-phenotype

association is replication between populations (CHANOCK *et al.* 2007). This practice increases the confidence that the observed association was not due to the specific population structure of the samples in the original study. Failure to replicate, however, does not necessarily indicate that a genotype-phenotype association is a false positive. The causal variant may be poorly linked to the marker in the second population. The causal variant may not be present in the second population. The success or failure of a genotype-phenotype association replication is not entirely dependent on the accuracy of that association.

Unlike many problems, increasing sample size does not resolve population structure issues. Current research suggests that genetic divergence between human populations is primarily dependent on geographic distance between individuals (NOVEMBRE *et al.* 2008). As a result, larger samples will have greater population structure. Larger sample sizes allow associations with smaller effects to be detected, but also increase the probability of spurious associations due to population structure.

Large association studies generally attempt to compensate for population structure. Cell culture methods have no specific, methodological advantages over human subjects when it comes to population structure. The ability to store lines do allow sample collections to be cumulatively built by adding samples over time from a constrained population. The accumulation of community knowledge, particularly genotypes, would allow the population structure of cell line samples to become progressively better defined over time.

Limited options for direct experimentation: Given an identified genotype-phenotype association, there are, currently, almost no options available to confirm the association by direct experimentation. A hypothetical case where the association leads to an effective disease prophylaxis or treatment may be considered practical evidence for the causal hypothesis, but it does not constitute a direct test. The lack of options for experimental confirmation has led to replication of genotype-phenotype associations between populations becoming the standard for genotype-phenotype association validation in humans (CHANOCK *et al.* 2007). This method of validation is not only susceptible to population structure (as discussed above), but is also vulnerable to the effects of statistical false negatives.

The small effect size of genotype-phenotype associations places many of these associations just over the threshold of statistical significance. Therefore, it would be expected that a number of these associations, even if true positives would not be replicated as significant by chance even if the same population, but not the same individuals, was sampled again. True positives in one test may be false negatives in another test. While replication between populations is an accessible and useful approach to validation, direct, experimental confirmation is necessary to rigorously distinguish true associations from false positives without generating an excessive number of false negatives.

In contrast to human subjects, cell culture does have options available for experimental confirmation of genotype-phenotype associations. LCLs can be transformed, permitting experiments using plasmids or short interfering RNA (siRNA).

Recently, Shukla *et al.* used siRNA in LCLs to confirm the hypothesis that CD44 expression affects sensitivity to carboplatin, a hypothesis generated from genome-wide genotype-phenotype associations (SHUKLA *et al.* 2009). Although this experimental methodology may not be applicable to all phenotypes, this study demonstrates that experimental confirmation is practical in cell culture.

It is also theoretically possible, due to the nature of cell culture phenotypes, to separate an untested set of samples to be used to confirm phenotypic predictions based on significant genotype-phenotype associations. Given the small average effect size and low predictive power of human QTL, this approach may not be practical and I am currently unaware of any applications of this method.

Because cell culture is not constrained by the same ethical limits as human subjects, it is reasonable to suggest that methods to directly test genotype-phenotype associations will develop more rapidly for cell culture than human subjects in the future. *Conclusion*: The complications addressed above are common to natural populations, although not necessarily of the same degree (ethical limitations vary depending on the species, with humans being among the most restrictive), especially when compared to idealized model organisms in the laboratory, or domesticated organisms in controlled settings. The inescapable fact is that genetic variation of greatest interest to humans is human genetic variation. Regardless of the difficulty of the system, the study of human genetic variation will continue to be a high priority and essential avenue of research. Due to these limitations, however, the study of human genetic variation would benefit from

complementary cell culture systems models that might address some of the most serious limitations of human cohort studies.

Cell culture as a model system for human genetic variation: The choice of cell lines to study and phenotypes to measure is of central importance to any study using cell culture to study human genetic variation. There are many available cell lines and collections with characteristics making each suitable for different applications. While the number of phenotypes that might conceivably be measured is effectively infinite, the number suited for measurement in cell culture is more limited. The applicability of the phenotypes chosen for measurement to cell culture may influence the reliability and applicability of experimental results.

One of the major, perceived advantages of the use of cell culture for the study of human genetic variation is the ability to control environmental variables. This factor is a substantial balancing factor to counteract the weaknesses of cell culture as a human analog. Understanding the components of non-genetic variation in phenotype in cell culture is critical to investigating the sources of genetic variation in phenotype.

First, we will discuss the available cell line collections that are available as experimental resources for the study of human genetic variation. Then, we will discuss phenotyping of these collections in other studies. Finally, we will discuss non-genetic variation in these cell lines.

Available cell line collections: LCLs are a particularly attractive cell line for the study of human genetic variation. LCLs are derived from human B cells that have been immortalized by infection with Epstein-Barr Virus (EBV). LCLs grow as non-adherent

cell lines in liquid media making LCLs particularly easy to culture and manpulate. Because EBV immortalization of B cells is a relatively standard and straightforward procedure, LCLs have been routinely isolated from humans across many categories, such as ethnicity and disease, as a source of DNA and for phenotypic characterization. An LCL accurately represents the genotype of the individual from which it was derived. The LCL collections, therefore, represent the genetic variation in the human populations from which they are drawn, making LCL collections a compelling resource for the use of cell culture to investigate human genetic variation.

Due to the presence of shared genotype data, LCL sets for the study of human genetic variation have primarily been drawn from two large, partially overlapping collections: the Foundation Jean Dausset-Centre d'Etude du Polymorphisme Humain (CEPH) (DAUSSET *et al.* 1990) and the International HapMap Project (2003; 2007). The LCLs in both were not collected to test genotype-phenotype associations, but to create a standard, renewable supply of genetic material for the identification and mapping of genetic variation in human populations (2003; 2007; DAUSSET *et al.* 1990; THE INTERNATIONAL HAPMAP 2005). The goal of the International HapMap Project (2003; THE INTERNATIONAL HAPMAP 2005) was to:

> …determine the common patterns of DNA sequence variation in the human genome, by characterizing sequence variants, their frequencies, and correlations between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The project will thus provide tools that will allow the indirect association approach to be

> applied readily to any functional candidate gene in the genome, to any region suggested by family-based linkage analysis, or ultimately to the whole genome for scans for disease risk factors. (2003)

It must be noted that the utility of this goal has been questioned (BODMER and BONILLA 2008; TERWILLIGER and HIEKKALINNA 2006). These collections were designed to represent diversity of human genomic sequence variation.

Because the current use, testing genotype-phenotype associations, was not the original purpose of these collections, the suitability of these collections for this application should be assessed. Their utility will be strongly influenced by the collection design.

Despite similar goals, these collections have different designs (2003; 2007; DAUSSET *et al.* 1990; THE INTERNATIONAL HAPMAP 2005), reflecting the technology available at the time the projects were implemented. The overall CEPH collection is an extremely diverse set of cell lines that includes all 1050 individuals from 51 populations comprising the Human Genome Diversity Cell Line Panel of the Human Genome Diversity Project (CANN *et al.* 2002). A subset of the larger CEPH collection, however, is generally used for testing genotype-phenotype associations. This subset normally consists of LCLs derived from complete, or nearly complete, three-generation pedigrees of Utah residents with large sibships in the final generation (CEPH-UT). These individuals are considered to representative of Americans of European-origin populations (MEUCCI *et al.* 2005), as well as Northern (BAUCHET *et al.* 2007; LAO *et al.* 2008; MEUCCI *et al.* 2005; SMITH *et al.* 2006) and Western European populations (BAUCHET *et al.* 2007; HE *et al.*

2009; LAO *et al.* 2008; MEUCCI *et al.* 2005). At the time of collection (circa 1980 (DAUSSET *et al.* 1990)), neither the human reference genome nor modern genotyping technologies were available. Mapping genetic variation required linkage. Linkage required pedigrees; hence, the pedigree structure of the collection. Publicly available genotypes exist for many CEPH-UT LCLs, but, because most of these data were generated by individual research groups, there is a great deal of variation in the type of data (e.g., microsatellite, SNP, etc.), quality, density, and number of polymorphisms.

A subset of the CEPH-UT LCLs are central to the International HapMap Project collection, but with a different structure. The International HapMap Project collection contains 270 individuals. Ninety individuals, representing 30 father-mother-adult child trios, come from the CEPH-UT collection (CEU). The International HapMap Project collection also contains LCLs from individuals in other world populations including 90 Yoruban in Ibadan, Nigeria (YRI) individuals (30 father-mother-adult child trios), 45 unrelated Han Chinese in Beijing, China (CHB) individuals, and 45 Japanese in Tokyo, Japan (JPT) individuals (2003; 2007; THE INTERNATIONAL HAPMAP 2005). Using microarray-based genotyping, these 270 individuals have been uniformly genotyped to generate a map of 3.1 million single nucleotide polymorphisms (SNPs). The density of SNPs genotypes increases the resolution of QTL mapping and the probability that a marker SNP will be tightly linked to a causal variant (limit case is the marker SNP being the causal variant) increasing the marker's associated effect size and with it the probability of detection. The number of potential genotype-phenotype associations

available for testing, however, imposes substantial multiple hypothesis testing penalties (BODMER and BONILLA 2008).

An estimate of the proportion of the total phenotypic variance that can be explained by genetic variance is critical for interpreting the results of GWA studies. The additive component of the genetic variance is of particular interest, as GWA studies must assume additive genetic models for genotype-phenotype associations (i.e., how much of the additive genetic variance is explained by significant QTL?). Genome-wide testing of interactions between pairs of loci is not possible for any reasonable sample size.

The additive genetic component of phenotypic variance is estimated by the narrow-sense heritability (h^2) of the phenotype within a population (FALCONER 1989). Methods for estimating narrow-sense heritability are based on estimating how well the mean progeny phenotypic value is predicted by the parental phenotype (FALCONER 1989; LYNCH and WALSH 1998), for example by regressing the mean progeny phenotypic values on the mid-parent phenotypic values (the mean phenotypic value of the two parents). The mean progeny phenotypic value will be estimated more accurately as the number of progeny measurements increase, increasing the confidence in narrow-sense heritability estimates.

The three-generation pedigree structure of the CEPH-UT collection is informative for narrow-sense heritability analysis. Each pedigree contains up to three midparent-mean progeny comparisons (not all grandparents are present in every pedigree), two first generation-second generation comparisons and one second generation-third generation

comparison. In addition, the large sibships in the third generation allow the mean progeny phenotypic value for that generation to be estimated with increased accuracy.

In the International HapMap Project collection, the narrow-sense heritability of a phenotype is estimated from thirty CEU and thirty YRI father-mother-adult child trios, which are useful for SNP quality control via Mendelian inheritance. The mean progeny phenotypic value is estimated from the phenotypic value of a single individual. The error inherent in the individual estimates of means, is accommodated by the number of trios assayed; but, in general, the narrow-sense heritability estimates from trios have more associated error and less confidence than those from large sibships (LYNCH and WALSH 1998).

The CEPH-UT collection minimizes complications due to population structure by sampling from a single population (BAUCHET *et al.* 2007; HE *et al.* 2009; LAO *et al.* 2008; MEUCCI *et al.* 2005; SMITH *et al.* 2006). In contrast, the International HapMap Project collection contains samples from four identified populations (2003; 2007; THE INTERNATIONAL HAPMAP 2005). Population structure is a potential, major complication if all samples are pooled in order to increase sample size. On the other hand, the population substructure in the International HapMap Project collection is explicitly defined and can be included in all analyses. Furthermore, inter-population comparisons are possible within the International HapMap Project collection, although the individual population sample sizes are rather small ($45 \leq n \leq 90$) (2003; 2007; THE INTERNATIONAL HAPMAP 2005).

The selection of a sample set will strongly influence the limitations of a research study. Based on the above characteristics, the CEPH-UT collection is suited for studies attempting to define the components of phenotypic variance (e.g., narrow-sense heritability estimates) or testing genotype-phenotype associations for a limited number of phenotypes of great interest (e.g., chemotherapeutic cytotoxicity). The International HapMap Project collection is suited for studies requiring higher resolution QTL (e.g., distinguishing *cis*- and *trans*-expression QTL [eQTL]) or testing a large number of phenotypes simultaneously (e.g., genome-wide transcript abundance) where a substantial per phenotype false negative rate can be tolerated, in order to reasonably constrain the false positive rate.

A 2008 study by Choy *et al.* that examined genetic variation in the International HapMap Project LCL lines for transcript abundance and drug response (CHOY *et al.* 2008) provides an example of the importance of sample selection. This study reported poor repeatability between replicate samples (discussed further below) and failed to identify any drug response QTL (CHOY *et al.* 2008). In particular, they failed to replicate the genotype-phenotype association from a 2004 study by Watters *et al.* (WATTERS *et al.* 2004) for 5-fluorouracil cytotoxicity (5-FOA) (CHOY *et al.* 2008; WATTERS *et al.* 2004).

The failure to replicate may be explained by the poor repeatability and the different number of replicates samples in Choy *et al.* (n=2) (CHOY *et al.* 2008) and Watters *et al.* (n=12) (WATTERS *et al.* 2004). The two studies used different sets of LCLs. Choy *et al.* used the International HapMap Project collection, with all populations combined (CHOY *et al.* 2008). The LCL panel of Watters *et al.* was primarily drawn from

the CEPH-UT collection (WATTERS *et al.* 2004). Population structure may have contributed to the discrepancy. Furthermore, Choy *et al*. lacked the power to detect significant narrow-sense heritability for cell physiology phenotypes when the additive genetic variance accounted for less than half of the total phenotypic variance (significant h^2>0.5) (CHOY *et al.* 2008). Watters *et al*. could detect significant narrow-sense heritabilities down to h^2=0.21 (WATTERS *et al.* 2004). While Choy *et al.* are pessimistic about the use of LCLs as an experimental system for the study of human genetics (CHOY *et al.* 2008), their results were substantially affected by technical difficulties and the choice of a sample set that was not suited to their research goals.

Not only does type of cell line chosen affect the utility of cell culture as a model system for human genetic variation, but the design of the sample set is also critical. *Phenotypes studied in lymphoblastoid cell lines:* Based on paper publications, the use of LCLs as a model system for human genetic variation is increasing (2 papers in 2003, 9 in 2009), likely due to the International HapMap Project. The phenotypes most frequently tested for genotype-phenotype association are transcript abundance (BERGEN *et al.* 2007; CHEUNG *et al.* 2003; CHEUNG and EWENS 2006; CHEUNG *et al.* 2005; CHOY *et al.* 2008; CORREA and CHEUNG 2004; DEUTSCH 2005; DUAN *et al.* 2007; DUAN *et al.* 2009; FORD *et al.* 2001; HUANG *et al.* 2007a; HUANG *et al.* 2008b; JEN and CHEUNG 2003; LI *et al.* 2008; LI *et al.* 2009; MONKS *et al.* 2004; MORLEY *et al.* 2004; PRICE *et al.* 2008; SMIRNOV *et al.* 2009; SPIELMAN *et al.* 2007; STRANGER *et al.* 2007; WANG *et al.* 2009; ZHANG *et al.* 2009), response to chemotherapeutics (BLEIBEL *et al.* 2009; CHOY *et al.* 2008; CLOOS *et al.* 1999; DOLAN *et al.* 2004; DUAN *et al.* 2007; HARTFORD *et al.* 2009;

HUANG *et al.* 2007a; HUANG *et al.* 2008a; HUANG *et al.* 2008b; HUANG *et al.* 2007b; HUANG *et al.* 2007c; LI *et al.* 2008; LI *et al.* 2009; SHUKLA *et al.* 2008; SHUKLA *et al.* 2009; WATTERS *et al.* 2004; WEI *et al.* 1996), and response to ionizing radiation (CORREA and CHEUNG 2004; FORD *et al.* 2001; HARRIS *et al.* 2005; JEN and CHEUNG 2003; SMIRNOV *et al.* 2009); but LCLs have also been used to study a variety of phenotypes including alternative splicing (DUAN *et al.* 2009; KWAN *et al.* 2007; ZHANG *et al.* 2009), infection susceptibility (LOEUILLET *et al.* 2008), ion transport (SCHORK *et al.* 2002), metabolism (ATLAS *et al.* 1976; DUAN *et al.* 2009; SCHORK *et al.* 2002), and micro RNAs (WANG *et al.* 2009). These studies draw on both the CEPH-UT and HapMap collections.

This history does not necessarily indicate widespread community interest in LCLs as a model system. Only a few research groups produce many of the studies (e.g., Cheung (CHEUNG *et al.* 2003; CHEUNG and EWENS 2006; CHEUNG *et al.* 2005; CORREA and CHEUNG 2004; JEN and CHEUNG 2003; MORLEY *et al.* 2004; PRICE *et al.* 2008; SMIRNOV *et al.* 2009; SPIELMAN *et al.* 2007), Dolan (BLEIBEL *et al.* 2009; DOLAN *et al.* 2004; DUAN *et al.* 2007; DUAN *et al.* 2009; HARTFORD *et al.* 2009; HUANG *et al.* 2007a; HUANG *et al.* 2008a; HUANG *et al.* 2008b; HUANG *et al.* 2007b; HUANG *et al.* 2007c; SHUKLA *et al.* 2008; SHUKLA *et al.* 2009; ZHANG *et al.* 2009), and the Mayo Clinic (HARRIS *et al.* 2005; LI *et al.* 2008; LI *et al.* 2009; WANG *et al.* 2009)). In addition, multiple publications from a group may exploit overlapping data sets. For example, 111 of 143 samples in a 2006 study from the Cheung group overlap (CHEUNG and EWENS 2006) with the 355 samples of a 2004 study (MORLEY *et al.* 2004) and 56 samples overlap with the 100 samples of a 2005 study (CHEUNG *et al.* 2005) and the 208 samples of a 2007 study

(SPIELMAN *et al.* 2007); the 2005 study (CHEUNG *et al.* 2005) shares 71 of 100 samples with the 2007 study (SPIELMAN *et al.* 2007). Paper publication rates are not necessarily indicative of the amount of community interest or of the amount of work that has been done.

Because LCLs represent a single cell type and have been immortalized, it is relevant to question the applicability of LCL phenotypes to studies of human phenotypes (SHUKLA and DOLAN 2005). For studies interested in understanding the relationship of genetic variation to phenotypic variation, a relevant phenotype is one where genetic variance was a major component of the total phenotypic variance (discussed in detail below). For studies primarily interested in extracting clinically useful biomedical information from genetic variation, phenotypic relevance is defined by how well LCL phenotypes, and associated genotypes, translate to human phenotypes.

While primary tissue disease models are certainly more specific to the condition of interest (GRUNDBERG *et al.* 2009), LCLs are significantly more accessible. Current research indicates that eQTL identified in LCLs often co-occur in primary tissue samples (BULLAUGHEY *et al.* 2009). A 2008 study examining genetic variation in HIV-1 susceptibility found that the LCL genotype-phenotype association was also associated with disease progression in one of two cohorts of HIV-1 infected human subjects (LOEUILLET *et al.* 2008). Although LCL phenotypes may not translate directly to human phenotypes, the variety and number of LCL phenotypes available for study suggests that LCLs are a useful complement to human subjects, especially as a platform for discovery and hypothesis development.

An under-explored class of phenotypes is localized protein expression levels. Despite proteins being the primary functional unit of the cell and the ready access to monoclonal antibodies specific to B cell surface antigens, few LCL studies of human genetic variation have quantitated protein expression (CHOY *et al.* 2008; LOEUILLET *et al.* 2008). Instead of being treated as a quantitative trait, immunological markers have been used to assess variation in B cell sub-type between lines (CHOY *et al.* 2008). While not as amenable to high-throughput analysis as transcript abundance, localized protein expression levels may integrate multiple sources of variation during localized expression (e.g., transcription, translation, post-translational modification, transport, etc.) with the potential to better represent how genetic variation generates phenotypic variation in humans.

Non-Genetic variance in lymphoblastoid cell lines: The contribution of non-genetic variance to the total phenotypic variance in LCLs is relevant to the study of genotype-phenotype associations, regardless of the biomedical importance of the phenotype. This topic, however, is not commonly addressed in the literature, and, then, only when results do not appear to conform to the expected strict control of environmental variance (CHOY *et al.* 2008).

There are a number of potential sources of non-genetic variance during the phenotyping of LCLs. Although some sources will be specific to individual phenotyping assays, the general points at which non-genetic variation might be introduced are constrained by the cell culture process. The LCL specific components of non-genetic

variance, discussed in greater detail below, are the sampling/immortalization variance, freeze/thaw variance, cell culture variance, and assay variance.

Sampling/Immortalization Variance: The sampling and immortalization process that establishes defined LCL stocks is not only the source of genetic variation between LCLs (samples are taken from different subjects with different genotypes), but it is also a potential source of non-genetic variation. Some sources are non-random in respect to that individual subject, such as life history and age. The CEPH collection, however, has limited information (only age at sampling, sex, and pedigree position) that would allow the impact of these variables on phenotypic variation to be understood. Sampling/immortalizaton variation will cause non-genetic differences between lines that cannot be distinguished from the genetic component of variation without samples that can be used to quantify the contribution of sampling/immortalizaton variance to phenotypic differences between lines.

Unlike additional sources of variation discussed below, both the non-random and random variation introduced during the sampling and immortalization processes become fixed variation between lines during this process. While the non-genetic variation introduced during the collection and immortalization process will contribute to phenotypic variation between lines, it does not contribute to variation within a line. Random variation within a sample is converted into systematic variation between lines.

Because this variation is non-genetic, it will reduce the power of a study to identify genotype-phenotype associations.

Freezing, thawing, and cell culture variance: Following the establishment of LCLs, there are additional sources of variation, once the LCLs have entered the individual research lab. To reduce costs and insure consistency within samples by avoiding repeatedly ordering LCLs from a cell line repository, standard procedure is to freeze multiple aliquots from individual LCLs that can be stored and thawed individually as experiments require. The steps of freezing, thawing, and cell culture that necessarily fill the methodological space between the sampling and immortalization process and the phenotyping assay are all potential sources of non-genetic variation.

Non-genetic variation within lines will affect the reliability of any phenotypic measurements made. There has, however only been one study, to date, that has attempted explicitly examined this question. The authors of this study concluded that LCLs are an unreliable experimental resource (CHOY *et al.* 2008). The reported correlation between two independent aliquots for drug response to three drugs and RNA transcript abundance was both low and inconsistent (ρ=0.39-0.82). Variation in response between aliquots was equal to the variation between drug treatments (CHOY *et al.* 2008). The methods used in this study, however, make it unclear whether the low correlation between aliquots was due to inherent variability in LCLs or the particular methods used.

Non-genetic variation between and within LCLs will obscure the relationship between genetic variation and phenotypic variation by reducing the accuracy of phenotypic measurements.

Assay variance: The final source of non-genetic variation in LCLs is variation in the phenotype assays. Random and non-random effects can affect the precision of the assays.

Random effects include sampling variance in the assay and variation introduced by the experimentalist. Non-random effects include one-time errors, systematic differences between equipment used at different locations, and tendencies of different experimentalists conducting the same assay.

Although assay variation in phenotyping assays is not a major contributor to phenotypic variation, this variation is independent of sources of variation discussed above and will further reduce the ability to associate genetic variation with phenotypic variation by reducing the accuracy of phenotypic measurements. Sufficient sampling can control the random effects. Experimental consistency, including use of the same reagents, single experimentalists, and the same equipment, can control the non-random effects. A recent study suggests that variation between technical replicates is not a major contributor to overall phenotypic variation (CHOY *et al.* 2008).

Conclusion: The numerous challenges facing human genetics require the use of multiple, complementary approaches, and that those approaches represent real human genetic variation. Cell culture as a model system for human genetic variation has obvious weakness, but also has strengths that address some of the challenges of studying human subjects. LCLs are a particularly attractive model system for the study of human genetic variation because they represent human genetic variation, can have a large sample sizes, are easy to use, and are publicly available. My efforts to test the utility of LCLs as a complementary resource for the study of human genetic variation are described in Chapter Three.

LITERATURE CITED

2003 The International HapMap Project. Nature **426:** 789-796.

2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449:** 851-861.

ADAMS, J., C. PAQUIN, P. W. OELLER and L. W. LEE, 1985 Physiological characterization of adaptive clones in evolving populations of the yeast, Saccharomyces cerevisiae. Genetics **110:** 173-185.

AIRD, I., H. H. BENTALL and J. A. ROBERTS, 1953 A relationship between cancer of stomach and the ABO blood groups. Br Med J **1:** 799-801.

ARJAN, J. A., M. VISSER, C. W. ZEYL, P. J. GERRISH, J. L. BLANCHARD *et al.*, 1999 Diminishing returns from mutation supply rate in asexual populations. Science **283:** 404-406.

ATLAS, S. A., E. S. VESELL and D. W. NEBERT, 1976 Genetic control of interindividual variations in the inducibility of aryl hydrocarbon hydroxylase in cultured human lymphocytes. Cancer Res **36:** 4619-4630.

AVERY, O. T., C. M. MACLEOD and M. MCCARTY, 1944 Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. J Exp Med **79:** 137-158.

BARRICK, J. E., and R. E. LENSKI, 2009 Genome-wide Mutational Diversity in an Evolving Population of Escherichia coli. Cold Spring Harb Symp Quant Biol.

BARRICK, J. E., D. S. YU, S. H. YOON, H. JEONG, T. K. OH *et al.*, 2009 Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature **461:** 1243-1247.

BAUCHET, M., B. MCEVOY, L. N. PEARSON, E. E. QUILLEN, T. SARKISIAN *et al.*, 2007 Measuring European Population Stratification with Microarray Genotype Data. The American Journal of Human Genetics **80:** 948-956.

BERGEN, A. W., A. BACCARELLI, T. K. MCDANIEL, K. KUHN, R. PFEIFFER *et al.*, 2007 Cis sequence effects on gene expression. BMC Genomics **8:** 296.

BLEIBEL, W. K., S. DUAN, R. S. HUANG, E. O. KISTNER, S. J. SHUKLA *et al.*, 2009 Identification of genomic regions contributing to etoposide-induced cytotoxicity. Hum Genet **125:** 173-180.

BLOUNT, Z. D., C. Z. BORLAND and R. E. LENSKI, 2008 Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc Natl Acad Sci U S A **105:** 7899-7906.

BODMER, W., and C. BONILLA, 2008 Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet **40:** 695-701.

BUCKLING, A., M. A. WILLS and N. COLEGRAVE, 2003 Adaptation limits diversification of experimental bacterial populations. Science **302:** 2107-2109.

BULLAUGHEY, K., C. I. CHAVARRIA, G. COOP and Y. GILAD, 2009 Expression quantitative trait loci detected in cell-lines are often present in primary tissues. Hum Mol Genet.

CALDWELL, J. G., E. V. PRICE, A. L. SCHROETER and G. F. FLETCHER, 1973 Aortic regurgitation in the Tuskegee study of untreated syphilis. J Chronic Dis **26:** 187-194.

CANN, H. M., C. DE TOMA, L. CAZES, M. F. LEGRAND, V. MOREL *et al.*, 2002 A human genome diversity cell line panel. Science **296:** 261-262.

CASTLE, W. E., 1951 Variation in the hooded pattern of rats, and a new allele of hooded. Genetics **36:** 254-266.

CHANOCK, S. J., T. MANOLIO, M. BOEHNKE, E. BOERWINKLE, D. J. HUNTER *et al.*, 2007 Replicating genotype-phenotype associations. Nature **447:** 655-660.

CHEUNG, V., L. CONLIN, T. WEBER, M. ARCARO, K. JEN *et al.*, 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet **33:** 422-425.

CHEUNG, V. G., and W. J. EWENS, 2006 Heterozygous carriers of Nijmegen Breakage Syndrome have a distinct gene expression phenotype. Genome Res **16:** 973-979.

CHEUNG, V. G., R. S. SPIELMAN, K. G. EWENS, T. M. WEBER, M. MORLEY *et al.*, 2005 Mapping determinants of human gene expression by regional and genome-wide association. Nature **437:** 1365-1369.

CHOY, E., R. YELENSKY, S. BONAKDAR, R. PLENGE, R. SAXENA *et al.*, 2008 Genetic Analysis of Human Traits In Vitro: Drug Response and Gene Expression in Lymphoblastoid Cell Lines. PLoS Genet **4:** e1000287.

CLIFTEN, P., P. SUDARSANAM, A. DESIKAN, L. FULTON, B. FULTON *et al.*, 2003 Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science **301:** 71-76.

CLOOS, J., E. J. NIEUWENHUIS, D. I. BOOMSMA, D. J. KUIK, M. L. VAN DER STERRE *et al.*, 1999 Inherited susceptibility to bleomycin-induced chromatid breaks in cultured peripheral blood lymphocytes. J Natl Cancer Inst **91:** 1125-1130.

COHAN, F. M., and A. A. HOFFMANN, 1986 Genetic divergence under uniform selection. II. Different responses to selection for knockdown resistance to ethanol among Drosophila melanogaster populations and their replicate lines. Genetics **114:** 145-164.

CONSORTIUM, W. T. C. C., 2010 Wellcome Trust Centre for Human Genetics - Research Projects, pp. http://www.well.ox.ac.uk/research-projects-5.

COOPER, T. F., D. E. ROZEN and R. E. LENSKI, 2003 Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli. Proc Natl Acad Sci U S A **100:** 1072-1077.

CORREA, C. R., and V. G. CHEUNG, 2004 Genetic variation in radiation-induced expression phenotypes. Am J Hum Genet **75:** 885-890.

DAUSSET, J., H. CANN, D. COHEN, M. LATHROP, J. M. LALOUEL *et al.*, 1990 Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics **6:** 575-577.

DE QUEIROZ, K., 2007 Species concepts and species delimitation. Syst Biol **56:** 879-886.

DE VISSER, J. A., and R. E. LENSKI, 2002 Long-term experimental evolution in Escherichia coli. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. BMC Evol Biol **2:** 19.

DE VISSER, J. A., and D. E. ROZEN, 2006 Clonal Interference and the Periodic Selection of New Beneficial Mutations in Escherichia coli. Genetics **172:** 2093-2100.
DEUTSCH, S., 2005 Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. Human Molecular Genetics **14:** 3741-3749.
DEUTSCHBAUER, A. M., and R. W. DAVIS, 2005 Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat Genet **37:** 1333-1340.
DOLAN, M. E., K. G. NEWBOLD, R. NAGASUBRAMANIAN, X. WU, M. J. RATAIN *et al.*, 2004 Heritability and linkage analysis of sensitivity to cisplatin-induced cytotoxicity. Cancer Res **64:** 4353-4356.
DUAN, S., W. BLEIBEL, R. HUANG, S. SHUKLA, X. WU *et al.*, 2007 Mapping Genes that Contribute to Daunorubicin-Induced Cytotoxicity. Cancer Research **67:** 5425-5433.
DUAN, S., R. S. HUANG, W. ZHANG, S. MI, W. K. BLEIBEL *et al.*, 2009 Expression and alternative splicing of folate pathway genes in HapMap lymphoblastoid cell lines. Pharmacogenomics **10:** 549-563.
DUNHAM, M. J., H. BADRANE, T. FEREA, J. ADAMS, P. O. BROWN *et al.*, 2002 Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A **99:** 16144-16149.
ESTES, S., P. C. PHILLIPS, D. R. DENVER, W. K. THOMAS and M. LYNCH, 2004 Mutation accumulation in populations of varying size: the distribution of mutational effects for fitness correlates in Caenorhabditis elegans. Genetics **166:** 1269-1279.
FALCONER, D. S., 1989 *Introduction to quantitative genetics*. Longman
Wiley, Burnt Mill, Harlow, Essex, England
New York.
FEREA, T. L., D. BOTSTEIN, P. O. BROWN and R. F. ROSENZWEIG, 1999 Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc Natl Acad Sci U S A **96:** 9721-9726.
FORD, B. N., D. WILKINSON, E. M. THORLEIFSON and B. L. TRACY, 2001 Gene expression responses in lymphoblastoid cells after radiation exposure. Radiat Res **156:** 668-671.
FRASER, C., E. J. ALM, M. F. POLZ, B. G. SPRATT and W. P. HANAGE, 2009 The bacterial species challenge: making sense of genetic and ecological diversity. Science **323:** 741-746.
GERKE, J., K. LORENZ and B. COHEN, 2009 Genetic interactions between transcription factors cause natural variation in yeast. Science **323:** 498-501.
GERKE, J. P., C. T. CHEN and B. A. COHEN, 2006 Natural isolates of Saccharomyces cerevisiae display complex genetic variation in sporulation efficiency. Genetics **174:** 985-997.
GIAEVER, G., A. M. CHU, L. NI, C. CONNELLY, L. RILES *et al.*, 2002 Functional profiling of the Saccharomyces cerevisiae genome. Nature **418:** 387-391.
GOFFEAU, A., B. G. BARRELL, H. BUSSEY, R. W. DAVIS, B. DUJON *et al.*, 1996 Life with 6000 genes. Science **274:** 546, 563-547.

GRIMBERG, B., and C. ZEYL, 2005a The effects of sex and mutation rate on adaptation in test tubes and to mouse hosts by Saccharomyces cerevisiae. Evolution Int J Org Evolution **59:** 431-438.

GRIMBERG, B., and C. ZEYL, 2005b The effects of sex and mutation rate on adaptation in test tubes and to mouse hosts by Saccharomyces cerevisiae. Evolution **59:** 431-438.

GRUNDBERG, E., T. KWAN, B. GE, K. LAM, V. KOKA *et al.*, 2009 Population Genomics in a Disease Targeted Primary Cell Model. Genome Research**:** 1-43.

HARRIS, S. L., G. GIL, H. ROBINS, W. HU, K. HIRSHFIELD *et al.*, 2005 Detection of functional single-nucleotide polymorphisms that affect apoptosis. Proc Natl Acad Sci U S A **102:** 16297-16302.

HARTFORD, C. M., S. DUAN, S. M. DELANEY, S. MI, E. O. KISTNER *et al.*, 2009 Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. Blood **113:** 2145-2153.

HE, M., J. GITSCHIER, T. ZERJAL, P. DE KNIJFF, C. TYLER-SMITH *et al.*, 2009 Geographical Affinities of the HapMap Samples. PLoS One **4:** e4684.

HEALTH, N. I. O., 2003 Grants Policy Statement, pp.

HEGRENESS, M., N. SHORESH, D. HARTL and R. KISHONY, 2006 An equivalence principle for the incorporation of favorable mutations in asexual populations. Science **311:** 1615-1617.

HERRING, C. D., A. RAGHUNATHAN, C. HONISCH, T. PATEL, M. K. APPLEBEE *et al.*, 2006 Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:** 1406-1412.

HEUVEN, H. C., R. H. VAN WIJK, B. DIBBITS, T. A. VAN KAMPEN, E. F. KNOL *et al.*, 2009 Mapping carcass and meat quality QTL on Sus Scrofa chromosome 2 in commercial finishing pigs. Genet Sel Evol **41:** 4.

HILTNER, S., 1973 The Tuskegee Syphilis Study under review. Christ Century **90:** 1174-1176.

HUANG, R. S., S. DUAN, W. K. BLEIBEL, E. O. KISTNER, W. ZHANG *et al.*, 2007a A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proc Natl Acad Sci U S A **104:** 9758-9763.

HUANG, R. S., S. DUAN, E. O. KISTNER, C. M. HARTFORD and M. E. DOLAN, 2008a Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. Mol Cancer Ther **7:** 3038-3046.

HUANG, R. S., S. DUAN, E. O. KISTNER, W. ZHANG, W. K. BLEIBEL *et al.*, 2008b Identification of genetic variants and gene expression relationships associated with pharmacogenes in humans. Pharmacogenet Genomics **18:** 545-549.

HUANG, R. S., S. DUAN, S. J. SHUKLA, E. O. KISTNER, T. A. CLARK *et al.*, 2007b Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. Am J Hum Genet **81:** 427-437.

HUANG, R. S., E. O. KISTNER, W. K. BLEIBEL, S. J. SHUKLA and M. E. DOLAN, 2007c Effect of population and gender on chemotherapeutic agent-induced cytotoxicity. Mol Cancer Ther **6:** 31-36.

JACOBS, J., B. L. GUTHRIE, G. A. MONTES, L. E. JACOBS, N. MICKEY-COLMAN *et al.*, 2006 Homeopathic combination remedy in the treatment of acute childhood diarrhea in Honduras. J Altern Complement Med **12:** 723-732.

JEN, K. Y., and V. G. CHEUNG, 2003 Transcriptional response of lymphoblastoid cells to ionizing radiation. Genome Res **13:** 2092-2100.

JEONG, H., V. BARBE, C. H. LEE, D. VALLENET, D. S. YU *et al.*, 2009 Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). J Mol Biol **394:** 644-652.

KAMPMEIER, R. H., 1972 The Tuskegee study of untreated syphilis. South Med J **65:** 1247-1251.

KAMPMEIER, R. H., 1974 Final report on the "Tuskegee syphilis study". South Med J **67:** 1349-1353.

KASSEN, R., and T. BATAILLON, 2006 Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet **38:** 484-488.

KWAN, T., D. BENOVOY, C. DIAS, S. GURD, D. SERRE *et al.*, 2007 Heritability of alternative splicing in the human genome. Genome Res **17:** 1210-1218.

LAO, O., T. T. LU, M. NOTHNAGEL, O. JUNGE, S. FREITAG-WOLF *et al.*, 2008 Correlation between Genetic and Geographic Structure in Europe. Current Biology **18:** 1241-1248.

LAURIE, C., 2004 The Genetic Architecture of Response to Long-Term Artificial Selection for Oil Concentration in the Maize Kernel. Genetics **168:** 2141-2155.

LAURIE, C. C., S. D. CHASALOW, J. R. LEDEAUX, R. MCCARROLL, D. BUSH *et al.*, 2004 The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. Genetics **168:** 2141-2155.

LENSKI, R. E., and M. TRAVISANO, 1994 Dynamics of Adaptation and Diversification: A 10,000-Generation Experiment with Bacterial Populations. PNAS **91:** 6808-6814.

LI, L., B. FRIDLEY, K. KALARI, G. JENKINS, A. BATZLER *et al.*, 2008 Gemcitabine and cytosine arabinoside cytotoxicity: association with lymphoblastoid cell expression. Cancer Res **68:** 7050-7058.

LI, L., B. L. FRIDLEY, K. KALARI, G. JENKINS, A. BATZLER *et al.*, 2009 Gemcitabine and Arabinosylcytosin Pharmacogenomics: Genome-Wide Association and Drug Response Biomarkers. PLoS One **4:** e7765.

LITI, G., D. M. CARTER, A. M. MOSES, J. WARRINGER, L. PARTS *et al.*, 2009 Population genomics of domestic and wild yeasts. Nature **458:** 337-341.

LOEUILLET, C., S. DEUTSCH, A. CIUFFI, D. ROBYR, P. TAFFE *et al.*, 2008 In vitro whole-genome analysis identifies a susceptibility locus for HIV-1. PLoS Biol **6:** e32.

LYNCH, M., and B. WALSH, 1998 *Genetics and analysis of quantitative traits*. Sinauer, Sunderland, Mass.

MAYR, E., 1942 *Systematics and the origin of species from the viewpoint of a zoologist*. Columbia University Press, New York,.

MEUCCI, M. A., S. MARSH, J. W. WATTERS and H. L. MCLEOD, 2005 CEPH individuals are representative of the European American population: implications for pharmacogenetics. Pharmacogenomics **6:** 59-63.

MONKS, S. A., A. LEONARDSON, H. ZHU, P. CUNDIFF, P. PIETRUSIAK *et al.*, 2004 Genetic inheritance of gene expression in human cell lines. Am J Hum Genet **75:** 1094-1105.

MORLEY, M., C. MOLONY, T. WEBER, J. DEVLIN, K. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430:** 743-747.

NOVEMBRE, J., T. JOHNSON, K. BRYC, Z. KUTALIK, A. R. BOYKO *et al.*, 2008 Genes mirror geography within Europe. Nature **456:** 98-101.

OLANSKY, S., L. SIMPSON and S. H. SCHUMAN, 1954 Environmental factors in the Tuskegee study of untreated syphilis. Public Health Rep **69:** 691-698.

OSTROWSKI, E. A., D. E. ROZEN and R. E. LENSKI, 2005a Pleiotropic effects of beneficial mutations in Escherichia coli. Evolution Int J Org Evolution **59:** 2343-2352.

OSTROWSKI, E. A., D. E. ROZEN and R. E. LENSKI, 2005b Pleiotropic effects of beneficial mutations in Escherichia coli. Evolution **59:** 2343-2352.

OSTROWSKI, E. A., R. J. WOODS and R. E. LENSKI, 2008 The genetic basis of parallel and divergent phenotypic responses in evolving populations of Escherichia coli. Proc Biol Sci **275:** 277-284.

PAQUIN, C., and J. ADAMS, 1983a Frequency of fixation of adaptive mutations is higher in evolving diploid than haploid yeast populations. Nature **302:** 495-500.

PAQUIN, C. E., and J. ADAMS, 1983b Relative fitness can decrease in evolving asexual populations of S. cerevisiae. Nature **306:** 368-370.

PELOSI, L., L. KUHN, D. GUETTA, J. GARIN, J. GEISELMANN *et al.*, 2006 Parallel Changes in Global Protein Profiles During Long-Term Experimental Evolution in Escherichia coli. Genetics **173:** 1851-1869.

PRICE, A. L., N. PATTERSON, D. C. HANCKS, S. MYERS, D. REICH *et al.*, 2008 Effects of cis and trans genetic ancestry on gene expression in African Americans. PLoS Genet **4:** e1000294.

REICH, D. E., and E. S. LANDER, 2001 On the allelic spectrum of human disease. Trends Genet **17:** 502-510.

RIEHLE, M. M., A. F. BENNETT, R. E. LENSKI and A. D. LONG, 2003 Evolutionary changes in heat-inducible gene expression in lines of Escherichia coli adapted to high temperature. Physiol Genomics **14:** 47-58.

RIFKIN, S. A., D. HOULE, J. KIM and K. P. WHITE, 2005 A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438:** 220-223.

ROCKWELL, D. H., A. R. YOBS and M. B. MOORE, JR., 1964 The Tuskegee Study of Untreated Syphilis; the 30th Year of Observation. Arch Intern Med **114:** 792-798.

SCHORK, N. J., J. P. GARDNER, L. ZHANG, D. FALLIN, B. THIEL *et al.*, 2002 Genomic association/linkage of sodium lithium countertransport in CEPH pedigrees. Hypertension **40:** 619-628.

SCHUMAN, S. H., S. OLANSKY, E. RIVERS, C. A. SMITH and D. S. RAMBO, 1955 Untreated syphilis in the male negro; background and current status of patients in the Tuskegee study. J Chronic Dis **2:** 543-558.

SHUKLA, S. J., and M. E. DOLAN, 2005 Use of CEPH and non-CEPH lymphoblast cell lines in pharmacogenetic studies. Pharmacogenomics **6:** 303-310.

SHUKLA, S. J., S. DUAN, J. A. BADNER, X. WU and M. E. DOLAN, 2008 Susceptibility loci involved in cisplatin-induced cytotoxicity and apoptosis. Pharmacogenet Genomics **18:** 253-262.

SHUKLA, S. J., S. DUAN, X. WU, J. A. BADNER, K. KASZA *et al.*, 2009 Whole-genome approach implicates CD44 in cellular resistance to carboplatin. Hum Genomics **3:** 128-142.

SILANDER, O. K., O. TENAILLON and L. CHAO, 2007 Understanding the Evolutionary Fate of Finite Populations: The Dynamics of Mutational Effects. PLoS Biology **5:** 94.

SMIRNOV, D. A., M. MORLEY, E. SHIN, R. S. SPIELMAN and V. G. CHEUNG, 2009 Genetic analysis of radiation-induced changes in human gene expression. Nature **459:** 587-591.

SMITH, E. M., X. WANG, J. LITTRELL, J. ECKERT, R. COLE *et al.*, 2006 Comparison of linkage disequilibrium patterns between the HapMap CEPH samples and a family-based cohort of Northern European descent. Genomics **88:** 407-414.

SPIELMAN, R. S., L. A. BASTONE, J. T. BURDICK, M. MORLEY, W. J. EWENS *et al.*, 2007 Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet **39:** 226-231.

STRANGER, B. E., A. C. NICA, M. S. FORREST, A. DIMAS, C. P. BIRD *et al.*, 2007 Population genomics of human gene expression. Nat Genet **39:** 1217-1224.

STUDIER, F. W., P. DAEGELEN, R. E. LENSKI, S. MASLOV and J. F. KIM, 2009 Understanding the differences between genome sequences of Escherichia coli B strains REL606 and BL21(DE3) and comparison of the E. coli B and K-12 genomes. J Mol Biol **394:** 653-680.

TAYLOR, D. R., C. ZEYL and E. COOKE, 2002 Conflicting levels of selection in the accumulation of mitochondrial defects in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A **99:** 3690-3694.

TERWILLIGER, J., and T. HIEKKALINNA, 2006 An utter refutation of the 'Fundamental Theorem of the HapMap'. Eur J Hum Genet **14:** 426-437.

THE INTERNATIONAL HAPMAP, C., 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

THOMAS, F. J., H. L. MCLEOD and J. W. WATTERS, 2004 Pharmacogenomics: the influence of genomic variation on drug response. Curr Top Med Chem **4:** 1399-1409.

WANG, L., A. L. OBERG, Y. W. ASMANN, H. SICOTTE, S. K. MCDONNELL *et al.*, 2009 Genome-wide transcriptional profiling reveals microRNA-correlated genes and biological processes in human lymphoblastoid cell lines. PLoS One **4:** e5878.

WATSON, J. D., and F. H. CRICK, 1953 The structure of DNA. Cold Spring Harb Symp Quant Biol **18:** 123-131.

WATTERS, J. W., A. KRAJA, M. A. MEUCCI, M. A. PROVINCE and H. L. MCLEOD, 2004 Genome-wide discovery of loci influencing chemotherapy cytotoxicity. Proc Natl Acad Sci U S A **101:** 11809-11814.

WEI, Q., M. R. SPITZ, J. GU, L. CHENG, X. XU *et al.*, 1996 DNA repair capacity correlates with mutagen sensitivity in lymphoblastoid cell lines. Cancer Epidemiol Biomarkers Prev **5:** 199-204.

WOODS, R., D. SCHNEIDER, C. L. WINKWORTH, M. A. RILEY and R. E. LENSKI, 2006 Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. Proc Natl Acad Sci U S A **103:** 9107-9112.

ZEYL, C., 2000 Budding yeast as a model organism for population genetics. Yeast **16:** 773-784.

ZEYL, C., 2004 Experimental studies on ploidy evolution in yeast. FEMS Microbiol Lett **233:** 187-192.

ZEYL, C., 2005 The number of mutations selected during adaptation in a laboratory population of Saccharomyces cerevisiae. Genetics **169:** 1825-1831.

ZEYL, C., 2006 Experimental evolution with yeast. FEMS Yeast Res **6:** 685-691.

ZEYL, C., B. ANDRESON and E. WENINCK, 2005a Nuclear-mitochondrial epistasis for fitness in Saccharomyces cerevisiae. Evolution **59:** 910-914.

ZEYL, C., and G. BELL, 1997 The advantage of sex in evolving yeast populations. Nature **388:** 465-468.

ZEYL, C., C. CURTIN, K. KARNAP and E. BEAUCHAMP, 2005b Antagonism between sexual and natural selection in experimental populations of Saccharomyces cerevisiae. Evolution **59:** 2109-2115.

ZEYL, C., and J. A. DEVISSER, 2001 Estimates of the rate and distribution of fitness effects of spontaneous mutation in Saccharomyces cerevisiae. Genetics **157:** 53-61.

ZEYL, C., M. MIZESKO and J. A. DE VISSER, 2001 Mutational meltdown in laboratory yeast populations. Evolution **55:** 909-917.

ZEYL, C., T. VANDERFORD and M. CARTER, 2003 An evolutionary advantage of haploidy in large yeast populations. Science **299:** 555-558.

ZHANG, W., S. DUAN, W. K. BLEIBEL, S. A. WISEL, R. S. HUANG *et al.*, 2009 Identification of common genetic variants that account for transcript isoform variation between human populations. Hum Genet **125:** 81-93.

CHAPTER TWO: COMPLEX GENETIC CHANGES IN STRAINS OF *SACCHAROMYCES CEREVISIAE* DERIVED BY SELECTION IN THE LABORATORY

STATEMENT OF EFFORT AND ATTRIBUTION

The work presented in this chapter was originally published in the July 2007 issue of *Genetics* (177:1) as "Complex genetic changes in strains of *Saccharomyces cerevisiae* derived by selection in the laboratory" authored by Joshua T. Witten, Christina T.L. Chen, and Barak A. Cohen (WITTEN *et al.* 2007). Witten and Cohen designed the experiments, analyzed the data, and wrote the paper. Witten conducted the experiments. Chen conducted the analysis of variance on transcript abundance data.

ABSTRACT

Selection of model organisms in the laboratory has the potential to generate useful substrates for testing evolutionary theories. These studies generally employ relatively long term selections with weak selective pressures to allow the accumulation of multiple adaptations. In contrast to this approach, we analyzed two strains of *Saccharomyces cerevisiae* that were selected for resistance to multiple stress challenges by a rapid selection scheme to test whether the variation between rapidly selected strains might also be useful in evolutionary studies. We found that resistance to oxidative stress is a multigene trait in these strains. Both derived strains possess the same major effect

adaptations to oxidative stress, but have distinct modifiers of the phenotype. Similarly, both derived strains have altered their global transcriptional responses to oxidative stress in similar ways, but do have at least some distinct differences in transcriptional regulation. We conclude that short term laboratory selections can generate complex genetic variation that may be a useful substrate for testing evolutionary theories.

INTRODUCTION

Selections of model organisms in the laboratory are a useful complement to the study of natural variation in wild populations aimed at understanding principles of adaptive evolution. Indeed, selections in the laboratory have previously been used to address important questions in evolutionary biology such as the existence of transgressive segregation (CASTLE 1951), the fitness effects of mutation (DE VISSER and LENSKI 2002; DE VISSER and ROZEN 2006; ESTES *et al.* 2004; HEGRENESS *et al.* 2006; KASSEN and BATAILLON 2006; LENSKI and TRAVISANO 1994; OSTROWSKI *et al.* 2005; SILANDER *et al.* 2007), the role of genome rearrangements (DUNHAM *et al.* 2002), the effects of haploidy and diploidy (PAQUIN and ADAMS 1983a; ZEYL *et al.* 2003), the frequency of parallelism versus convergence (BUCKLING *et al.* 2003; COHAN and HOFFMANN 1986; COOPER *et al.* 2003; HERRING *et al.* 2006; LENSKI and TRAVISANO 1994), the effects of asexual and sexual lifestyles (DE VISSER and ROZEN 2006; GRIMBERG and ZEYL 2005), and evolutionary change in gene expression (COOPER *et al.* 2003; FEREA *et al.* 1999; PELOSI *et al.* 2006; RIEHLE *et al.* 2003; RIFKIN *et al.* 2005). Microorganisms are particularly attractive for these kinds of experiments due to the ease of culturing, control of environment, short generation time, and genetic tractability (ESTES *et al.* 2004; ZEYL

2000). Several studies have employed the yeast, *Saccharomyces cerevisiae* as a model for natural selection (DUNHAM *et al.* 2002; FEREA *et al.* 1999; GRIMBERG and ZEYL 2005; PAQUIN and ADAMS 1983a; PAQUIN and ADAMS 1983b; ZEYL *et al.* 2005; ZEYL *et al.* 2001; ZEYL *et al.* 2003).

One goal of these studies is to generate diversity that resembles, in some respects, the diversity found in natural populations. Most phenotypic variation in nature is continuous and results from the segregation of alleles at multiple genetic loci, as well as from environmental effects. Beneficial alleles that confer a fitness advantage tend to increase in frequency and eventually fix in natural populations. One major question is how many different combinations of alleles can be beneficial when a continuously varying trait comes under selection. The relative contribution of structural and regulatory changes to adaptive evolution is also unknown. Genetic diversity generated from selections in the laboratory could be a useful tool for addressing these questions.

Genetic diversity generated from selections in the laboratory must be sufficiently complex in order to serve as a useful model of natural variation. Intuitively, strong selections applied over short intervals are expected to produce single genetic changes of large effect that show simple monogenic inheritance, while weak selections applied over long periods of time are expected to produce multiple genetic changes of smaller effect that show complex inheritance (ELENA and LENSKI 2003). Empirically, weak selection regimes applied over long periods of time do allow the accumulation of multiple genetic changes with smaller effect sizes (DE VISSER and LENSKI 2002; LENSKI and TRAVISANO 1994), and may better mimic natural variation. The practical difficulty, however, of

continually and accurately maintaining multiple lines over the course of long term selections presents a barrier to performing certain types of experiments. For example, to address mechanisms of convergence, it is necessary to maintain a very large number of independent lines over a long period of time, an experiment that is not currently practical in many systems (CASTLE 1951; LAURIE *et al.* 2004; LENSKI and TRAVISANO 1994; PAQUIN and ADAMS 1983a; PAQUIN and ADAMS 1983b). This barrier might be overcome if stronger selections conducted over shorter periods of time can generate comparable genetic variation.

As a complement to existing studies on yeast selections, we analyzed two strains from a selection in the laboratory designed to produce stress resistant clones for industrial purposes (CAKAR *et al.* 2005). We sought to determine whether the relatively short duration of selection, the stringency of selection, and the use of a mutagen would bias the selection toward less genetically complex phenotypes, or whether such selections could produce strains with genetically complex traits that can be used to study adaptive evolution.

MATERIALS AND METHODS

Strains, plasmids, and primers: Strains were cultured on solid yeast extract-peptone-dextrose (YPD) medium or in liquid synthetic complete (SC) medium unless otherwise noted.

The haploid *S. cerevisiae* strain CEN.PK 113-14A (*MATa*, *MAL2-8*^{c}, *SUC2*) was used as the ancestor for the selections performed. JWY100 is a clonal isolate of the *MATα* CEN.PK 113-14A haploid with the *MET14* locus replaced by the kanamycin

resistance cassette (*KanMX4*) (GOLDSTEIN and MCCUSKER 1999) (amplified using primers 5'-ACGACGCCTTGGCAATGTAGCA-3' and 5'-GCAAAGCACGCCTCAAAATCTGGT-3') from the *met14Δ0* homozygous diploid from the systematic deletion collection (GIAEVER *et al.* 2002). All transformations were performed as described in (GIETZ and WOODS 2002).

Strain JWY101 is a clonal isolate of population $H_1T_2N_3$ that has been transformed with a plasmid containing a nourseothricin (Nat) resistance cassette (GOLDSTEIN and MCCUSKER 1999) (Mark Johnston, Washington University School of Medicine, St. Louis, MO). Strain JWY102 is a clonal isolate of population H_1H_2 that has been transformed with a plasmid containing a Nat resistance cassette. In order to cross JWY101 and JWY102 we transformed the same clonal isolate of H_1H_2 as JWY102 with a plasmid containing both a Kan resistance cassette and the HO locus (John McCusker, Duke University, Durham, NC). We used dual selection (geneticin and Nat) to select hybrid diploids from all crosses.

Quantitative 96-well growth assay: We arrayed samples in 96-well plate format. Two days prior to the assay, frozen samples were pinned onto solid YPD and incubated overnight at 30°C. One day prior to the assay, single colonies were suspended in 500μL SC medium and incubated overnight at 30°C with shaking at 325rpm in deep-well 96-well plates (Corning #3960). We diluted 10μL of each sample in 490μL fresh SC medium and incubated at 30°C for 5 hours with shaking at 325rpm.

After incubation, all samples were diluted in SC medium to 5.26×10^6 cells /mL in 190μL total volume. We divided each dilute culture into two 95μL samples in adjacent

wells of a 96-well microtiter plate (Corning #3595). We added 5μL of either SC medium or 20mM H_2O_2 in SC (1mM H_2O_2 final) to each well. Samples were incubated on a BioTek Synergy HT plate reader (Winooski, VT) for 20 hours at 30°C with shaking at approximately 1200rpm (level 3) for 20 seconds every four minutes. The reader measured the cell density (absorbance at 600 nm: A_{600}) immediately after every shaking period (299 total measurements).

In order to calculate the growth constant for a sample, we Log_2 transformed the A_{600} measurements and analytically determined the best fit least squares linear regression through all time points where $-2.5 < Log_2(A_{600}) < -2$. The best fit least squares linear regression assumes the form $y = mx + b$, where y is $Log_2(A_{600})$, x is time in seconds, b is the intercept, and m is the growth constant in $\log_2(OD_{600})\bullet s^{-1}$ units.

Biometric analysis of tetrads: Spores were identified as having the phenotype of either the ancestral parent (JWY100), the derived parent (JWY101 or JWY102), or neither using the z-test (α=0.05) to determine from which phenotypic distribution an individual spore measurement was drawn. Tetrads with two spores with the ancestral phenotype and two with the derived phenotype were identified as having 2:2 segregation. The probability of a false positive (a tetrad with 2:2 segregation identified as not having 2:2 segregation due to mislabeling of a spore), if the phenotype is monogenic and all tetrads segregate 2:2, can be determined from the binomial distribution

$$P(x) = \sum_{i}^{N} p^i q^{N-i} \binom{N}{i},$$

where p=0.05 is the probability of incorrectly calling a spore derived from a parental distribution as being distinct, q=0.95 is the probability of correctly calling the spore, and N=4 is the number of spores in the tetrad. This analysis assumes that the parental distributions are Gaussian and the probability of sampling one parent sample from the other parental distribution is infinitesimal (JWY100 vs. JWY101: P=1.83x10^{-44}; JWY100 vs. JWY102: P=6.38x10^{-48}; two-tailed student's t-test). If any one of the four spores is incorrectly labeled, then the tetrad will also be incorrectly labeled. Therefore, the probability, P_I, of incorrectly labeling a tetrad that segregates 2:2 is

$$P_I = 1 - P(x = 0),$$

where $P(x=0)$ is the probability of making no incorrect calls on the spores in a tetrad. The expected number of incorrect calls, if every tetrad is segregating 2:2, is

$$E(n) = P_I n,$$

where n is the number of tetrads assayed.

Additionally, we determined the distribution of growth constants in H_2O_2 for eight replicates of each spore from a single tetrad from the JWY100 x JWY101 and JWY100 x JWY102 crosses as above. We ranked the spores in a tetrad by their average growth constant in H_2O_2 and defined 2:2 segregation as the null hypotheses where

$$H_0 : \mu_D = \mu_\alpha = \mu_\beta$$

$$H_0^{'} : \mu_A = \mu_\gamma = \mu_\delta,$$

where μ is the mean growth constant (A, ancestral parent; D, derived parent; α, highest ranked spore; β, second ranked spore; γ, third ranked spore; δ, lowest ranked spore).

Deviation from either null hypothesis indicates non-2:2 segregation. We used a single-factor analysis of variance (ANOVA) (α=0.05) to test the 2:2 segregation null hypotheses.

We calculated the segregational variance (σ_s^2) as described elsewhere (LYNCH and WALSH 1998). Broad sense heritability (H^2) for each cross was calculated as described elsewhere (MOORE and MCCABE 2006),

$$H^2 = 1 - \frac{\sigma_e^2}{\sigma_s^2},$$

where σ_e^2 is the variance due to environment, which is equal to the pooled variance of the parents (LYNCH and WALSH 1998).

To test for epistasis, we used an adaptation, from (BREM and KRUGLYAK 2005), of the Δ–statistic of (LYNCH and WALSH 1998).

Transcription profiling: Three clones each of JWY100, JWY101, and JWY102 were suspended in 3mL SC medium and incubated overnight at 30°C with shaking at 325rpm. On the next day, we diluted the 3mL overnight cultures in 125mL fresh SC medium. We incubated these cultures at 30°C for six hours at 325rpm until all cultures were in exponential-phase growth. The cell density of each culture was measured on an Eppendorf BioPhotometer (Westbury, NY). We diluted each culture to $8x10^6$ cells/mL in SC medium (50mL total volume) twice. Either 7μL SC medium or 8.8M H_2O_2 (1.2mM H_2O_2 final) was added to each replicate. Both treatments were incubated at 30°C for one hour at 325rpm.

RNA extractions, sample labeling, microarray hybridizations, and mixed-model ANOVA were performed as described by (GERKE *et al.* 2006) with the following changes.

The first mixed ANOVA model removes the genomewide effects of strain and condition on transcript levels,

$$Log_2(Y_{ijkm}) = \mu + S_i + C_j + A_{k(ij)} + SC_{ij} + \varphi_{(ijk)m},$$

where Y_{ijkm} is the median of ratios for each spot, μ is the baseline expression at each spot independent of all other factors, S_i is the average strain main effect, C_j is the average condition main effect, $A_{k(ij)}$ is the average array main effect, SC_{ij} is the average strain-by-condition interaction effect, and $\varphi_{(ijk)m}$ is the residual.

The second mixed ANOVA model, applied separately to each transcript using the residuals of the first model with outliers removed,

$$\varphi_{g(ijk)m} = \gamma_g + \gamma S_{gi} + \gamma C_{gj} + \gamma A_{gk(ij)} + \gamma SC_{gij} + \varepsilon_{(gijk)m},$$

where $\varphi_{g(ijk)m}$ is the residual from the first mixed ANOVA model for each spot, γ_g is the average gene expression for each gene *g*, γS_{gi} is the expression due to strain *i*, γC_{gj} is the expression due to condition *j*, $\gamma A_{gk(ij)}$ is the expression due to array *k*, γSC_{gij} is the expression due to the strain-by-condition interaction effect, and $\varepsilon_{(gijk)m}$ is the residual. Genes with significant effects were determined using a false discovery rate (FDR) of 0.05 as described in (GERKE *et al.* 2006). For genes with significant strain main effects,

differences in expression between pairs of strains were identified by comparing the least-square means of each strain using the *t*-test.

The unrooted tree and estimated branch lengths describing the relationships among transcription profiles were found using the CONTML from PHYLIP (Phylogeny Inference Package) version 3.6 (FELSENSTEIN 1985; FELSENSTEIN 1989; FELSENSTEIN 2005). Bootstrap support (FELSENSTEIN 1985) for this tree was determined using 1000 pseudo-datasets by CONSENSE from PHYLIP version 3.6.

RESULTS

Oxidative stress resistance in ancestral and derived strains: In order to test whether the variation between rapidly selected strains might be useful in evolutionary studies, we isolated individual clones from two different populations of yeast that were selected for tolerance to multiple stresses (CAKAR *et al.* 2005) (Figure 1). JWY101 is an individual clone from the $H_1T_2N_3$ population that was first selected for tolerance to the oxidizing agent hydrogen peroxide (H_2O_2) and then subjected to both a high temperature selection and a freeze/thaw selection. JWY102 is a clone from the H_1H_2 population, which was split from the H1T2N3 population after the initial selection in H_2O_2 and then subjected to additional selection in H_2O_2. The ancestral strain for both populations is CEN.PK 113-14A.

We quantified the oxidative stress resistance phenotype of JWY100, an isogenic derivative of the ancestral strain, CEN.PK 113-14A, and of the two derived strains, JWY101 and JWY102. We measured the growth constants ($Log_2(A_{600})\bullet s^{-1}$) of each of these strains in SC medium in either the presence or the absence of 1mM H_2O_2. In

untreated SC medium, the growth curves for all three strains were similar (Figure 2A) and there was no statistical difference in growth constants in untreated media between the strains (P>0.044 for all comparisons, two-tailed student's t-test, Figure 2B). When challenged with H_2O_2, the two derived strains grew more rapidly than the ancestral strain (Figure 3A). The growth constants of both JWY101 and JWY102 in oxidative stress are approximately two-fold greater than that of their ancestor, JWY100 (Figure 3B). The differences between growth constants in oxidative stress for the two derived strains is not significant (P=0.087, two-tailed student's t-test).

Evidence that oxidative stress resistance segregates as a multigene trait: The derivation of JWY101 and JWY102 used a strong selection applied over a relatively small number of generations. We sought to determine if this scheme allowed time for multiple, adaptive, genetic changes to fix in these strains or if a single adaptive change could account for the phenotypic variation in the F_2 generations of crosses between the ancestral strain, JWY100, and the derived strains, JWY101 and JWY102.

The phenotypic distribution of the progeny of the JWY100 x JWY101 cross is not strictly bimodal, suggesting the phenotype is controlled by more than one locus (Figure 4A). The distribution of progeny phenotypes, however, is not significantly different from a composite of the two parental distributions (two-tailed P=0.1145, Mann-Whitney-Wilcoxon test). We, therefore, sought additional lines of evidence to determine whether oxidative stress resistance is a monogenic or a multigenic trait in JWY101.

We took advantage of the fact that, in *S. cerevisiae*, the four meiotic products are encased in a structure called a tetrad and can be analyzed individually. In the case where

a single adaptive change is responsible for the oxidative stress resistance of a derived strain, the resistance phenotype should segregate 2:2 in tetrads. If more than one adaptive change is involved, the spores in tetrads will show a more complex pattern of segregation.

We determined whether oxidative stress resistance segregates 2:2 in tetrads derived from the JWY100 x JWY101 cross. Among the progeny, the oxidative stress resistance phenotype does not segregate 2:2 in 44 of 54 tetrads (10/54 expected if all tetrads 2:2). The phenotypes of the spores from a representative tetrad are shown in Figure 4B. In this tetrad, there is significant variation in the mean oxidative stress resistance between JWY100 and the two lowest ranked spores (P=5.95x10^{-4}, single-factor ANOVA) and between JWY101 and the two highest ranked spores (P=0.024, single-factor ANOVA). This segregation pattern is not consistent with oxidative stress resistance arising from a single genetic change in JWY101. In addition, the progeny mean phenotypic value is located between the mean of JWY101 and the mid-parent value, suggesting epistasis between segregating loci. In this cross, a *t*-test for epistasis was significant (P=0.024).

Oxidative stress resistance is also a multigenic trait in JWY102. The progeny distribution from the JWY100 x JWY102 cross is not strictly bimodal suggesting the segregation of multiple involved genes (Figure 5A). This distribution of phenotypes in the progeny is significantly different from a composite of the JWY100 and JWY102 phenotype distributions (two-tailed P=0.0007, Mann-Whitney-Wilcoxon test). Similarly, oxidative stress resistance does not segregate 2:2 in 38 of 48 tetrads (9/48 expected if all

tetrads 2:2) from the JWY100 x JWY102 cross. The phenotypes of the spores from a representative tetrad are shown in Figure 5B. In this tetrad, there is significant variation in the mean oxidative stress resistance between JWY100 and the two lowest ranked spores (P=2.16x10^{-8}, single-factor ANOVA), but there is not significant variation between JWY102 and the two highest ranked spores (P=0.667, single-factor ANOVA). This segregation pattern strongly suggests that oxidative stress resistance is not a monogenic trait in JWY102. In addition, the progeny mean phenotypic value is located between the mean of JWY102 and the mid-parent value, suggesting epistasis between segregating loci. A *t*-test for epistasis was significant (P<0.0001).

JWY101 and JWY102 have the same major effect adaptations: We sought to determine whether JWY101 and JWY102 contain the same or different genetic adaptations. We compared the segregational variance (the additional variation in the F_2 progeny due to the segregation of parental genes (LYNCH and WALSH 1998)) from a cross between JWY101 and JWY102 ($\sigma_s^2 = 2.10x10^{-10}$) (Figure 6) to the variance observed in the crosses of these strains to the ancestral strain (JWY100 x JWY101: $\sigma_s^2 = 5.69x10^{-10}$ and JWY100 x JWY102: $\sigma_s^2 = 6.48x10^{-10}$) (Figures 4A & 5A). There is a three-fold reduction in segregational variance in the cross between the derived lines, suggesting that fewer parental genes are segregating in the cross between the derived lines and that JWY101 and JWY102 have the same major effect adaptations for oxidative stress resistance.

The broad sense heritability (H^2) represents the portion of the variance in the progeny phenotypes that can be explained by variance in their genotypes. The heritability in the progeny of the JWY101 x JWY102 cross (H^2=0.490) indicates that additional loci of small effect, which differ between JWY101 and JWY102, are also segregating in this cross. Because a *t*-test for epistasis was highly significant (P<0.0001), these additional minor effect loci appear to have epistatic interactions with each other or with the major effect loci in this cross.

Transcription profiling of derived strains: The oxidative stress response in yeast includes dramatic changes in transcriptional regulation (GASCH *et al.* 2000; JAMIESON 1998). Accordingly, the transcriptional response to H_2O_2 treatment can be used to assess the level of similarity in the oxidative stress response of JWY100, JWY101, and JWY102. In order to assess the similarity of transcriptional regulatory responses, we extracted mRNA from three replicate cultures of JWY100, JWY101, and JWY102 grown in both the presence and the absence of H_2O_2. We used a mixed model ANOVA to identify transcripts with significant (FDR<0.05) strain main effects, condition main effects, and strain-by-condition main effects (Figure 7A).

Down regulation of oxidative phosphorylation and increased proteolysis characterize the response to oxidative stress in yeast (GASCH *et al.* 2000; JAMIESON 1998). Transcripts with significant strain-by-condition effects, that are expressed similarly in JWY101 and JWY102, but differently in JWY100, are enriched for Gene Ontology (GO) Function terms (BERRIZ *et al.* 2003) related to oxidative phosphorylation

and protein catabolism. The derived strains appear to have an exaggerated version of the typical response to oxidative stress.

We evaluated the relationship of the transcriptional response to oxidative stress among the three strains by constructing an unrooted tree from all eighteen samples, using transcripts with either a significant strain main effect or strain-by-condition interaction effect (Figure 7B). We did not include transcripts with only a significant condition main effect, because these transcripts represent the response to oxidative stress that is common among the three strains. Samples group by condition. Within conditions, there is strong bootstrap support indicating that the transcription profiles of the two derived strains, JWY101 and JWY102, are closer to each other than to JWY100, in both conditions. The transcription profiles of the derived strains have the greatest divergence from the ancestral strain in the presence of H_2O_2 (Figure 7B).

Although the expression profiles of the derived strains are far more similar to each other then they are to the ancestral strain, there are at least some distinct differences. In particular, the transcriptional response of JWY102 appears to have diverged further from the ancestral response then JWY101. We therefore tested the hypothesis that JWY101 has significantly fewer mRNAs transcribed differently from the ancestral line than JWY102. The results of this test indicate that the transcriptional profile of JWY101 has diverged less from the ancestral strain than that of JWY102 ($P=6.36 \times 10^{-40}$, χ^2- test).

DISCUSSION

It has been previously shown that long term, weak selections in the laboratory can produce variation similar to that observed in natural isolates. The traits generated are

generally genetic, quantitative, increase directionally throughout the period of selection, and involve both regulatory and structural adaptations (BUCKLING *et al.* 2003; CASTLE 1951; COHAN and HOFFMANN 1986; COOPER *et al.* 2003; CROZAT *et al.* 2005; DE VISSER and LENSKI 2002; DE VISSER and ROZEN 2006; DUNHAM *et al.* 2002; ELENA and LENSKI 2003; FEREA *et al.* 1999; GRIMBERG and ZEYL 2005; LAURIE *et al.* 2004; LENSKI and TRAVISANO 1994; PAQUIN and ADAMS 1983a; PAQUIN and ADAMS 1983b; PELOSI *et al.* 2006; RIEHLE *et al.* 2003; RIFKIN *et al.* 2005; ZEYL *et al.* 2005; ZEYL *et al.* 2003). These studies, however, can be intensive in both labor and resources. As a result, these studies are limited in the number of replicate strains and conditions that can be explored. If comparable variation can be generated from short selections in the laboratory, the reduction in labor and cost would allow key questions to be explored more systematically, such as the frequency of convergence relative to parallelism or the relationship of the strength of selection to the genetic complexity of adaptation.

We examined the adaptation to oxidative stress in two strains derived in the laboratory to determine whether a short intense selection can generate variation comparable to long term, weak selections. A key element of this selection was the use of ethyl methanesulfonate (EMS) mutagenesis to create genetic variation in the starting population, therefore reducing the time required for mutation accumulation.

In natural populations, quantitative variation is the result of segregation of alleles at multiple loci. A primary concern with short term selections in the laboratory is that they are likely to produce adaptive phenotypes arising from a single genetic change. We sought to determine whether a short term selection for stress resistance produced

multigenic phenotypes in the derived lines. The oxidative stress phenotype in JWY102 is clearly multigenic. Both biometric and tetrad analysis indicated the presence of multiple involved genes segregating among the progeny of a cross between JWY102 and the ancestral strain. The results with JWY101 were less clear cut. Biometric analysis showed no significant deviation from the progeny distribution that would be expected if only one gene caused the oxidative stress phenotype, but tetrad analysis showed clear deviation from 2:2 segregation, suggesting more than one involved locus. The evidence for epistasis in the JWY100 x JWY101 cross and transgressive segregation in the JW101 x JWY102 cross also suggests the involvement of more than one gene in the oxidative stress resistance phenotype of JWY101. Our interpretation of these data is that oxidative stress resistance phenotype is a relatively simple complex trait, perhaps with as few as one major effect locus and a few additional minor effect loci. Our results suggest that short term selection in the laboratory can generate traits with a complex genetic basis, but that sometimes these complex traits will have relatively simple architectures.

Given that the ancestral strain was selected for oxidative stress resistance before it was split into two strains (Figure 1), we asked whether JWY101 and JWY102 had the same adaptations to oxidative stress. To answer this question, we crossed the two derived strains and examined the segregation of the phenotype in the F_2 haploid progeny. We observed a three-fold reduction in segregational variance in the progeny of this cross compared to the progeny of either strain backcrossed to the ancestral strain. This result suggests that JWY101 and JWY102 have the same major effect adaptations to oxidative stress. It also suggests that the major effect adaptations fixed in the population early,

before the strains were split at H_1 (Figure 1). The transcriptional profiles of JWY101 and JWY102 also show marked similarities in both treated and untreated cells that are distinct from the ancestral strain, JWY100. These differences in expression may result, directly or indirectly, from the influence of shared major effect loci.

The two derived strains also have distinct differences that likely accumulated after the culture was split. The heritability of stress resistance in the JWY101 x JWY102 cross (H^2=0.490) and evidence for epistatic interactions between loci indicate that unique modifiers of the phenotype have arisen in each strain after they were split from their common ancestor. Likewise there are small, but significant, differences in the expression profiles between the two derived strains. Overall, the transcription data are in agreement with the genetic data. The transcriptional responses of JWY101 and JWY102 are very similar, with some differences (Figures 6 & 7B); but, both strains differ substantially from JWY100 (Figures 4A, 5A, & 7B).

Regulatory effects have been shown to be important for evolutionary changes in several different lineages (CARROLL 2005; WRAY 2007). The set of transcripts with significant strain-by-condition effects were enriched for functions related to oxidative phosphorylation and protein degredation. These classes of genes are known to be regulated as part of the typical response to oxidative stress (GASCH *et al.* 2000; JAMIESON 1998). The fact that the regulation of these classes of genes is altered in the derived strains suggests that they are likely part of the adaptation to oxidative stress.

Here we have shown that a short term selection on a mutagenized population can be used to develop genetically complex adapted strains. These strains may be suitable

substrates for testing theories of adaptive evolution. While long term selection in the laboratory will continue to play an important role in testing evolutionary theories, short term selections will be a useful complement to these studies, especially when the study design requires generating large numbers of parallel strains.

ACKNOWLEDGEMENTS

We thank Petek Çakar and Uwe Sauer for providing strains, Mark Johnston and John McCusker for providing plasmids, and Justin Fay and members of the Cohen Lab for advice and discussion. This work was supported by grants from the American Cancer Society (RSG-06-039-01-GMC) and the National Science Foundation (0543156).

SUPPLEMENTAL FILES

Full transcript abundance data for all samples are available upon request or from *Genetics* (http://www.genetics.org/cgi/content/full/genetics.107.077859/DC1).

FIGURE LEGENDS

Figure 1.—Derivation and phenotype of strains. (A) Schematic of selection used to derive JWY101 and JWY102 after (CAKAR *et al.* 2005). JWY100 is an isogenic, clonal derivative of CEN.PK 113-14A. (CAKAR *et al.* 2005) performed the selection for stress resistant populations as follows. The ancestral population was EMS mutagenized and then treated with H_2O_2 for one hour. The surviving population recovered overnight in minimal medium (YMM). This population was EMS mutagenized and split into two lines. One line was exposed to heat stress followed by overnight recovery in YMM and EMS mutagenesis and freeze/thaw stress ($H_1T_2N_3$) from which a clonal derivative was isolated (JWY101). The second line was exposed to H_2O_2 and allowed to recover overnight in YMM. From this population (H_1H_2), a clonal derivative was isolated (JWY102).

Figure 2.—Growth in untreated media. (A) Growth in untreated SC medium for JWY100, JWY101, and JWY102 over 20 hours. (B) Growth constants in untreated SC medium of JWY100, JWY101, and JWY102. Growth constants in untreated SC medium are approximately equal in the three strains.

Figure 3.—Growth in oxidative stress conditions. (A) Growth in SC medium treated with 1mM H_2O_2 for JWY100, JWY101, and JWY102. JWY101 and JWY102 show greater growth in the presence of H_2O_2 than the ancestral strain, JWY100. (B) Growth constants in SC medium treated with 1mM H_2O_2 of JWY100, JWY101, and JWY102. Growth constants for JWY101 and JWY102 are approximately equal, but are significantly greater than that of JWY100 (***: $P<<0.001$).

Figure 4.—Segregation of oxidative stress resistance in JWY100 x JWY101 cross. (A) Frequency distribution of growth constants in SC medium treated with 1mM H_2O_2 of JWY100, JWY101, and the F_2 progeny of the JWY100 x JWY101 cross. (B) Growth constants in SC medium treated with 1mM H_2O_2 for JWY100, JWY101, and the four spores (α, β, γ, and δ) of a representative tetrad from the JWY100 x JWY101 cross. Oxidative stress resistance does not segregate 2:2 in this tetrad.

Figure 5.—Segregation of oxidative stress resistance in JWY100 x JWY102 cross. (A) Frequency distribution of growth constants in SC medium treated with 1mM H_2O_2 of JWY100, JWY102, and the F_2 progeny of the JWY100 x JWY102 cross. (B) Growth constants in SC medium treated with 1mM H_2O_2 for JWY100, JWY102, and the four spores (α, β, γ, and δ) of a representative tetrad from the JWY100 x JWY102 cross. Oxidative stress resistance does not segregate 2:2 in this tetrad.

Figure 6.—Segregation of oxidative stress resistance in JWY100 x JWY102 cross. (A) Frequency distribution of growth constants in SC medium treated with 1mM H_2O_2 of JWY101, JWY102, and the F_2 progeny of the JWY101 x JWY102 cross.

Figure 7.—Transcription profiling. (A) Venn diagram showing the number of transcripts with significant effects from mixed model ANOVA. (B) Heat map of relative transcript abundance of transcripts with a significant strain main effect or a strain-by-condition main effect (1635 transcripts). Relationship between transcription profiles shown by an unrooted tree (empty circles are nodes). Bootstrap support for each branch is indicated above branch and estimated branch lengths below.

Figure 1

A

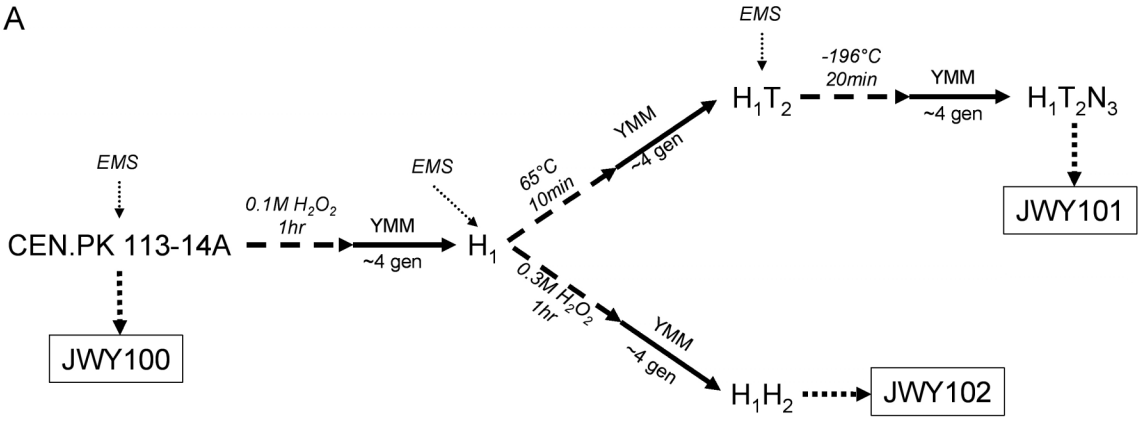

EMS
CEN.PK 113-14A
JWY100
0.1M H_2O_2
1hr
YMM
~4 gen
EMS
H_1
65°C
10min
YMM
~4 gen
EMS
H_1T_2
-196°C
20min
YMM
~4 gen
$H_1T_2N_3$
JWY101
0.3M H_2O_2
1hr
YMM
~4 gen
H_1H_2
JWY102

Figure 2

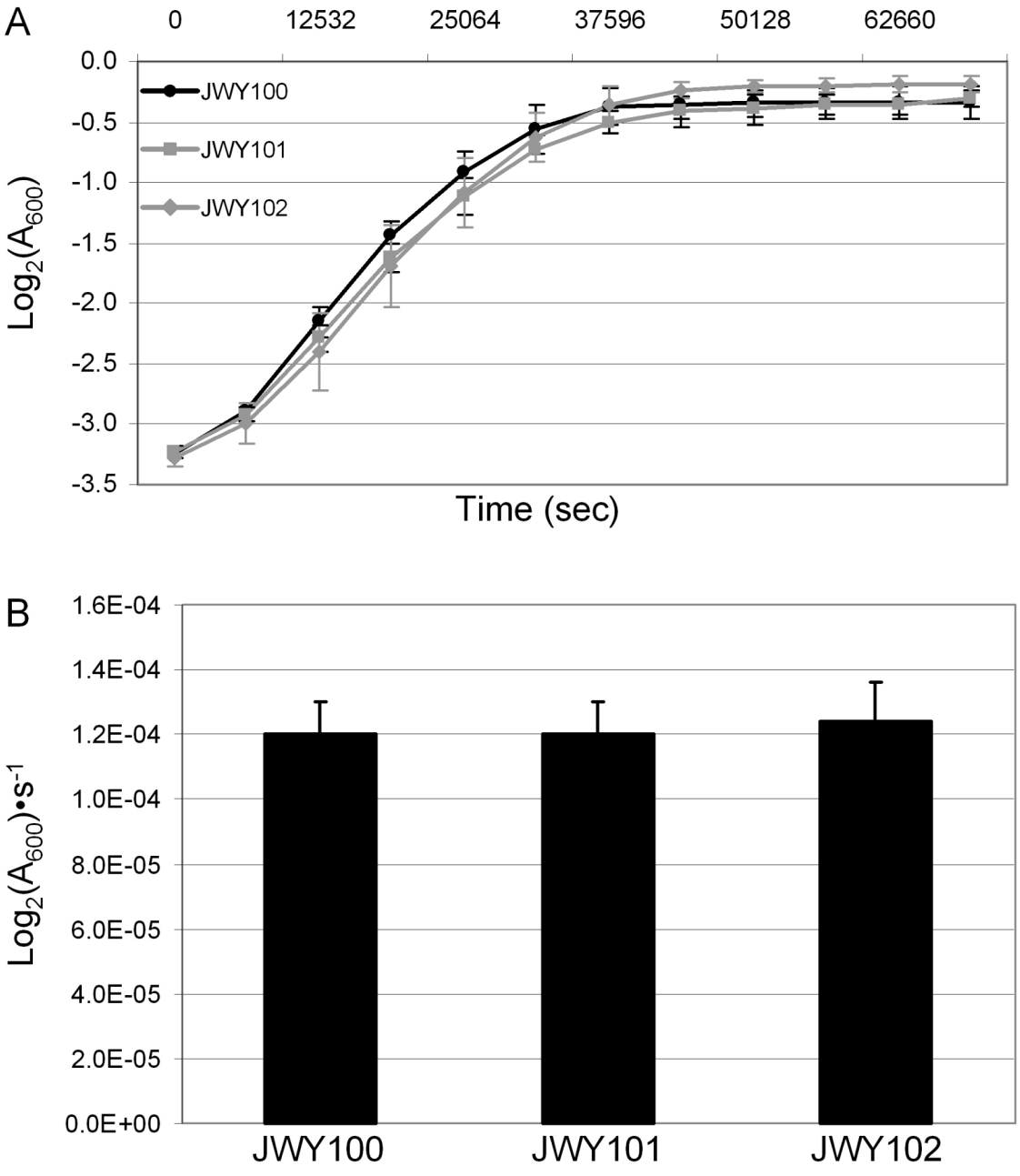

A
0
12532
25064
37596
50128
62660
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
JWY100
JWY101
JWY102
Log2(A600)
Time (sec)
B
1.6E-04
1.4E-04
1.2E-04
1.0E-04
8.0E-05
6.0E-05
4.0E-05
2.0E-05
0.0E+00
Log2(A600)•s-1
JWY100
JWY101
JWY102

Figure 3

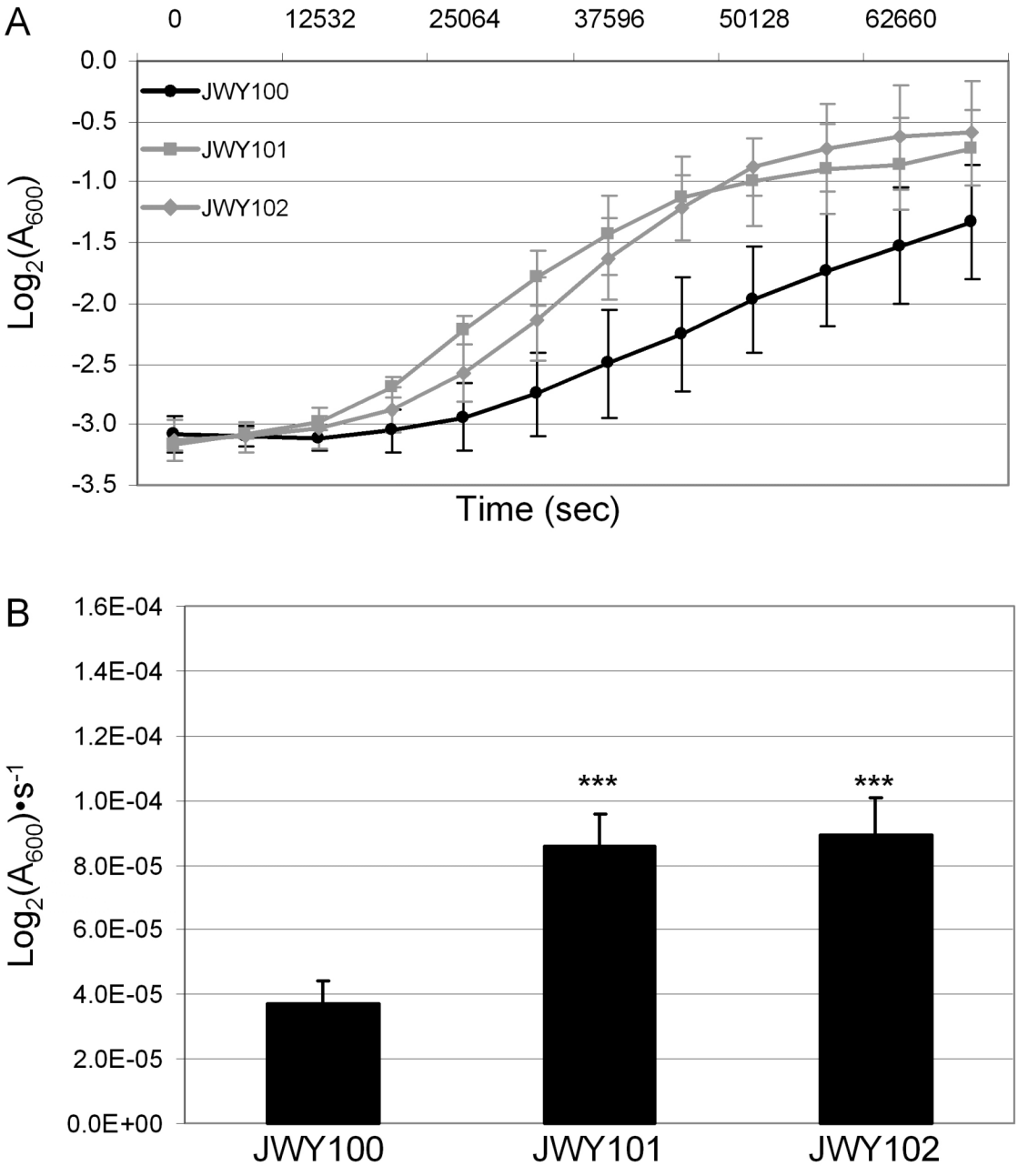
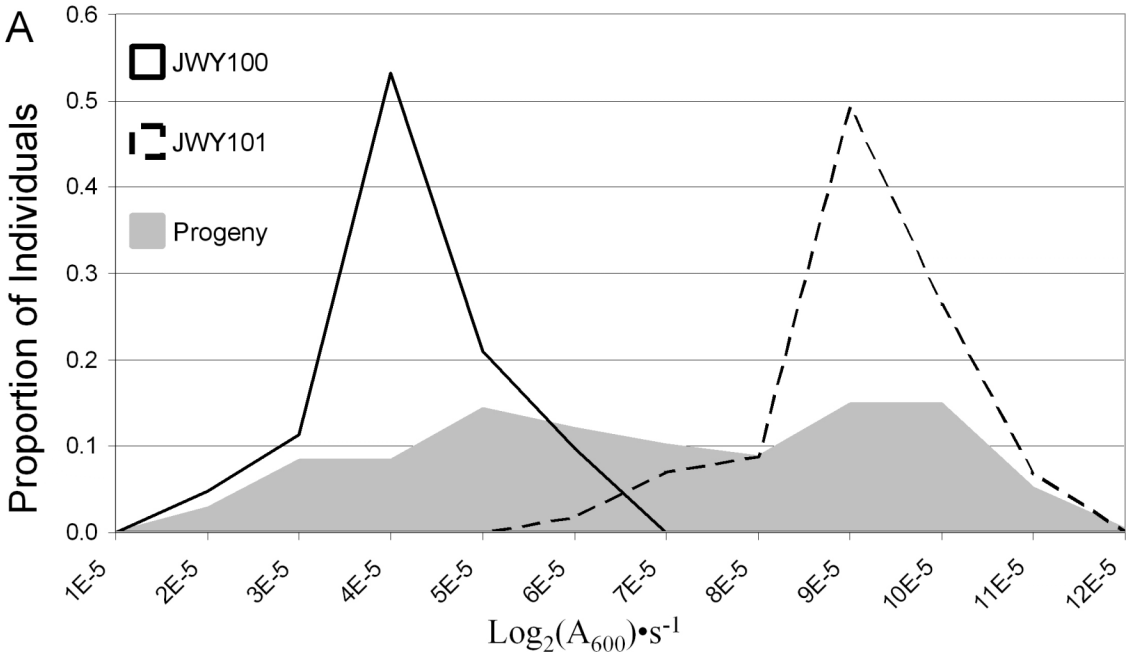

A
0
12532
25064
37596
50128
62660
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
JWY100
JWY101
JWY102
Log2(A600)
Time (sec)
B
1.6E-04
1.4E-04
1.2E-04
1.0E-04
8.0E-05
6.0E-05
4.0E-05
2.0E-05
0.0E+00
Log2(A600)•s-1
***
***
JWY100
JWY101
JWY102

Figure 4


A
JWY100
JWY101
Progeny
Proportion of Individuals
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1E-5
2E-5
3E-5
4E-5
5E-5
6E-5
7E-5
8E-5
9E-5
10E-5
11E-5
12E-5
Log2(A600)•s-1

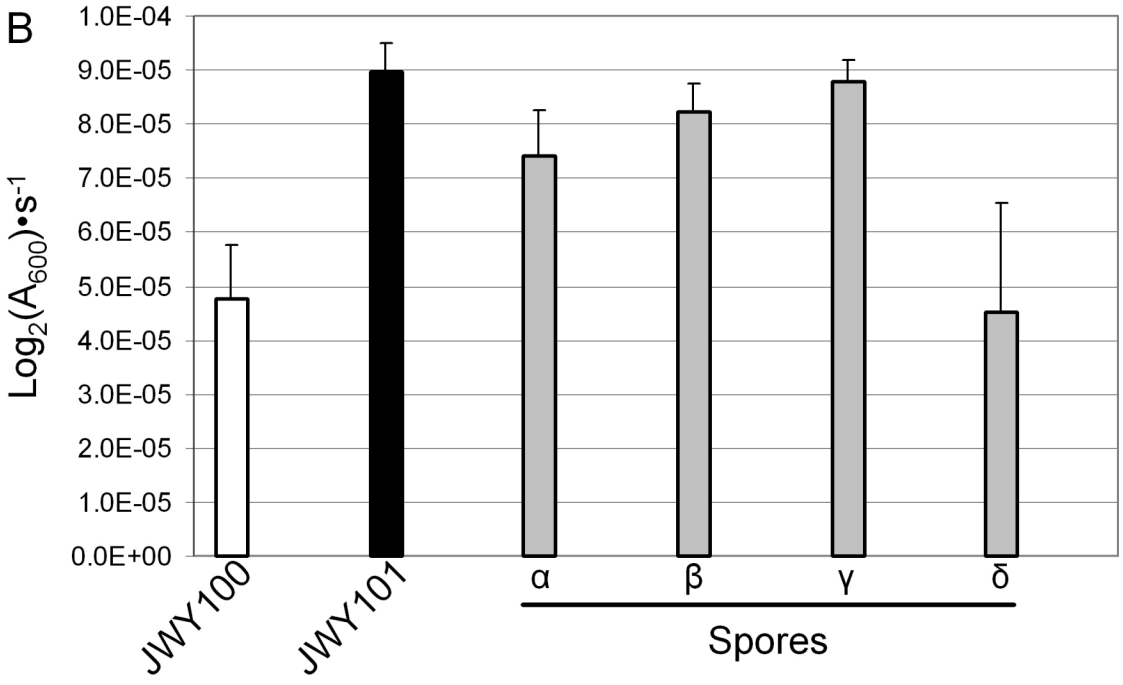
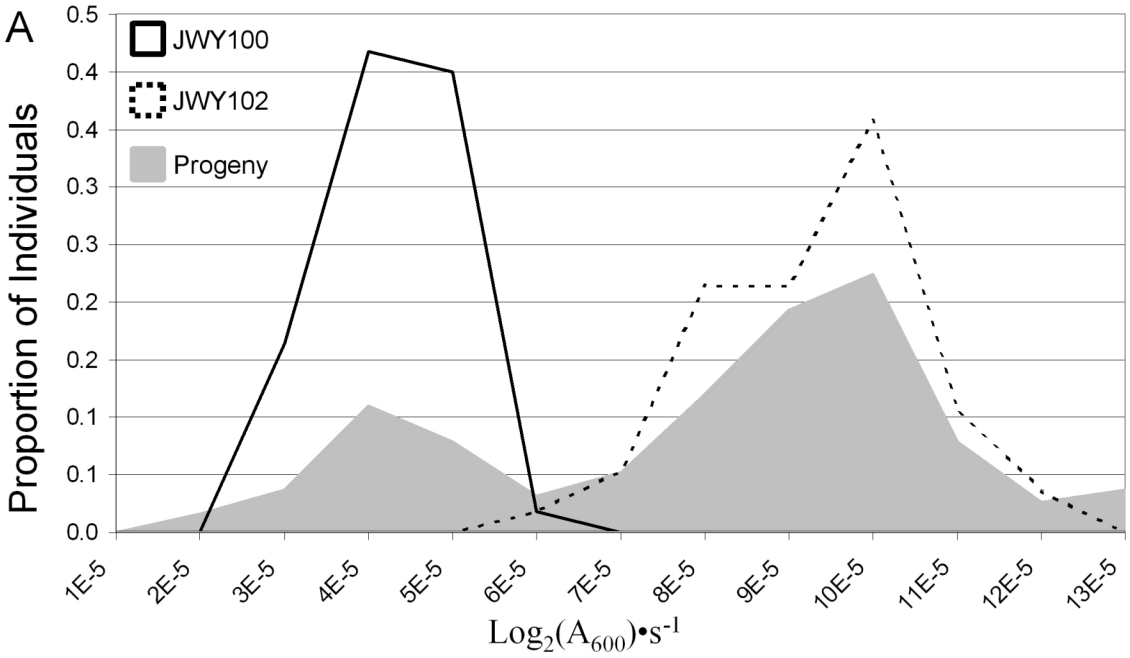

B
1.0E-04
9.0E-05
8.0E-05
7.0E-05
6.0E-05
5.0E-05
4.0E-05
3.0E-05
2.0E-05
1.0E-05
0.0E+00
Log2(A600)•s-1
JWY100
JWY101
α
β
γ
δ
Spores

Figure 5


A
JWY100
JWY102
Progeny
Proportion of Individuals
0.5
0.4
0.4
0.3
0.3
0.2
0.2
0.1
0.1
0.0
1E-5
2E-5
3E-5
4E-5
5E-5
6E-5
7E-5
8E-5
9E-5
10E-5
11E-5
12E-5
13E-5
Log2(A600)•s-1

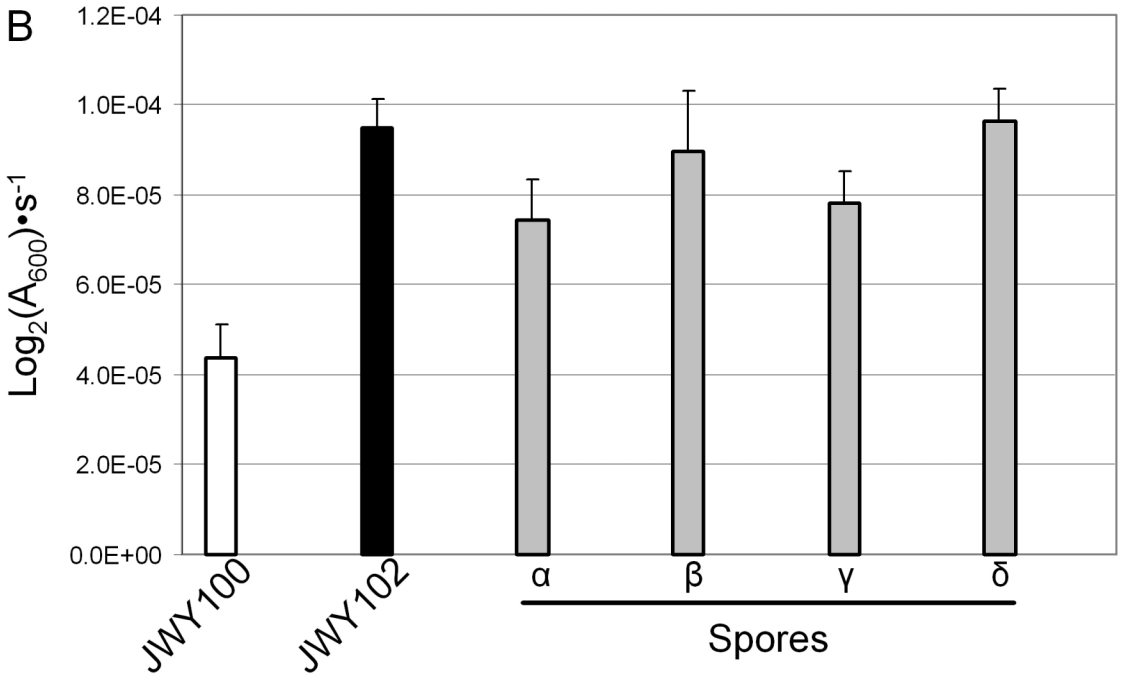

B
1.2E-04
1.0E-04
8.0E-05
6.0E-05
4.0E-05
2.0E-05
0.0E+00
Log2(A600)•s-1
JWY100
JWY102
α
β
γ
δ
Spores

Figure 6

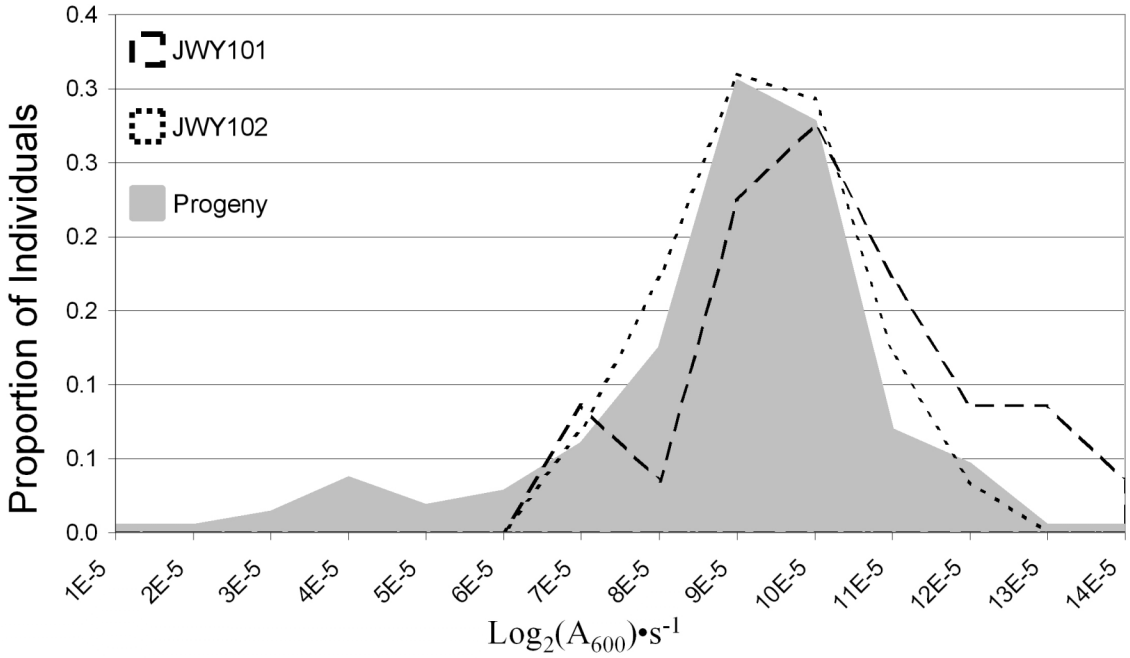
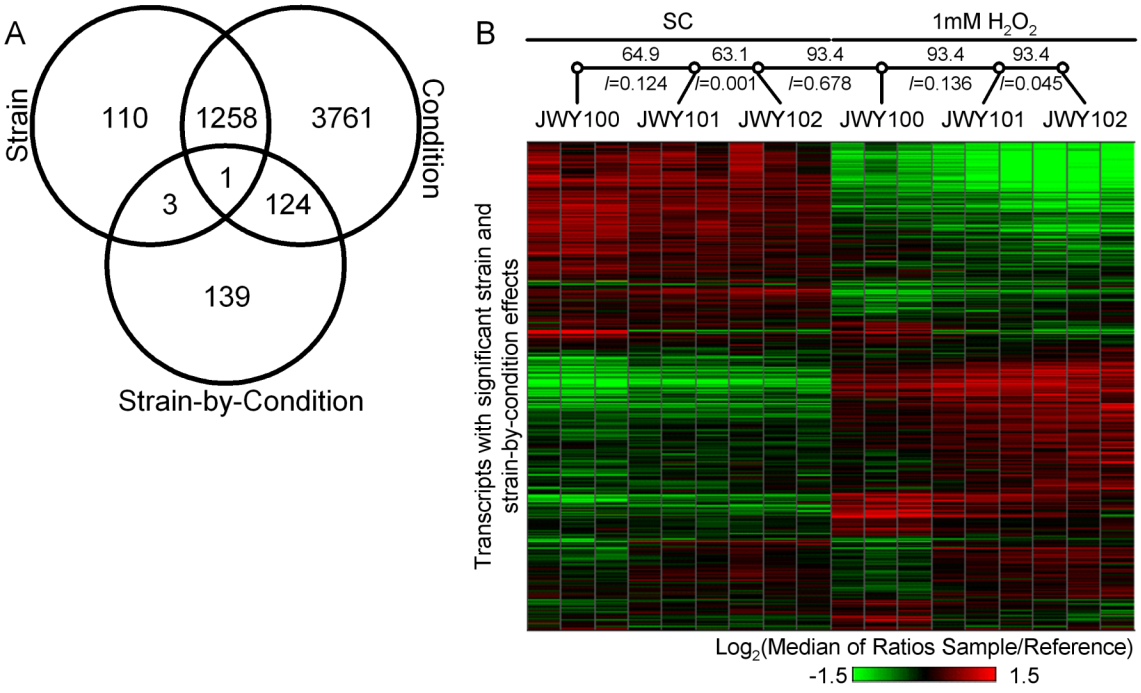

0.4
0.3
0.3
0.2
0.2
0.1
0.1
0.0
JWY101
JWY102
Progeny
Proportion of Individuals
1E-5
2E-5
3E-5
4E-5
5E-5
6E-5
7E-5
8E-5
9E-5
10E-5
11E-5
12E-5
13E-5
14E-5
Log2(A600)•s-1

Figure 7


A
Strain
Condition
Strain-by-Condition
110
1258
3761
1
3
124
139
B
SC
1mM H2O2
64.9
63.1
93.4
93.4
93.4
/=0.124
/=0.001
/=0.678
/=0.136
/=0.045
JWY100
JWY101
JWY102
JWY100
JWY101
JWY102
Transcripts with significant strain and strain-by-condition effects
Log2(Median of Ratios Sample/Reference)
-1.5
1.5

LITERATURE CITED

BERRIZ, G. F., O. D. KING, B. BRYANT, C. SANDER and F. P. ROTH, 2003 Characterizing gene sets with FuncAssociate. Bioinformatics **19:** 2502-2504.

BREM, R. B., and L. KRUGLYAK, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. PNAS **102:** 1572-1577.

BUCKLING, A., M. A. WILLS and N. COLEGRAVE, 2003 Adaptation limits diversification of experimental bacterial populations. Science **302:** 2107-2109.

CAKAR, Z. P., U. O. SEKER, C. TAMERLER, M. SONDEREGGER and U. SAUER, 2005 Evolutionary engineering of multiple-stress resistant Saccharomyces cerevisiae. FEMS Yeast Res **5:** 569-578.

CARROLL, S. B., 2005 Evolution at two levels: on genes and form. PLoS Biol **3:** e245.

CASTLE, W. E., 1951 Variation in the hooded pattern of rats, and a new allele of Hooded. Genetics **36:** 254-266.

COHAN, F. M., and A. A. HOFFMANN, 1986 Genetic divergence under uniform selection. II. Different responses to selection for knockdown resistance to ethanol among Drosophila melanogaster populations and their replicate lines. Genetics **114:** 145-164.

COOPER, T. F., D. E. ROZEN and R. E. LENSKI, 2003 Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli. Proc Natl Acad Sci U S A **100:** 1072-1077.

CROZAT, E., N. PHILIPPE, R. E. LENSKI, J. GEISELMANN and D. SCHNEIDER, 2005 Long-term experimental evolution in Escherichia coli. XII. DNA topology as a key target of selection. Genetics **169:** 523-532.

DE VISSER, J. A., and R. E. LENSKI, 2002 Long-term experimental evolution in Escherichia coli. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. BMC Evol Biol **2:** 19.

DE VISSER, J. A., and D. E. ROZEN, 2006 Clonal Interference and the Periodic Selection of New Beneficial Mutations in Escherichia coli. Genetics **172:** 2093-2100.

DUNHAM, M. J., H. BADRANE, T. FEREA, J. ADAMS, P. O. BROWN *et al.*, 2002 Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A **99:** 16144-16149.

ELENA, S. F., and R. E. LENSKI, 2003 Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet **4:** 457-469.

ESTES, S., P. C. PHILLIPS, D. R. DENVER, W. K. THOMAS and M. LYNCH, 2004 Mutation accumulation in populations of varying size: the distribution of mutational effects for fitness correlates in Caenorhabditis elegans. Genetics **166:** 1269-1279.

FELSENSTEIN, J., 1985 Confidence limits on phylogenies: An approach using the bootstrap. Evolution **39:** 783-791.

FELSENSTEIN, J., 1989 PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics **5:** 164-166.

FELSENSTEIN, J., 2005 PHYLIP (Phylogeny Inference Package) version 3.66, pp. http://evolution.genetics.washington.edu/phylip.html. Department of Genome Sciences, University of Washington, Seattle.

FEREA, T. L., D. BOTSTEIN, P. O. BROWN and R. F. ROSENZWEIG, 1999 Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc Natl Acad Sci U S A **96:** 9721-9726.

GASCH, A. P., P. T. SPELLMAN, C. M. KAO, O. CARMEL-HAREL, M. B. EISEN *et al.*, 2000 Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell **11:** 4241-4257.

GERKE, J. P., C. T. L. CHEN and B. A. COHEN, 2006 Natural Isolates of Saccharomyces cerevisiae Display Complex Genetic Variation in Sporulation Efficiency. Genetics **174:** 985-997.

GIAEVER, G., A. M. CHU, L. NI, C. CONNELLY, L. RILES *et al.*, 2002 Functional profiling of the Saccharomyces cerevisiae genome. Nature **418:** 387-391.

GIETZ, R. D., and R. A. WOODS, 2002 Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Methods Enzymol **350:** 87-96.

GOLDSTEIN, A. L., and J. H. MCCUSKER, 1999 Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. Yeast **15:** 1541-1553.

GRIMBERG, B., and C. ZEYL, 2005 The effects of sex and mutation rate on adaptation in test tubes and to mouse hosts by Saccharomyces cerevisiae. Evolution Int J Org Evolution **59:** 431-438.

HEGRENESS, M., N. SHORESH, D. HARTL and R. KISHONY, 2006 An equivalence principle for the incorporation of favorable mutations in asexual populations. Science **311:** 1615-1617.

HERRING, C. D., A. RAGHUNATHAN, C. HONISCH, T. PATEL, M. K. APPLEBEE *et al.*, 2006 Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38:** 1406-1412.

JAMIESON, D. J., 1998 Oxidative stress responses of the yeast Saccharomyces cerevisiae. Yeast **14:** 1511-1527.

KASSEN, R., and T. BATAILLON, 2006 Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet **38:** 484-488.

LAURIE, C. C., S. D. CHASALOW, J. R. LEDEAUX, R. MCCARROLL, D. BUSH *et al.*, 2004 The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. Genetics **168:** 2141-2155.

LENSKI, R. E., and M. TRAVISANO, 1994 Dynamics of Adaptation and Diversification: A 10,000-Generation Experiment with Bacterial Populations. PNAS **91:** 6808-6814.

LYNCH, M., and B. WALSH, 1998 *Genetics and analysis of quantitative traits*. Sinauer, Sunderland, Mass.

MOORE, D. S., and G. P. MCCABE, 2006 *Introduction to the pratice of statistics*. W.H. Freeman and Co., New York.

OSTROWSKI, E. A., D. E. ROZEN and R. E. LENSKI, 2005 Pleiotropic effects of beneficial mutations in Escherichia coli. Evolution Int J Org Evolution **59:** 2343-2352.

PAQUIN, C., and J. ADAMS, 1983a Frequency of fixation of adaptive mutations is higher in evolving diploid than haploid yeast populations. Nature **302:** 495-500.

PAQUIN, C. E., and J. ADAMS, 1983b Relative fitness can decrease in evolving asexual populations of S. cerevisiae. Nature **306:** 368-370.

PELOSI, L., L. KUHN, D. GUETTA, J. GARIN, J. GEISELMANN *et al.*, 2006 Parallel Changes in Global Protein Profiles During Long-Term Experimental Evolution in Escherichia coli. Genetics **173:** 1851-1869.

RIEHLE, M. M., A. F. BENNETT, R. E. LENSKI and A. D. LONG, 2003 Evolutionary changes in heat-inducible gene expression in lines of Escherichia coli adapted to high temperature. Physiol Genomics **14:** 47-58.

RIFKIN, S. A., D. HOULE, J. KIM and K. P. WHITE, 2005 A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438:** 220-223.

SILANDER, O. K., O. TENAILLON and L. CHAO, 2007 Understanding the Evolutionary Fate of Finite Populations: The Dynamics of Mutational Effects. PLoS Biology **5:** 94.

WITTEN, J. T., C. T. CHEN and B. A. COHEN, 2007 Complex genetic changes in strains of Saccharomyces cerevisiae derived by selection in the laboratory. Genetics **177:** 449-456.

WRAY, G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nat Rev Genet **8:** 206-216.

ZEYL, C., 2000 Budding yeast as a model organism for population genetics. Yeast **16:** 773-784.

ZEYL, C., C. CURTIN, K. KARNAP and E. BEAUCHAMP, 2005 Antagonism between sexual and natural selection in experimental populations of Saccharomyces cerevisiae. Evolution Int J Org Evolution **59:** 2109-2115.

ZEYL, C., M. MIZESKO and J. A. G. M. D. VISSER, 2001 Mutational meltdown in laboratory yeast populations. Evolution **55:** 909-917.

ZEYL, C., T. VANDERFORD and M. CARTER, 2003 An evolutionary advantage of haploidy in large yeast populations. Science **299:** 555-558.

CHAPTER THREE: USE OF LYMPHOBLASTOID CELL LINES IN THE STUDY OF HUMAN GENETIC VARIATION

ABSTRACT

Understanding human genetic variation is one of the most important goals of biomedical research and one of the most constrained by ethical considerations and technical limitations. The study of lymphoblastoid cell lines derived from human blood samples has been suggested as a complementary approach that may avoid some of these complications. The reliability of these cell lines, not only as a representation of human biology, but also a consistent experimental resource has recently been called into question.

To examine these issues, I quantified variation in the expression of cell surface proteins amongst a panel of 255 lymphoblastoid cell lines using flow cytometry. I found that lymphoblastoid cell lines are a reliable experimental platform that can be used for consistently repeatable phenotyping, providing that sensible approaches to cell culture, experiment design, and data processing are adopted. Almost all of the phenotypic variation can be explained by intrinsic differences between lines, a small, but significant, fraction of which can be explained by additive genetic variation, suggesting that it will be possible to identify the genomic location of some putatively, causal variants by linkage-association mapping.

While acknowledging concerns that lymphoblastoid cell lines only represent a fraction of the diversity of human cell types, I conclude that lymphoblastoid cell lines, as

well as other cell culture methods, are reliable experimental resources with distinct strengths and weaknesses. When used with careful consideration of those strengths and weaknesses, cell culture methods should be an integral part of a complementary program to study human genetic variation.

INTRODUCTION

Certain classes of experiments necessary to discover the genetic differences between humans, such as unpredictable variation in response to cytotoxic chemotherapeutics, critical for functional, personalized medicine cannot be ethically conducted in healthy humans. Technical successes in identifying genotype-phenotype association for both disease risk and general traits in humans contrast with the low predictive power conferred by these associations to highlight the issue of validation. Direct experimentation to confirm genotype-phenotype associations are rarely ethical or technically feasible in human subjects. In response to these considerations, lymphoblastoid cell lines (LCLs) - circulating B cells immortalized by Epstein-Barr virus (EBV) infection - have been suggested as a potential, *ex vivo* model system for the study of human genetic variation.

LCLs have a number of characteristics that make them an attractive model for human genetic variation. Genetic variation in LCLs represents genetic variation that exists in human populations (BAUCHET *et al.* 2007; HE *et al.* 2009; LAO *et al.* 2008; MEUCCI *et al.* 2005; SMITH *et al.* 2006). As cell culture lines, LCLs are not subject to the same ethical restrictions as human subjects, allowing discovery based pharmacogenomic studies that would otherwise be unacceptable. There are also options for direct

experimental confirmation of genetic hypotheses in LCLs that are not available in human subjects, such as siRNA (LI *et al.* 2009). These options are certain to expand for LCLs with technological advancements, while the options for human subjects may be fundamentally constrained by ethical considerations.

In microorganisms, where it is possible to directly test genotype-phenotype associations, the nature of the genetic variation controlling phenotypic variation does not agree with the observations of genotype-phenotype association studies in humans, which indicate that human genetic variation is controlled by many loci with very small effect sizes. For example, variation sporulation efficiency between two isolates of the budding yeast *Saccharomyces cerevisiae* can be explained by only a few QTL resolved to the nucleotide level (quantitative trait nucleotides, QTN) of relatively large effect. The QTN effects are primarily additive, but there are significant epistatic interactions between QTN and with the genetic background. The effects of these QTN have been confirmed by direct experimentation (GERKE *et al.* 2009; GERKE *et al.* 2006). In contrast, height, a highly heritable human trait, has had 47 QTL associated with it by GWA studies, all of small effect. Although some of these QTL cover genes known to affect stature and the QTL can predict inclusion at the extremes of the height distribution, these QTL do not predict height for the vast majority of the population (LETTRE 2009).

It is not known whether this discrepancy arises due to fundamental differences in the relationship of genotype and phenotype, such as different amounts of epistasis, between microorganisms and large, multicellular organisms, an artifact of technical challenges in human genetic studies, or a combination of both.

To date, quantitative genetic studies using LCLs have not provided a resolution to this question. The few QTL identified in LCLs in both pharmacogenomic and transcript abundance studies also explain only a fraction of the phenotypic variation, suggesting that phenotypic variation in humans at the level of transcription, cell survival, and organismal disease is controlled by many loci of small effect. Again, it is not clear whether the small effect sizes reflect the true biology of human genetic variation, a result of LCLs not being a representative human analog, or an artifact of measuring "phenotypes" with potentially small organismal phenotypic effects (i.e., transcript abundance).

The reliability of LCLs as an experimental resource, however, has recently been called into question (CHOY *et al.* 2008). Estimating the heritability of phenotypes and the identification of genetic variation associated or linked to phenotypic variation is dependent on the reliability of phenotypic quantification. If phenotypic quantification is unreliable as suggested (CHOY *et al.* 2008), then not only must the use of LCLs as a resource for identifying genetic variation that controls phenotypic variation in humans be questioned, but so must the results of a number of studies using LCLs for this purpose (ATLAS *et al.* 1976; BERGEN *et al.* 2007; BLEIBEL *et al.* 2009; CHEUNG *et al.* 2003; CHEUNG and EWENS 2006; CHEUNG *et al.* 2005; CLOOS *et al.* 1999; CORREA and CHEUNG 2004; DEUTSCH 2005; DOLAN *et al.* 2004; DUAN *et al.* 2007; DUAN *et al.* 2009; FORD *et al.* 2001; HARRIS *et al.* 2005; HARTFORD *et al.* 2009; HUANG *et al.* 2007a; HUANG *et al.* 2008a; HUANG *et al.* 2008b; HUANG *et al.* 2007b; HUANG *et al.* 2007c; JEN and CHEUNG 2003; KWAN *et al.* 2007; LI *et al.* 2008; LI *et al.* 2009; LOEUILLET *et al.* 2008; MONKS *et*

al. 2004; MORLEY *et al.* 2004; SCHORK *et al.* 2002; SHUKLA and DOLAN 2005; SHUKLA *et al.* 2008; SHUKLA *et al.* 2009; SMIRNOV *et al.* 2009; WANG *et al.* 2009a; WATTERS *et al.* 2004; WEI *et al.* 1996; ZHANG *et al.* 2009). This experimental variation may not only reduce the ability to identify genetic variants controlling phenotypic variation (i.e., false negatives), but also may cause spurious genetic correlations (i.e., false positives), if the experimental variation is correlated to unmeasured, potentially heritable phenotypes like growth rate.

While there are a number of steps during the establishment of LCLs that may contribute to non-genetic variation between lines, there are three general processes in LCL phenotyping that could contribute to experimental variation using established lines (e.g., the CEPH collection): variation in the freezing and thawing of individual sample aliquots (freeze/thaw variance), variation in cell culturing on technical replicates of the same sample preparation (culture variance), and variation in the phenotype assay (assay variance). The ability to make repeatable measurements of LCL phenotypes is critical to both the correlation of genetic variation with phenotypic variation and confidence in those correlations. Accordingly, I have examined both the repeatability of LCL phenotyping and the correlation of genetic variation with phenotypic variation using this cell culture based resource.

In this study, I have used a panel of 255 LCLs in 19 three-generation pedigrees from a population representative of Northern and Western European decent (BAUCHET *et al.* 2007; HE *et al.* 2009; LAO *et al.* 2008; MEUCCI *et al.* 2005; SMITH *et al.* 2006) to identify the genetic underpinnings of phenotypic variation in humans. In doing so, I have

explicitly tested the reliability of LCLs as an experimental resource to study human genetic variation. I have also explored whether the use of alternative phenotyping strategies, namely quantifying cell surface expression of proteins instead of mRNA transcript abundance, enables larger proportions of phenotypic variation to be explained by genotype.

RESULTS

Description of sample pedigrees: In order to study the genetic variation in cell surface expression of proteins using LCLs, I established a panel of 255 LCLs (Supplemental Table 1) from 19 three-generation families from the Utah population in the Foundation Jean Dausset-Centre d'Etude du Polymorphisme Humain collection (CEPH-UT) (DAUSSET *et al.* 1990). Both parents (second generation) were present in all pedigrees ($\overline{N} = 2$). All four grandparents (first generation) were present in most pedigrees ($\overline{N} = 3.5$). Third generation sibships were large ($\overline{N} = 7.9$).

Description of phenotypes: Phenotypes were measured by flow cytometry following paraformaldehyde fixation (fixes cells without permeablizing the cell membrane) and labeling with phycoerythrin (PE) labeled monoclonal antibodies (mAb) to targeted proteins (Materials and Methods). Targeted proteins were CD4, CD19, CD23 (or FCER2), CD38, CD40, CD45RA (or Protein Tyrosine Phosphotase, Receptor Type C), CD86, Intercellular Adhesion Molecule 1 (ICAM-1), and Toll-like Receptor 9 (TLR-9).

The CD4 and TLR-9 antibodies are used as controls for the other antibodies. CD4 is a major histocompatibility complex class II co-receptor typically observed only on T_H1 and T_H2 helper T cells, an expression pattern that is routinely used to identify T cells.

Therefore, it is not expected to observe CD4 on the B cell derived LCLs. Indeed, measured CD4 transcript abundance is extremely low in LCLs (MORLEY *et al.* 2004). Here, the CD4 mAb is used as an isotype control for level of background signal resulting from the combination of auto-fluorescence and non-specific binding. TLR-9 is expressed in B cells and recognizes CpG DNA, typically from bacteria, but is localized to an intracellular endosomal compartment. Therefore, the TLR-9 mAb is an isotype control for cell surface localization of labeling. Comparison to CD4 and TLR-9 staining allows for the detection of staining over background levels.

CD38 and CD86 are expressed on mature activated B-cells. CD38 has a role in B cell proliferation and is expressed both in early and germinal center B cells. CD38 expression has been implicated as both risk factor and prognostic marker for chronic lymphocytic leukemia (DEAGLIO *et al.* 2008; JAMROZIAK *et al.* 2009). CD86 is a ligand for CD28 and CTLA-4 and is expressed on activated B cells. CD86 has been suggested to be a factor in the development of asthma (CORYDON *et al.* 2007; ZHU *et al.* 2004). While these proteins are generally expressed in activated B cells, it is mature, inactivated B cells that are immortalized to form LCLs. EBV immortalization, however, may confer an activated phenotype to the LCLs (POKROVSKAJA *et al.* 1996; SATOH *et al.* 2003), and inappropriate activation of B cells could contribute to autoimmune disorders. Assays with CD38 and CD86 will yield information on the genetic differences in unactivated B-cells, which could contribute to conditions where inappropriate activation of B-cells is part of the etiology.

The remaining targeted proteins are expressed on the surface of the mature, unactivated B cells from which LCLs are derived. CD19 is a component of the B cell co-receptor with CD21 and CD81. CD19 overexpression has been associated with loss of tolerance, production of autoantibodies, and systematic sclerosis (TAYLOR *et al.* 2006). CD23 is a low-affinity receptor for IgE, regulates IgE synthesis, and is a ligand for the B cell co-receptor. CD23 expression has been connected with the development of B-chronic lymphocytic leukemia (CALAMINICI *et al.* 2004; JURISIC *et al.* 2008; SCHWARZMEIER *et al.* 2005). CD40 is the receptor for a co-stimulatory, activating signal from CD154 (CD40 ligand) on helper T cells. Variation at the *CD40* locus is a risk factor for rheumatoid arthritis (RAYCHAUDHURI *et al.* 2008). CD45RA is a B cell specific isoform of CD45, which enhances the signal from the antigen receptor. Variation at the *CD45* locus had been associated with multiple sclerosis (BALLERINI *et al.* 2002), but this association has since been refuted by more comprehensive studies (GOMEZ-LIRA *et al.* 2003; SZVETKO *et al.* 2009). ICAM-1 helps create a tight junction between B cells and T cells following antigen specific binding of the cells. Variation at the *ICAM-1* locus has been associated with differentiation of colorectal cancer (WANG *et al.* 2009b).

To obtain linear and informative measurements of this diverse set of cell surface expression phenotypes by flow cytometry, both “low” and “high” values were used for the photomultiplier tube (PMT) settings (Materials and Methods). CD4, CD38, CD40, CD45RA, and ICAM-1 were assayed using the “low” settings. CD4, CD19, CD23, CD86, and TLR-9 were assayed using the “high” settings. All flow cytometry measurements are reported below in arbitrary fluorescence units (AFU).

Mean CD4 cell surface expression was not significantly different from auto-fluorescence in the 1 PBS/0.2% BSA control ($\mu_{PBS} = 22.28;\ \mu_{CD4} = 21.01;\ P = 0.39$, two-tailed student's t-test) suggesting that the effects of non-specific binding are low in this assay. Mean TLR-9 cell surface expression was significantly greater than CD4 ($P = 6.25 \times 10^{-73}$, one-tailed, paired student's t-test; Figure 1A) indicating that there are either low levels of cell surface expression of TLR-9 or that the fixed cell membrane is somewhat permeable to the PE-mAbs.

Mean CD19 cell surface expression was significantly greater than $CD4_{high}$ ($\mu_{CD4_{high}} = 59.30;\ \mu_{CD19} = 63.67;\ P = 1.19 \times 10^{-20}$, one-tailed, paired student's t-test), but was not greater than TLR-9 ($\mu_{TLR-9} = 81.06;\ \mu_{CD19} = 63.67;\ P = 5.93 \times 10^{-66}$, one-tailed, paired student's t-test). A bimodal distribution of CD19 cell surface expression levels would suggest that CD19 expression is not a binary character where some lines are actively expressing (i.e., equal to or greater than TLR-9) and others are not (i.e., equal to $CD4_{high}$. The distribution of CD19 cell surface expression levels is unimodal (Figure 1B), indicating that significant CD19 cell surface expression cannot be detected in this assay using standardized settings and protocols.

The mean cell surface expression of all other phenotypes measured was significantly greater than the relevant controls ($P \leq 5.00 \times 10^{-30}$, one-tailed, paired student's t-test; Figure 1C-H). For the majority of phenotypes, the flow cytometry assay was able to quantify cell surface expression levels above the noise of auto-fluorescence and non-specific binding.

Experimental components of phenotypic variance: Previous work did not separate freeze/thaw variance and cell culture/assay variance (CHOY *et al.* 2008), making it difficult to identify factors contributing to poor replication. I separated freeze/thaw variance from cell culture/assay variance, by measuring cell surface expression of all targeted proteins and a no antibody control (1X phosphate buffered saline/0.2% bovine serum albumin [1X PBS/0.2% BSA]) for three independent cultures that were grown from a single frozen aliquot of each LCL in the panel. A subset ($\bar{n}_{Ab} = 20.6$, $18 \le n_{Ab} \le 22$) of the LCL panel was measured for each phenotype. Poorly growing samples were removed from analysis (Materials and Methods). Cell surface expression was repeatable between cell culture/assay replicates for control phenotypes ($\bar{r} = 0.76$; $0.35 \le r \le 0.92$) and experimental phenotypes ($\bar{r} = 0.86$; $0.73 \le r \le 0.95$).

Due to the number of culture replicates measured (n=3), these data were highly sensitive to outlier data points. This supposition was supported by the fact that the lowest correlation ($r = 0.35$) was an isolate event that was an extreme deviation from all other observations ($0.73 \le r \le 0.95$) and appeared to be due to a single set of replicates (data not shown). I removed outlier data points (Materials and Methods) and repeated the analysis.

I found that cell surface expression was repeatable between cell culture/assay replicates for both control phenotypes ($\bar{r} = 0.93$; $0.88 \le r \le 0.96$) and experimental phenotypes ($\bar{r} = 0.94$; $0.91 \le r \le 0.97$). Despite the potential increase in variance due to the culturing process, this result compares favorably with the assay repeatability reported previously for both drug response (Spearman's rank correlation: $0.86 \le \rho \le 0.99$) (CHOY

et al. 2008) and transcript abundance ($\bar{r} = 0.98;\ 0.94 \le r \le 0.99$) (MORLEY *et al.* 2004). These results suggest that cell culture/assay variance only makes a minor contribution to total phenotypic variation (Figure 2).

In order to determine the total experimental variation, the cell surface expression levels of the full panel were quantified for all phenotypes using four independent freeze/thaw replicates of each line. The number of freeze/thaw replicates assayed for each phenotype ($n = 4$) and experience from the cell culture/assay variance discovery indicated that the phenotypic mean would be highly sensitive to individual outlier data points.

To test the impact of outlier data points, I quantified the repeatability before and after the removal of outlier data points (as above). Before the removal of outlier data points, freeze/thaw replicates had substantially lower repeatability for control ($\bar{r} = 0.48;\ 0.32 \le r \le 0.64$) and experimental phenotypes ($\bar{r} = 0.65;\ 0.57 \le r \le 0.71$) relative to the cell culture/assay replicates. These repeatablities for the experimental phenotypes were still at the upper limit of the range reported previously (CHOY *et al.* 2008). This result would suggest that experimental variance makes up a large portion of the total phenotypic variance.

After the removal of outlier data points, however, freeze/thaw repeatability improved dramatically for both control ($\bar{r} = 0.84;\ 0.77 \le r \le 0.89$; Table 1) and experimental phenotypes ($\bar{r} = 0.90;\ 0.87 \le r \le 0.92$; Table 1). These results suggest that experimental variance is not a major component of the total phenotypic variance and that freeze/thaw variance does add substantially to cell culture/assay variance (Figure 2).

Because the median is robust to outliers, the median phenotype for each line was used in all subsequent analyses, instead of the mean following outlier removal with which the median is highly correlated ($\bar{r} = 0.98$).

Covariate components of phenotypic variance: The CEPH-UT LCLs are associated with limited information on the individuals from which they were derived (pedigree position, sex, and age at sampling). Females and males had approximately equal representation in the full panel ($female = 47.8\%$; $male = 52.1\%$). Due to the structure of the pedigrees, ages (years) were not evenly distributed, but cluster into three groups corresponding to generation within the pedigree ($\tilde{g}_1 = 70.5$; $\tilde{g}_2 = 45$; $\tilde{g}_3 = 16$; Figure 3A). In addition, the LCLs in the panel were randomly distributed into three 96-well sampling plates (Materials and Methods). Pedigree position, sex, and age are all potential covariates that may have non-genetic effects on phenotype.

I used mixed model analysis of variance (ANOVA) to assess the significance or the effects of the potential covariates of generation in the pedigree, sex, age at sampling, and plate:

$$\phi_{ij} = \mu_i + \mathbf{a}_i\left(gen_j\right) + \mathbf{b}_i\left(sex_j\right) + \mathbf{c}_i\left(age_j\right) + \mathbf{d}_i\left(plate_j\right) + \varepsilon_{ij},$$

where ϕ_{ij} is the median value for the i^{th} phenotype of the j^{th} line, μ_i is the population mean value for the i^{th} phenotype, gen_j is the generation in the pedigree of the j^{th} line, sex_j is the sex of the j^{th} line, age_j is the age at sampling of the j^{th} line, $plate_j$ is the plate of the j^{th} line, and ε_{ij} is the residual for the i^{th} phenotype of the j^{th} line. Generation in the pedigree, sex, and plate were treated as class variables. Age was treated

as a continuous variable. As noted above, generation in the pedigree and age are not independent variables.

Generation, sex, and age had a significant effect ($\alpha = 0.05$) on a minority of phenotypes. Generation (Figure 3B; Table 2) had a significant effect on TLR-9 ($P = 8.38 \times 10^{-3}$; Figure 3C) and CD38 ($P = 1.83 \times 10^{-2}$; Figure 3D). Sex (Figure 4A; Table 2) had a significant effect on ICAM-1 ($P = 1.02 \times 10^{-2}$; Figure 4B). Age (Figure 5; Table 2) had a significant effect on TLR-9 ($P = 4.43 \times 10^{-2}$; Figure 5B). Only the effect of generation on TLR-9 cell surface expression remained significant after Bonferroni correction for multiple hypothesis testing ($\beta = 5 \times 10^{-3}$).

Plate assignment had a significant effect (Bonferroni correction) on half of the phenotypes measured ($P < 5 \times 10^{-3}$; Table 2). Lines were randomized to sampling plates prior to any other steps (Materials and Methods). A given line had the same plate assignment in all replicates and all phenotype assays. Therefore, the significant plate effect could be due to experimental error or to differences in the phenotype value distributions between plates, as a result of the random assortment of a finite number of samples. If experimental error caused the plate effect, statistical correction of the plate effect would be necessary. If differences in phenotype distributions caused the plate effect, statistical correction would be undesirable, as it would alter accurate phenotype measurements unnecessarily. Therefore, determining the cause of the plate effect is important for all further data analyses.

Comparing of replicate variation between plates could suggest the cause of the plate effect. Because all three sampling plates were handled at the same time for each

independent freeze/thaw replicate, the experimental error between replicates of a sampling plate is expected to be the same as the set of all plates assayed. If experimental error caused the plate effect, variation between sampling plate replicates would be expected to be the same as the variation between plates. If differences in phenotype distributions caused the plate effect, variation between sampling plate replicates would be expected to be less than the variation between plates.

Because the scale phenotypic measurements vary by more than an order of magnitude, I used the coefficient of variation (CV) for this comparison. For phenotypes with a significant plate effect, the mean sampling plate CV ($\bar{c}_v = 0.05$) was the same as phenotypes without a significant plate effect ($\bar{c}_v = 0.05$), with one exception. The mean CV of $CD4_{high}$ ($\bar{c}_v = 0.18$) was inflated by an easily identified outlier replicate (Figure 6), the removal of which reduces the mean CV to the same level as other phenotypes ($c_v = 0.05$). In contrast, the sampling plate CV ($\bar{c}_v = 0.07$) was greater than the replicate CV ($\bar{c}_v = 0.05$) for phenotypes with a significant plate effect, but not for phenotypes without a significant plate effect ($\bar{c}_v = 0.02$). The difference between sampling plate CV and replicate CV increases with the significance of the plate effect (e.g., CD40: $P_{plate} = 6.33 \times 10^{-12}$; $c_v^{plate} = 0.19$; $c_v^{thaw} = 0.05$; Figure 6B). These results suggest that the plate effect is most likely the result of uneven distribution of certain phenotypes across the sampling plates during randomization and should not be statistically corrected as it primarily represents actual phenotypic variation between lines, not the effect of a covariate and experimental error.

Differences in cell surface expression between lines: The demonstration of consistency in phenotypic measurements between replicates is not sufficient for identifying the genetic basis of phenotypic variation. There must also be variation between lines that is intrinsic to the lines themselves, which is distinct from the background variation due to experimental noise. The proportion of the total phenotypic variation that can be explained by differences between lines was estimated:

$$L = 1 - \frac{\overline{\sigma}_E^2}{\sigma_P^2},$$

where L is the proportion of phenotypic variance explained by , σ_P^2 is the total phenotypic variance, and $\overline{\sigma}_E^2$ is the mean environmental variance. Intrinsic differences between lines explain the vast majority of variation in all phenotypes assayed ($0.75 \leq L \leq 0.91$; $\tilde{L} = 0.89$; Table 1).

Because there are known, plausible non-genetic components to variation between lines that are undefined, the proportion of the phenotypic variance explained by differences between lines does not necessarily represent the proportion of the phenotypic variance explained by differences in genotypes, but does represent the upper bound on the variation that might be explained by genetic variation. Almost all of the phenotypic variation is due to intrinsic difference between lines both for every phenotype assayed (Figure 2).

Narrow-sense heritability of cell surface protein expression: The observation that differences between lines explain most variation for all phenotypes suggests that a portion of the total phenotypic variation may be under the control of additive genetic

variation. The proportion of the total phenotypic variation controlled by additive genetic variation can be estimated as the narrow-sense heritability (FALCONER 1989).

The narrow-sense heritabilities of all phenotypes assayed were estimated (Table 1) with collaboration of Aldi Kraja and Michael Province from the Washington University School of Medicine Division of Statistical Genomics (Materials and Methods). The narrow-sense heritabilities ranged from $h^2 = 0.03 \pm 0.07$ for CD23 to $h^2 = 0.30 \pm 0.11$ ($P = 3.54 \times 10^{-5}$) for CD38. These heritabilities were significant for seven of ten phenotypes (Table 1), but are on the low end of the range observed for common phenotypes in human subjects (e.g., $0.06 \leq h^2 \leq 0.52$ for lung function) (OBER *et al.* 2001). Although most of the phenotypic variation is due to difference between lines, only a fraction of that variation is controlled by additive genetic variation (Figure 2).

Comparison to pilot study: I compared the results of the full panel to those of a pilot study that had been restricted to measuring auto-flourescence (1X PBS), CD40 cell surface expression, and CD86 cell surface expression. The pilot's sample set (Supplemental Table 2) partially overlapped ($N_{pilot} = 108$; $N_{\cap} = 79$) with the full panel (Supplemental Table 1). CD40 cell surface expression was more heritable in the pilot ($h^2_{CD40} = 0.42$) than in the full panel ($h^2_{CD40} = 0.19$). CD86 cell surface expression was less heritable in the pilot ($h^2_{CD86} = 0.12$) than in the full panel ($h^2_{CD86} = 0.24$).

In order to understand the causes of the divergence in narrow-sense heritability estimates, I compared measurements of cell surface expression of CD40 and CD86 in the samples shared samples by both the full panel and the pilot. Although cell surface

expression is highly repeatable within both the full panel (e.g., $r_{CD40} = 0.89$; Figure 7A) and the pilot (e.g., $r_{CD40} = 0.81$; Figure 7B), cell surface expression of CD40 ($r_{CD40} = 0.59$) and CD86 ($r_{CD86} = 0.69$) in the pilot was not highly correlated with the full panel. These incomplete correlations may reflect methodological differences between the full panel and the pilot.

Because methodological differences, especially when phenotyping by flow cytometry, may cause nonlinear changes in the phenotypic measurement of a line, but should not change the phenotypic rank order of the line, I confirmed the divergence between the two sets by calculating Spearman's rank correlation between the pilot and full panel for CD40 ($\rho_{CD40} = 0.61$; Figure 7C) and CD86 ($\rho_{CD86} = 0.72$; Figure 7D). The slight improvement in the rank correlation suggests that there may be a slight contribution of nonlinear effects. In light of the precision of these assays, as demonstrated above, these results likely represent real differences between the pilot and the full panel. Due to the methodological differences, it may be most accurate to consider the phenotypes, which are nominally the same, as being measured under different conditions and, therefore, not equivalent phenotypes.

DISCUSSION

Due to the challenges inherent in the study of genetic variation in humans, complementary approaches that do not require repeated recruitment or consent of large numbers of human subjects may have great utility. LCLs are non-adherent cell lines that are grown with simple culturing procedures allowing many, repeated samples to be taken

under conditions that are either technically or ethically impossible in human subjects. Most importantly LCLs are cell lines derived from human subjects. The genetic variation in LCLs is the genetic variation that is in the human population (BAUCHET *et al.* 2007; HE *et al.* 2009; LAO *et al.* 2008; MEUCCI *et al.* 2005; SMITH *et al.* 2006). These characteristics make LCLs an attractive model system for the study of human genetic variation.

Yet, these characteristics alone are not sufficient to make LCLs a productive model system for the study of human genetic variation. The identification of genetic variants that control phenotypic variation is dependent on the ability to accurately phenotype individuals and to identify variation between individuals in those phenotypes.

There have been previous efforts to describe some components of phenotypic variance in LCLs. Assay variance has been a focus in transcript abundance phenotypes, in order to define a significant change relative to background noise (MORLEY *et al.* 2004). While quantification of assay variance is of technical import, it is not relevant to the suitability of LCLs individually as a model system – assuming that the utility of the model system is not strictly limited by the reliability of the assays available for that system. Understanding the relative contributions of both experimental and genetic components to the total phenotypic variance is relevant to the evaluation of LCLs as a model system for human genetic variation. This work represents the first reported effort that separates the components of the phenotypic variance in a way that allows intrinsic differences between LCLs to be explicitly separated from experimental sources of variation (i.e., freeze/thaw and cell culture/assay variance).

A prior report by Choy *et al.* raised the concern that LCLs are unreliable due to excessive amounts of variation between freeze/thaw aliquots, but in their study freeze/thaw variance was confounded with cell culture/assay variance. They reported that phenotypic differences between aliquots within lines were larger than the differences between lines (CHOY *et al.* 2008). In contrast, having experimentally separated cell culture/assay variance and freeze/thaw variance, I found that both culture/assay replicates and freeze/thaw replicates were highly repeatable. The discrepancy between these conclusions has several possible explanations based on cell culture methods, phenotypes assayed, and data processing.

The cell culture methods used by Choy *et al.* may have contributed to increased experimental error relative to this work. The marginal decrease in repeatability from cell culture/assay replicates ($\bar{r} = 0.94$) to freeze/thaw replicates ($\bar{r} = 0.90$) in this work is consistent with the expectation that experimental error will increase with the number of manipulations. The cell culture methods used here involve no manipulations of the samples between thawing of a frozen aliquot and fixation of the culture (Materials and Methods). Choy *et al.*, however, counted cell densities daily and adjusted cell densities to a target population size over a seven-day cell culture period. If each manipulations represents an independent opportunity to introduce additional experimental error, then the decrease in experimental error here would be expected based on the cell culture methods used.

Data processing made a large, definable contribution to the improvement in repeatability relative to Choy *et al*. Any collection of large numbers of data points will

likely experience both technical failures and outliers, which are categorically distinct from the effects of covariates. Outliers may be due to statistical chance or experimental error. Repeatability of all phenotypes ($\bar{r} = 0.65$) was substantially improved by the identification of outliers in the data ($\bar{r} = 0.90$).

There are several strategies for dealing with outiers. Large replicate sample size can minimize the influence of outliers on the data without explicit identification of the outlying data points (WATTERS *et al.* 2004). Such sample sizes are not always possible, due to factors like cost or availability of samples. In these cases, outliers may be identified and removed, or the median, which is robust to outliers, may be used preferentially to the mean. Choy *et al*. only had two replicate samples per measurement, at which sample size outliers cannot be identified and the robust median becomes equivalent to the sensitive mean. Furthermore, the impact of outliers increases as the sample size decreases. The ability, due to experimental design, to identify outliers (for the calculation of experimental repeatability) and to use the median may explain, in part, the substantial improvement here relative to previous work with LCLs (CHOY *et al.* 2008).

The high repeatability of cells surface expression levels indicates that the majority of phenotypic variance was due to intrinsic, but not necessarily genetic, differences, between lines (Figure 2). This variance due to differences between lines may be divided into a genetic and non-genetic component. One potential source of non-genetic variation is potential covariates associated with lines in the panel, such as generation in the pedigree, sex, age, and plate assignment of the sample. The mixed model ANOVA demonstrated that, overall, the few known, potential covariates did not have significant

effects on phenotype. The significant plate effect is more likely due to differences in phenotype distributions (i.e., differences between plate median values are true differences) than due to bias or experimental error. Known, potential covariates were not significant contributors of non-genetic variation to the differences between lines.

While, all components of the genetic differences between lines, like epistatic and dominance variance, cannot be estimated, the additive component may be estimated due to the pedigree structure of the full panel. The narrow-sense heritability estimates for most phenotypes were significant, but fell on the low end of the range of heritability estimates for common human phenotypes (OBER *et al.* 2001). Only a fraction of the differences between lines is composed of additive genetic variance. The remainder is composed of undefined contributions from epistasis, dominance, and non-genetic components.

The pilot identified substantially greater narrow-sense heritability of CD40 cell surface expression compared to the full panel, which could indicate that the methods used in the pilot increase the additive genetic component of the differences between lines. Although, the narrow-sense heritability of CD86 cell surface expression was substantially greater in the full panel. Some of the methodological differences between the full panel and the pilot – the use of fluorescein (FITC) conjugated monoclonal antibodies, instead of phycoerithrin (PE) were used, requiring the use of different flow cytometer settings, antibody dilutions, and antibody incubation times - could cause non-linear scaling differences between the full panel and the pilot, resulting in imperfect correlation between the full panel and pilot. The rank correlation between the full panel and pilot,

however, is not substantially larger than the standard correlation, suggesting that the imperfect correlation is the result of real differences in the phenotypes measured in the full panel and pilot, not non-linear scaling effects. Because samples in the pilot were cultured in flasks, not plates, as in the full panel, it is possible that the phenotypes measured in the full panel and the pilot are not precisely equivalent, with a larger additive genetic component to CD40 cell surface expression and a smaller additive genetic component to CD86 cell surface expression when grown in flasks.

Unfortunately, the samples necessary to assess the other components of the non-genetic differences between lines, such as sampling and immortalization variation, do not exist for either the CEPH-UT or the International HapMap Project (2003; FRAZER *et al.* 2007; THE INTERNATIONAL HAPMAP 2005) LCL collections. Because the non-genetic components are undefined, it is not possible to estimate the non-additive genetic components. Contributions to the non-genetic variation from sampling and immortalization variation could be defined by analyzing multiple LCLs derived from a single sample and LCLs derived from multiple samples from a single individual. These samples are absent from current collections because these collections were designed for genotype mapping, not studying phenotypic effect of genetic variation. They would, however, be an important addition to future collections.

The ability to identify genotype-phenotype associations is dependent on the proportion of phenotypic variance that is controlled by additive genetic variance. While the non-additive component of the genetic variation is a product of the genetics of the system, the non-genetic variation can be minimized through the use of highly repeatable

experimental methods, such as those described here. Variation in the cell surface expression of the proteins measured was due to differences between lines. Because high experimental repeatability of cell surface expression is due to consistency between freeze/thaw replicates, I expect that this high repeatability will also be observed in other phenotypes, provided that similar cell culture methods, sample sizes, and data processing steps are used.

With appropriate handling and experimental design to control for experimental sources of variation, LCLs can be used as a reliable resource for the study of genetic variation. This effort would benefit from quantification of the effects of sampling and immortalization, as well as increased sample sizes, requiring extensive additional sampling within the pedigree structures that made it possible to accurately estimate components contributing to variation between LCLs. These studies may also benefit from expanding genetic variation studies in cell culture to other cell types, perhaps through the use of induced pluripotent stem cell methods.

It is not clear that there are any advantages to using localized protein expression as a phenotype when compared with transcript abundance in terms of the degree that phenotypic variation is controlled by additive genetic variance and the identification of QTL. The QTL identified for localized protein expression may still be more likely to be associated with traits in human subjects, like disease risk. Direct experimental confirmation of the causal hypotheses Establishing this relationship is an important goal for future research.

The measurement of phenotypes in LCLs is both highly repeatable within individual lines and variable between lines, important characteristics for any model system of genetic variation. The high repeatability of phenotypes measured here (cell surface expression of protein) is expected to extend to other phenotypes, including those that cannot ethically or technically be studied in human subjects. Provided that rigorous methods, reasonable data processing, and sensible experimental design are used, these conclusions may extend to other cell lines, like induced pluripotent stem cells, which would allow human genetic variation to be investigated in the cell type most relevant to the particular question. These characteristics, in combination with the developing capacity to test genotype-phenotype associations (SHUKLA *et al.* 2009) and the accessibility to independent research groups, strongly recommend cell culture as a complementary approach to the study of human genetic variation.

MATERIALS AND METHODS

Cell lines and culture conditions in the full panel: A panel of 255 lymphoblastoid cell lines (LCLs) was established from the CEPH-UT collection (Supplemental Table 1). Prior to any assays, a single, 1mL, frozen (-140°C) aliquot of each cell line was thawed in a 37°C water bath for five minutes. The thawed aliquot was diluted in 10mL 37°C 1X phosphate buffered saline (PBS) and centrifuged (290x*g* for five minutes at room temperature [RT]) to pellet the cells. The supernatant was aspirated and the cells were suspended in 1mL 37°C standard LCL media (media): RPMI-1640 media (Gibco), 15% fetal bovine serum (FBS) (Gibco), 2mM L-glutamine, 50μg/mL gentamycin (Gibco). The 1mL samples were diluted in 10mL 37°C media in 25cm^2 flasks (Corning) and

incubated at 37°C with 5% CO2 and 95% humidity. At five days post-thaw, 10mL 37°C media was added to each sample. At nine days post-thaw, samples were transferred into larger 75cm2 flasks (Corning) and an additional 10mL 37°C media was added. At twelve days post-thaw, 20mL 37°C was added for a total volume of 50mL.

On day 15 post-thaw, cell line samples were frozen as follows. 50mL samples were transferred to conical vials and pelleted by centrifugation (290x*g* for five minutes at RT). The supernatant was aspirated and the cells were suspended in 10mL freezing media (RPMI-1640 media, 20% FBS, and 10% DMSO). 1mL aliquots were distributed into 96-well deep well plates. Plates were incubated for 48 hours at -80°C and then at -140°C afterwards.

To prepare samples for phenotyping, frozen (-140°C) plates were thawed by incubation in a 37°C water bath for 30 minutes. 20μL samples were diluted in 180μL 37°C media in tissue culture treated 96-well flat bottom plates with low evaporation lids. Samples were incubated at 37°C with 5% CO_2 and 95% humidity prior to antibody labeling.

Quantification of cell surface expression in the full panel: On Day 5 after aliquot thawing, samples were fixed by adding a 150μL sample to 50μL 4% paraformaldehyde (PFA) in 96-well 2μM filter plates (Corning) and incubated for 10 minutes at RT. Media was removed using a vacuum manifold (Millipore) with a pressure between -5 mmHg and -10 mmHg. Cells were washed once with 100μL 1X PBS and then suspended in 100 μL of R-phycoerythrin (PE) conjugated monoclonal antibody (mAb) diluted 1:500 in 1XPBS/0.2% bovine serum albumin (BSA). All antibodies were from US Biological,

raised in mouse, and were IgG1 isotype. An αCD4-PE (IgG1) mAb was used as an isotype control. Samples were incubated for two hours at 4°C in the dark. Following antibody labeling, media was removed with a vacuum manifold. Cells were washed once in 100μL 1X PBS, suspended in 180μL 1X PBS, and transferred to 96-well round bottom assay plates. Samples were stored at 4°C in the dark until assayed by flow cytometry.

Quantification of cell surface fluorescence was performed using a Cytomics FC500MPL flow cytometer and data was processed using FlowJo 8.8 software. Individual samples were measured for 30 seconds using the low flow setting. PE fluorescence was measured in a channel defined by a 565±10nm band pass filter using one of two settings (low: PMT=750, gain=1.0; high: PMT=900, gain=1.0. Cell surface expression was defined as the median fluorescence magnitude for a sample.

Data were initially screened for samples that failed to grow. Any sample with fewer than a threshold number of cell counts during flow cytometry (n=75) was removed from analysis.

Outlier data points were identified as follows. For any phenotype (i) measurement for a line (j) combinations with three or more replicate samples after cell count filtering, a scaled deviation score (δ_{ijk}) was calculated for each sample based on the line median for each phenotype:

$$\delta_{ijk} = \frac{\left| x_{ijk} - \tilde{x}_{ij} \right|}{\tilde{x}_{ij}},$$

where δ_{ijk} is the scaled deviation for the phenotype of the k^{th} replicate of the j^{th} line for the i^{th} phenotype, x_{ijk} is the phenotypic measurement for the k^{th} replicate of the j^{th} line for

the i^{th} phenotype, and $\tilde{x}_{ij}$ is the median phenotypic measurement for the j^{th} line for the i^{th} phenotype. This metric was compared to a scaled deviation threshold ($\Delta = 0.13$). For any set of line measurements for a phenotype where at least one replicate's scaled deviation exceeded the scaled deviation threshold ($\delta_{ijk} > \Delta$), the replicate with the greatest scaled deviation was identified as the outlier and eliminated from further analysis.

Cell lines, culture conditions, and quantification of cell surface expression in the pilot: A panel of 108 lymphoblastoid cell lines (LCLs) was established from the CEPH-UT collection (Table 4). On Day 0, a single, 1mL, frozen aliquot was thawed for five minutes at 37°C in a water bath. The sample was diluted in 10mL 1X PBS, centrifuged at 290x*g* for five minutes at room temperature and the supernatant was removed. The pellet was resuspended in 1mL media. The sample was diluted in 10mL media in a 25cm^2 flask and incubated at 37°C with 5% CO_2 and 95% humidity for four days.

On Day 4, samples were fixed by aliquoting 75μL samples into 96-well 2μM filter plates and adding 25μL 4% PFA. Plates were incubated at room temperature for five minutes and then spun at 1000x*g* for ten minutes at 4°C. Samples were washed once in 100μL 1X PBS, resuspended in 90μL 1X PBS, and stored at 4°C until assayed.

Prior to the phenotype assay, 10μL of either 1X PBS or a 1:10 dilution of a flouroscein (FITC) conjugated mAb. Antibodies were specific to CD40 and CD86. Samples were incubated with mAb for ten minutes at room temperature in the dark and then spun at 1000x*g* for ten minutes at 4°C. Samples were washed once in 100μL 1X PBS, resuspended in 170μL 1X PBS, and 150μL was transferred to 96-well round bottom assay plates.

Quantification of cell surface fluorescence was performed as above with the following exceptions. FITC fluorescence was measured in a channel defined by a 565±10nm band pass filter using the following settings: PMT=1000, gain=1.0.

Estimation of narrow sense heritability: Narrow-sense heritability estimates used the multivariate and multilocus, variance components method in SEGPATH (PROVINCE *et al.* 2003) as described previously (WATTERS *et al.* 2004). Due to high kurtosis in the CD19 cell surface expression, median CD19 cell surface expression levels were $\log_{10}$ transformed prior to analysis.

SUPPLEMENTAL FILES

Complete data for all replicates of all lines for all phenotypes is available upon request.

Supplemental Table 1: Details of the full 255 LCL panel with phenotypic values.

Supplemental Table 2: Details of 79 LCLs from pilot that are shared with the full 255 LCL panel.

TABLES

Table 1: Phenotype distributions and variance components

Phenotype	N	$\bar{x}$ (AFU)	σ_P^2	σ_E^2	L	h^2	$SE(h^2)$	$P(h^2)$
$CD4_{low}$	249	17.57	13.93	1.67	0.88	0.20	0.09	1.78E-03
$CD4_{high}$	248	59.41	180.40	29.04	0.84	0.20	0.10	3.49E-03
CD19	251	64.37	178.61	15.62	0.91	0.20	0.11	5.50E-03
CD23	247	115.20	512.04	67.89	0.87	0.03	0.07	3.04E-01
CD38	249	23.83	27.62	2.57	0.91	0.30	0.11	3.54E-05
CD40	250	171.95	2585.54	277.39	0.89	0.19	0.09	1.55E-03
CD45RA	250	40.73	93.69	9.53	0.90	0.22	0.09	6.91E-04
CD86	249	93.07	351.28	38.43	0.89	0.24	0.12	2.76E-03
ICAM-1	250	46.66	160.24	24.40	0.85	0.09	0.09	1.21E-01
TLR-9	254	81.22	205.54	52.38	0.75	0.07	0.07	1.42E-01

Table 2: *P*-values for covariate effects from mixed model ANOVA

Phenotype	Generation	Sex	Age	Plate Assignment
$CD4_{low}$	0.316	0.473	0.467	*1.51E-02*
$CD4_{high}$	0.343	0.603	0.593	***5.19E-13***
CD19	*8.38E-03*	0.693	*4.43E-02*	***2.81E-03***
CD23	0.367	0.139	0.969	***9.11E-05***
CD38	*1.83E-02*	0.337	0.528	0.435
CD40	0.549	0.306	0.306	***6.33E-12***
CD45RA	0.676	0.951	7.89E-02	*8.42E-03*
CD86	0.159	0.522	0.361	***2.95E-03***
ICAM-1	7.20E-02	*1.02E-02*	0.641	5.32E-02
TLR-9	0.228	0.104	0.339	0.683

*Italicized text indicates nominal significance ($\alpha = 0.05$).

**Bold and italicized text indicates significance after Bonferroni correction ($\beta = 0.005$).

FIGURE LEGENDS

Figure 1: Distribution of cell surface expression of proteins compared to background. Histograms of the distribution of cell surface expression levels in arbitrary fluorescence units (AFU) across the entire panel relative to background controls. (A) TLR-9 (brown) control for cell surface localization versus $CD4_{high}$ (green). (B) CD19 (pink) versus $CD4_{high}$ (green) and TLR-9 (brown). (C) CD23 (pink) versus $CD4_{high}$ (green) and TLR-9 (brown). (D) CD86 (pink) versus $CD4_{high}$ (green) and TLR-9 (brown). (E) CD38 (pink) versus $CD4_{low}$ (green). (F) CD40 (pink) versus $CD4_{low}$ (green). (G) CD45RA (pink) versus $CD4_{low}$ (green). (H) ICAM-1 (pink) versus $CD4_{low}$ (green).

Figure 2: Components of phenotypic variance. Bars represent the cumulative contributions of different variance components (y-axis) to the total variance for each phenotype assayed (x-axis). The summed contributions of all variance components must equal the total phenotypic variance (i.e., 100% phenotypic variance explained) for each phenotype assayed. Therefore, the bars do not indicate the relative amounts of total variance of phenotypes in comparison to each other. Line variance (pink) indicates the amount of phenotypic variance that is due to differences between lines. Variance between cell culture/assay replicates is indicated by the cell culture/assay variance (teal). (B) Line variance (pink) indicates the amount of phenotypic variance that is due to differences between lines, including additive genetic variance (green). Error bars indicate standard error of narrow-sense heritability estimates. Statistically significant narrow-sense heritability estimates are indicated (*). Variation between freeze/thaw replicates is indicated by the freeze/thaw variance (teal), which includes cell culture/assay variance.

Differences between lines explain the majority of variation for all phenotypes assayed, but additive genetic variance does not explain a majority of the line variance for these phenotypes.

Figure 3: Effect of generation on phenotype. (A) Histogram of distribution of ages in the panel. Due to generation structure within the pedigrees in the panel, the distribution of ages is not uniform. (B) Scatter plot comparing median cell surface expression (AFU) of indicated proteins of the first generation (grandparents; x-axis) with the median cell surface expression of the second (parents; green) and third (progeny; pink) generations (y-axis). Error bars indicate the 25^{th} and 75^{th} percentile values for the respective generation. Phenotypes with a significant generation effect from the mixed model ANOVA are indicated (*: $P < 0.05$). (C) Histograms of TLR-9 cell surface expression (AFU), which had a significant generation effect ($P = 8.38 \times 10^{-3}$), in the full panel for the first (grandparents; blue), second (parents; green) and third (progeny; pink) generations. (D) Histograms of CD38 cell surface expression (AFU), which had a significant generation effect ($P = 1.83 \times 10^{-2}$), in the full panel for the first (grandparents; blue), second (parents; green), and third (progeny; pink) generations.

Figure 4: Effect of sex on phenotype. (A) Scatter plot compares the median cell surface expression level of proteins (as indicated) in arbitrary fluorescence units (AFU) in the panel for females (x-axis) and males (y-axis). Error bars indicate the 25^{th} and 75^{th} percentile values of their respective sex. Phenotypes with a significant generation effect from the mixed model ANOVA are indicated (*: $P < 0.05$). (B) Histograms of ICAM-1

cell surface expression (AFU), which had a significant sex effect ($P = 1.02 \times 10^{-2}$), in the full panel for females (green) and males (green) generations.

Figure 5: Effect of age on phenotype. (A) Relationship between age (years) and cell surface expression (AFU) shown as the best-fit linear regression of cell surface expression on age. (B) Scatter plot comparison of age (years; x-axis) with TLR-9 cell surface expression (brown; y-axis). TLR-9 cell surface expression had a significant age effect ($P = 4.43 \times 10^{-2}$) in the mixed model ANOVA. Best-fit linear regression of TLR-9 cell surface expression on age (blue) represents general trend of gradual increase in expression with age ($m = 0.052$ AFU/yr).

Figure 6: Effect of plate assignment on phenotype. Median cell surface expression values (AFU; y-axis) for both sampling plates (large, empty circles) and freeze/thaw replicates of sampling plates (small, filled circles) are shown. Sampling plates are distinguished by color (plate A: pink, plate B: green, and plate C: blue; see Supplemental Table 1 for plate assignments). Each phenotype is in a separate column (x-axis). Phenotypes with a significant plate assignment effect from the mixed model ANOVA are indicated by the phenotype label (*: $P < 0.05$; and after Bonferroni correction **: $P < 0.005$).

Figure 7: Comparison of the full panel and pilot. (A) Scatter plot of CD40 cell surface expression level (AFU) for two, representative replicates in the full panel. (B) Scatter plot of CD40 cell surface expression level (AFU) for two, representative replicates in the pilot. (C) Scatter plot of CD40 cell surface expression rank in the respective sample set for LCLs shared by the pilot (x-axis) and the full panel (y-axis). (D) Scatter plot of CD86

cell surface expression rank in the respective sample set for LCLs shared by the pilot (x-axis) and the full panel (y-axis).

Figure 1

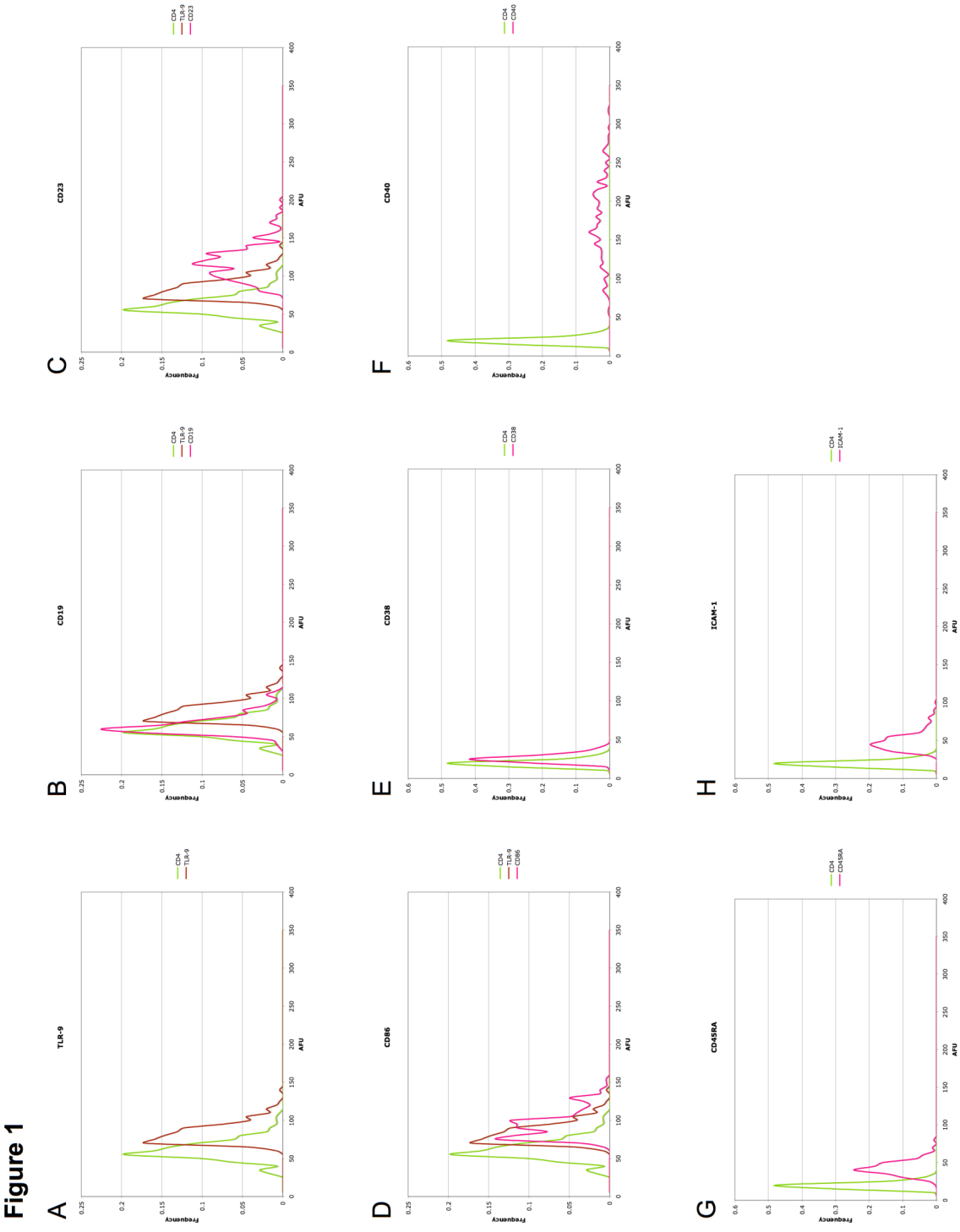
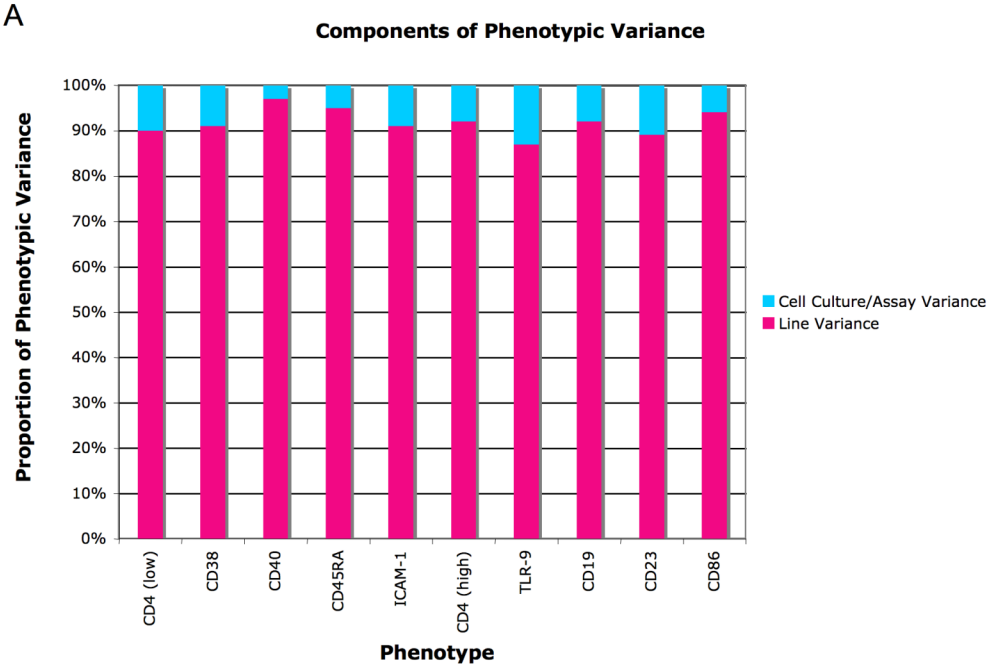

A
TLR-9
CD4
TLR-9
B
CD19
CD4
TLR-9
CD19
C
CD23
CD4
TLR-9
CD23
D
CD86
CD4
TLR-9
CD86
E
CD38
CD4
CD38
F
CD40
CD4
CD40
G
CD45RA
CD4
CD45RA
H
ICAM-1
CD4
ICAM-1
Frequency
AFU

Figure 2

A


Components of Phenotypic Variance
Proportion of Phenotypic Variance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
CD4 (low)
CD38
CD40
CD45RA
ICAM-1
CD4 (high)
TLR-9
CD19
CD23
CD86
Phenotype
Cell Culture/Assay Variance
Line Variance

B

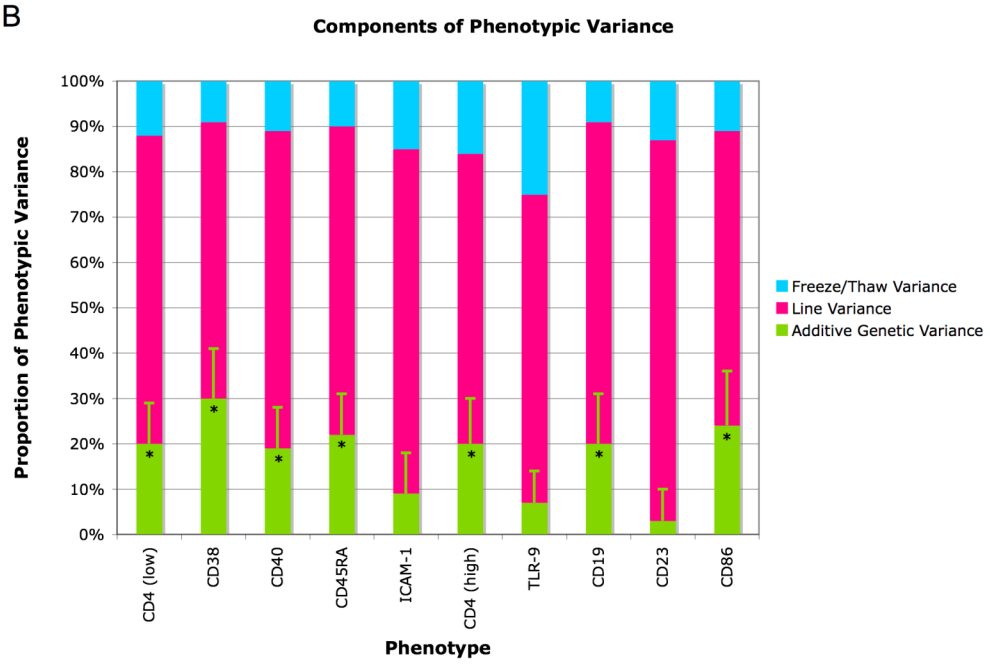

Components of Phenotypic Variance
Proportion of Phenotypic Variance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
CD4 (low)
CD38
CD40
CD45RA
ICAM-1
CD4 (high)
TLR-9
CD19
CD23
CD86
Phenotype
Freeze/Thaw Variance
Line Variance
Additive Genetic Variance

Figure 3

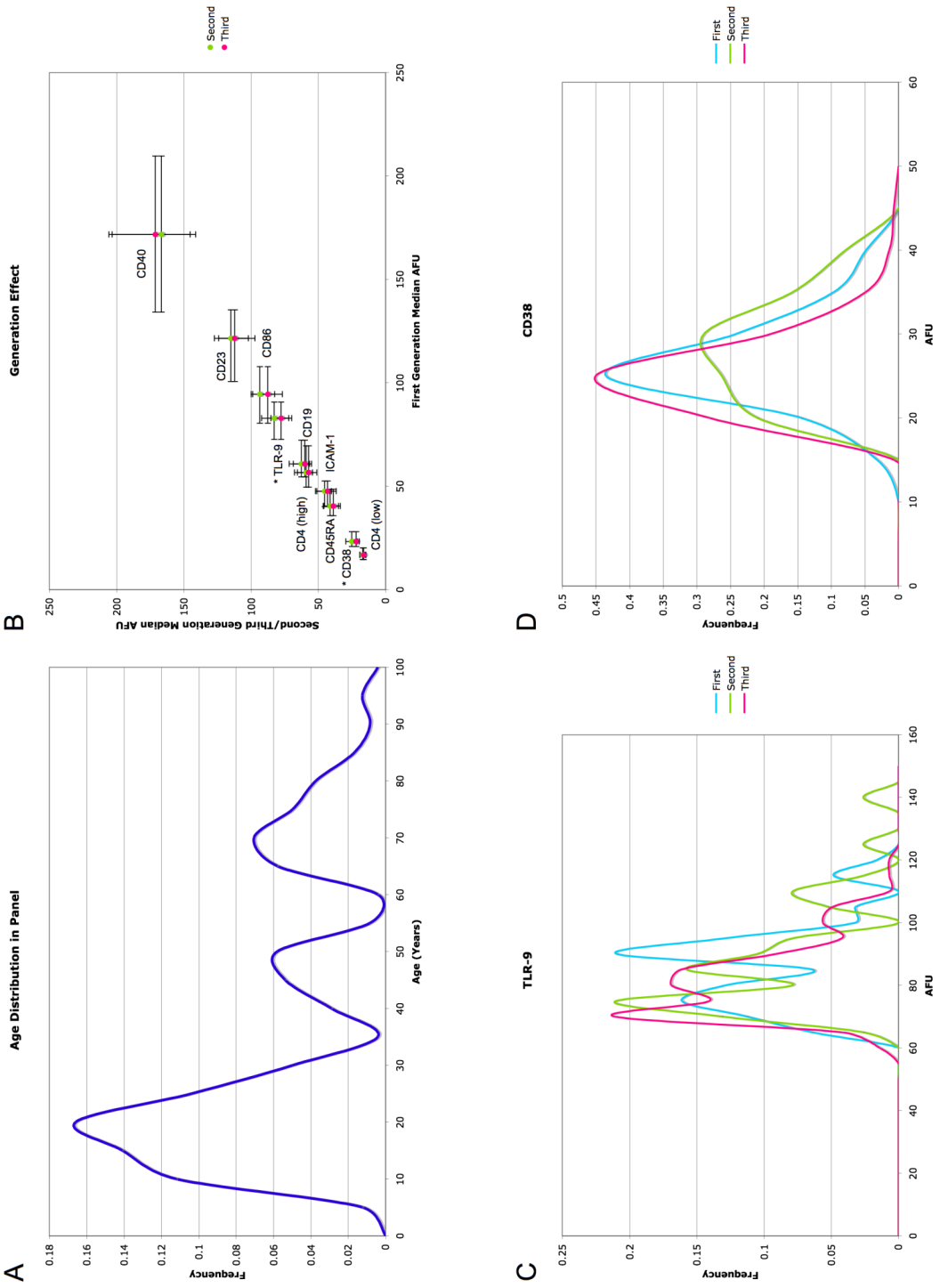
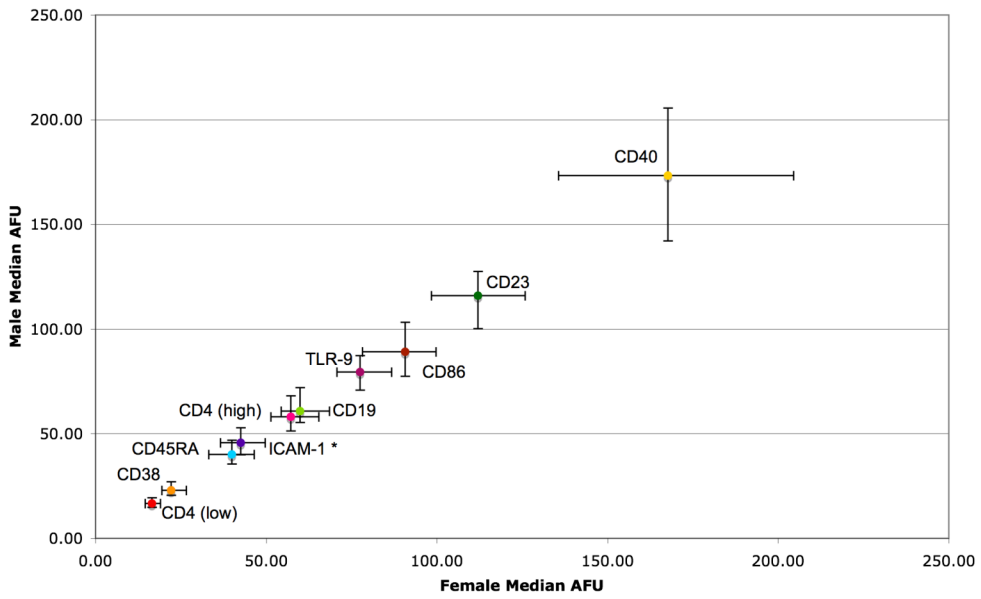

A
Age Distribution in Panel
Frequency
Age (Years)
B
Generation Effect
Second
Third
Second/Third Generation Median AFU
First Generation Median AFU
CD40
CD23
CD86
* TLR-9
CD19
CD4 (high)
ICAM-1
CD45RA
* CD38
CD4 (low)
C
TLR-9
First
Second
Third
Frequency
AFU
D
CD38
First
Second
Third
Frequency
AFU

Figure 4

A


Sex Effect
Male Median AFU
Female Median AFU
250.00
200.00
150.00
100.00
50.00
0.00
0.00
50.00
100.00
150.00
200.00
250.00
CD40
CD23
TLR-9
CD86
CD4 (high)
CD19
CD45RA
ICAM-1 *
CD38
CD4 (low)

B

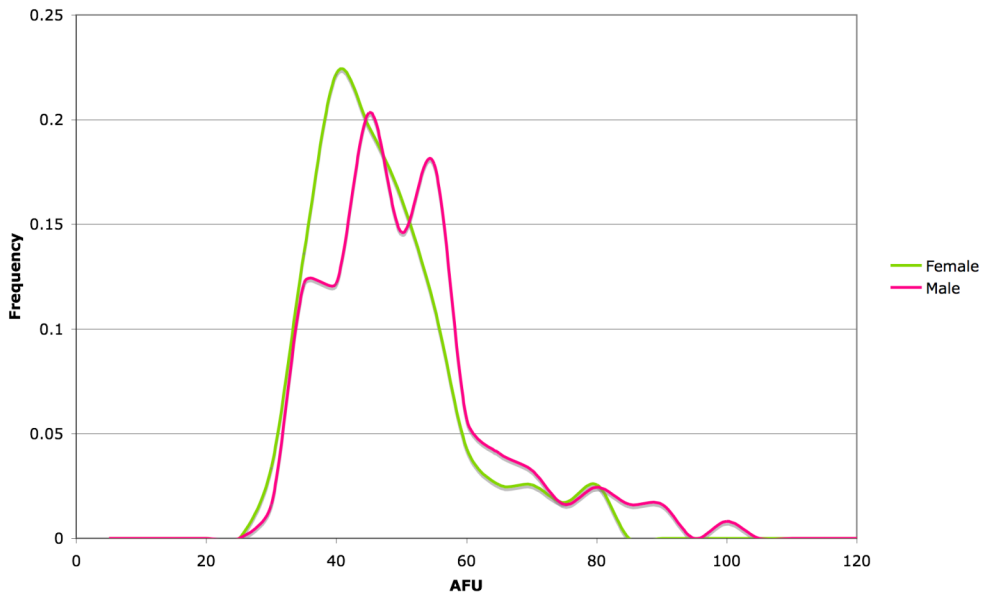

ICAM-1
Frequency
AFU
0.25
0.2
0.15
0.1
0.05
0
0
20
40
60
80
100
120
Female
Male

Figure 5

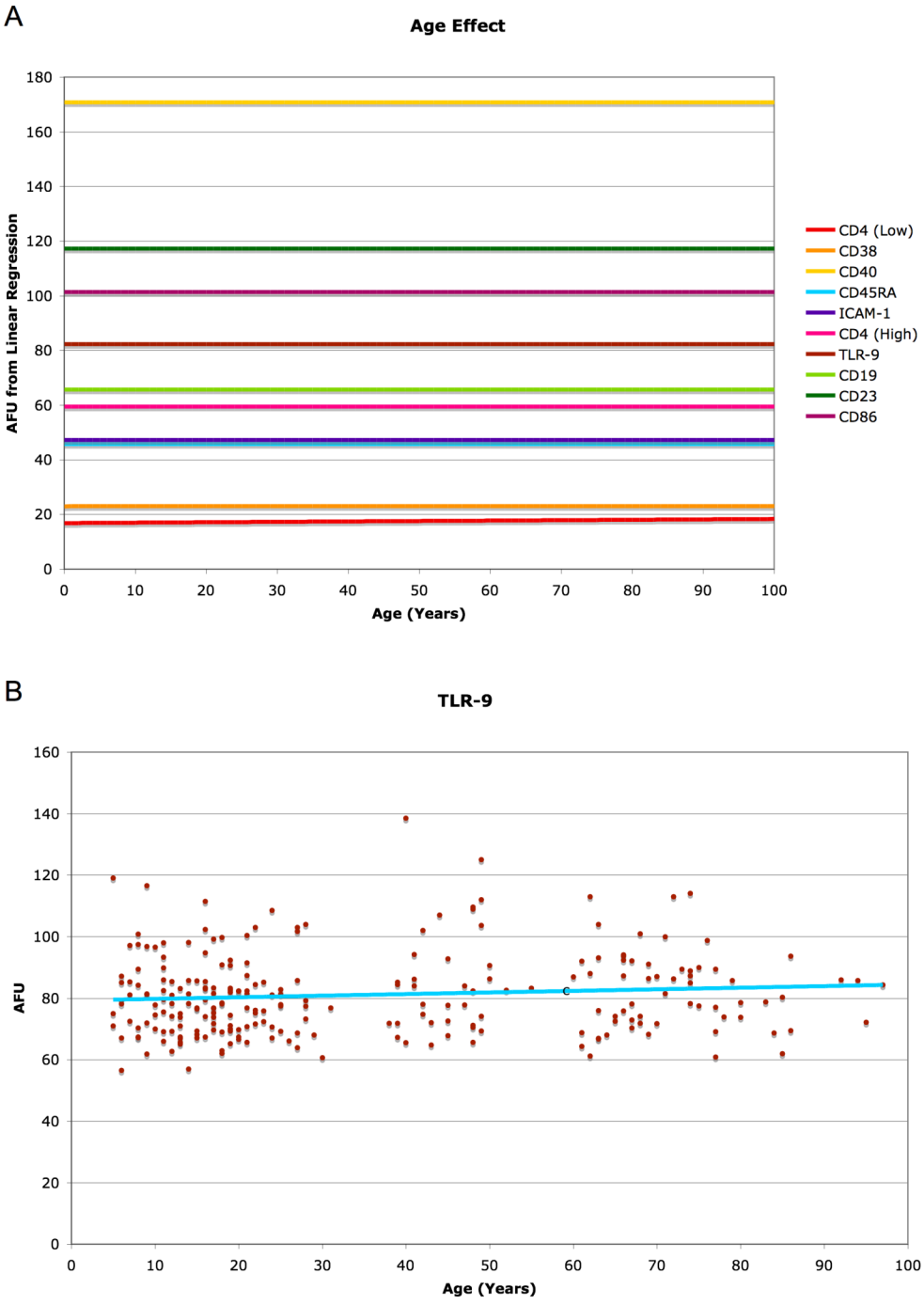

A
Age Effect
AFU from Linear Regression
Age (Years)
CD4 (Low)
CD38
CD40
CD45RA
ICAM-1
CD4 (High)
TLR-9
CD19
CD23
CD86
B
TLR-9
AFU
Age (Years)

Figure 6

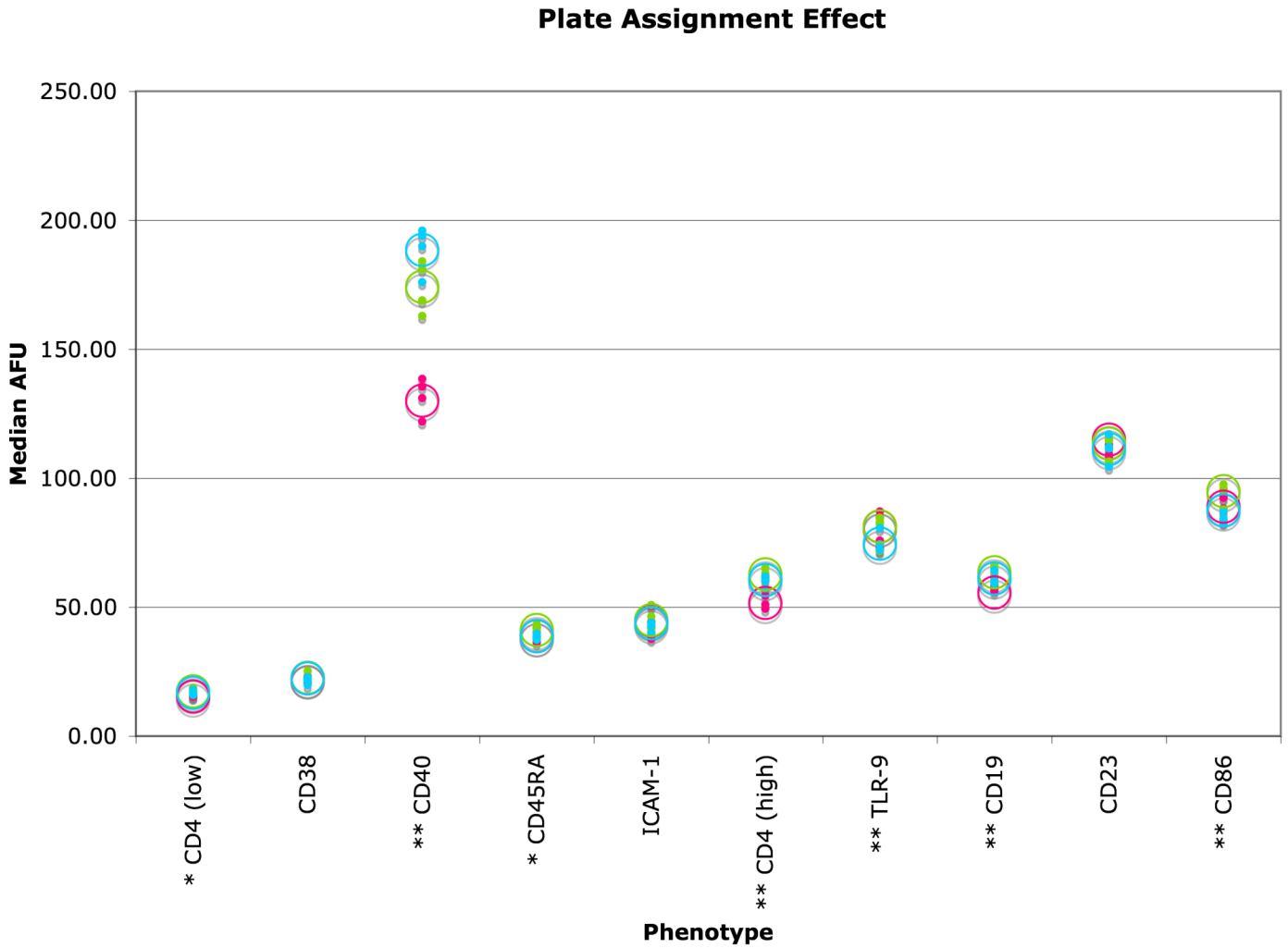
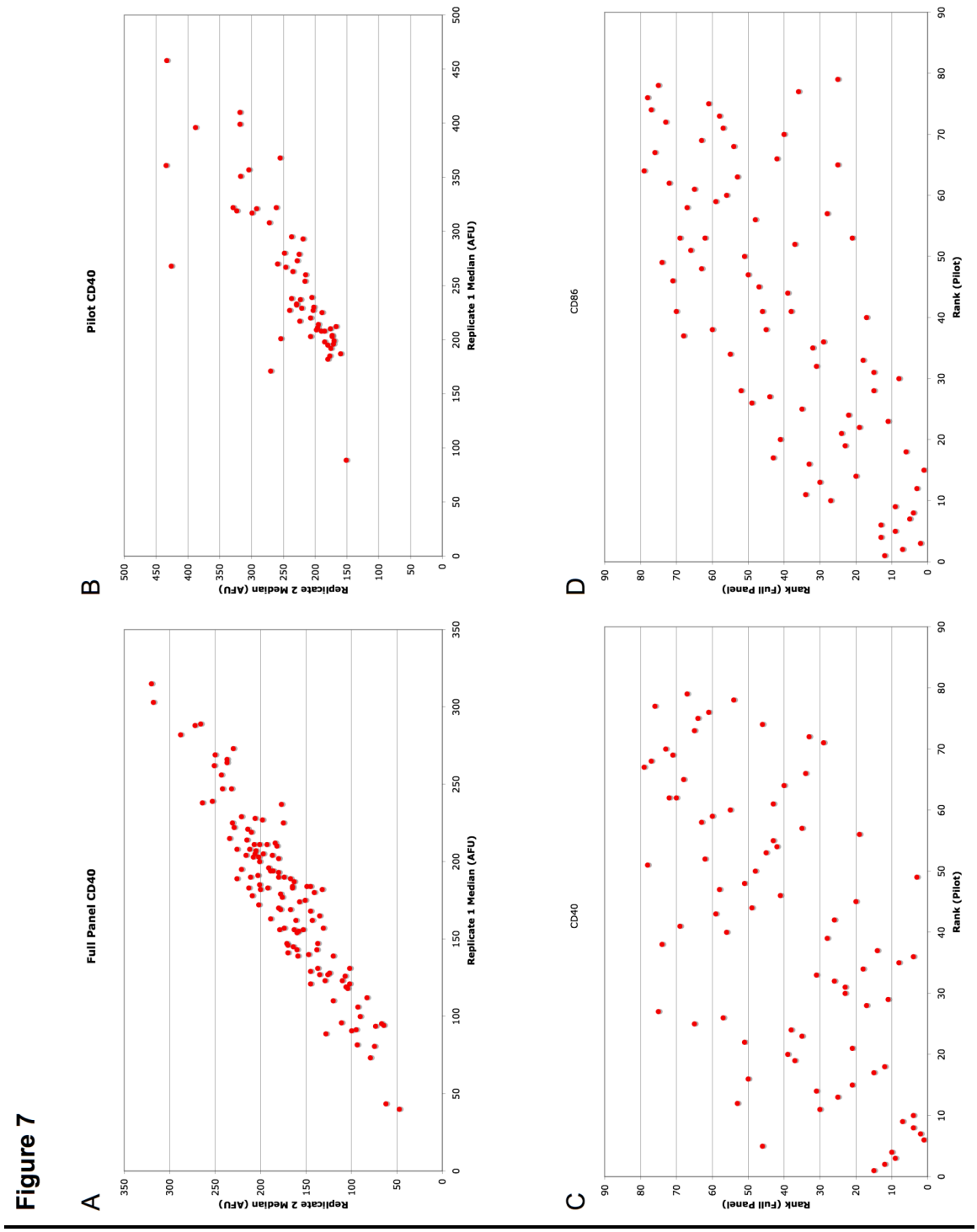


Plate Assignment Effect
250.00
200.00
150.00
100.00
50.00
0.00
Median AFU
* CD4 (low)
CD38
** CD40
* CD45RA
ICAM-1
** CD4 (high)
** TLR-9
** CD19
CD23
** CD86
Phenotype

Figure 7


A
Full Panel CD40
Replicate 2 Median (AFU)
Replicate 1 Median (AFU)
B
Pilot CD40
Replicate 2 Median (AFU)
Replicate 1 Median (AFU)
C
CD40
Rank (Full Panel)
Rank (Pilot)
D
CD86
Rank (Full Panel)
Rank (Pilot)

LITERATURE CITED

2003 The International HapMap Project. Nature **426:** 789-796.

ATLAS, S. A., E. S. VESELL and D. W. NEBERT, 1976 Genetic control of interindividual variations in the inducibility of aryl hydrocarbon hydroxylase in cultured human lymphocytes. Cancer Res **36:** 4619-4630.

BALLERINI, C., E. ROSATI, M. SALVETTI, G. RISTORI, S. CANNONI *et al.*, 2002 Protein tyrosine phosphatase receptor-type C exon 4 gene mutation distribution in an Italian multiple sclerosis population. Neurosci Lett **328:** 325-327.

BAUCHET, M., B. MCEVOY, L. N. PEARSON, E. E. QUILLEN, T. SARKISIAN *et al.*, 2007 Measuring European Population Stratification with Microarray Genotype Data. The American Journal of Human Genetics **80:** 948-956.

BERGEN, A. W., A. BACCARELLI, T. K. MCDANIEL, K. KUHN, R. PFEIFFER *et al.*, 2007 Cis sequence effects on gene expression. BMC Genomics **8:** 296.

BLEIBEL, W. K., S. DUAN, R. S. HUANG, E. O. KISTNER, S. J. SHUKLA *et al.*, 2009 Identification of genomic regions contributing to etoposide-induced cytotoxicity. Hum Genet **125:** 173-180.

CALAMINICI, M., K. PIPER, A. M. LEE and A. J. NORTON, 2004 CD23 expression in mediastinal large B-cell lymphomas. Histopathology **45:** 619-624.

CHEUNG, V., L. CONLIN, T. WEBER, M. ARCARO, K. JEN *et al.*, 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet **33:** 422-425.

CHEUNG, V. G., and W. J. EWENS, 2006 Heterozygous carriers of Nijmegen Breakage Syndrome have a distinct gene expression phenotype. Genome Res **16:** 973-979.

CHEUNG, V. G., R. S. SPIELMAN, K. G. EWENS, T. M. WEBER, M. MORLEY *et al.*, 2005 Mapping determinants of human gene expression by regional and genome-wide association. Nature **437:** 1365-1369.

CHOY, E., R. YELENSKY, S. BONAKDAR, R. PLENGE, R. SAXENA *et al.*, 2008 Genetic Analysis of Human Traits In Vitro: Drug Response and Gene Expression in Lymphoblastoid Cell Lines. PLoS Genet **4:** e1000287.

CLOOS, J., E. J. NIEUWENHUIS, D. I. BOOMSMA, D. J. KUIK, M. L. VAN DER STERRE *et al.*, 1999 Inherited susceptibility to bleomycin-induced chromatid breaks in cultured peripheral blood lymphocytes. J Natl Cancer Inst **91:** 1125-1130.

CORREA, C. R., and V. G. CHEUNG, 2004 Genetic variation in radiation-induced expression phenotypes. Am J Hum Genet **75:** 885-890.

CORYDON, T. J., A. HAAGERUP, T. G. JENSEN, H. G. BINDERUP, M. S. PETERSEN *et al.*, 2007 A functional CD86 polymorphism associated with asthma and related allergic disorders. J Med Genet **44:** 509-515.

DAUSSET, J., H. CANN, D. COHEN, M. LATHROP, J. M. LALOUEL *et al.*, 1990 Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics **6:** 575-577.

DEAGLIO, S., S. AYDIN, T. VAISITTI, L. BERGUI and F. MALAVASI, 2008 CD38 at the junction between prognostic marker and therapeutic target. Trends Mol Med **14:** 210-218.

DEUTSCH, S., 2005 Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. Human Molecular Genetics **14:** 3741-3749.

DOLAN, M. E., K. G. NEWBOLD, R. NAGASUBRAMANIAN, X. WU, M. J. RATAIN *et al.*, 2004 Heritability and linkage analysis of sensitivity to cisplatin-induced cytotoxicity. Cancer Res **64:** 4353-4356.

DUAN, S., W. BLEIBEL, R. HUANG, S. SHUKLA, X. WU *et al.*, 2007 Mapping Genes that Contribute to Daunorubicin-Induced Cytotoxicity. Cancer Research **67:** 5425-5433.

DUAN, S., R. S. HUANG, W. ZHANG, S. MI, W. K. BLEIBEL *et al.*, 2009 Expression and alternative splicing of folate pathway genes in HapMap lymphoblastoid cell lines. Pharmacogenomics **10:** 549-563.

FALCONER, D. S., 1989 *Introduction to quantitative genetics*. Longman

Wiley, Burnt Mill, Harlow, Essex, England

New York.

FORD, B. N., D. WILKINSON, E. M. THORLEIFSON and B. L. TRACY, 2001 Gene expression responses in lymphoblastoid cells after radiation exposure. Radiat Res **156:** 668-671.

FRAZER, K. A., D. G. BALLINGER, D. R. COX, D. A. HINDS, L. L. STUVE *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449:** 851-861.

GERKE, J., K. LORENZ and B. COHEN, 2009 Genetic interactions between transcription factors cause natural variation in yeast. Science **323:** 498-501.

GERKE, J. P., C. T. CHEN and B. A. COHEN, 2006 Natural isolates of Saccharomyces cerevisiae display complex genetic variation in sporulation efficiency. Genetics **174:** 985-997.

GOMEZ-LIRA, M., M. LIGUORI, C. MAGNANI, D. BONAMINI, A. SALVIATI *et al.*, 2003 CD45 and multiple sclerosis: the exon 4 C77G polymorphism (additional studies and meta-analysis) and new markers. J Neuroimmunol **140:** 216-221.

HARRIS, S. L., G. GIL, H. ROBINS, W. HU, K. HIRSHFIELD *et al.*, 2005 Detection of functional single-nucleotide polymorphisms that affect apoptosis. Proc Natl Acad Sci U S A **102:** 16297-16302.

HARTFORD, C. M., S. DUAN, S. M. DELANEY, S. MI, E. O. KISTNER *et al.*, 2009 Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. Blood **113:** 2145-2153.

HE, M., J. GITSCHIER, T. ZERJAL, P. DE KNIJFF, C. TYLER-SMITH *et al.*, 2009 Geographical Affinities of the HapMap Samples. PLoS One **4:** e4684.

HUANG, R. S., S. DUAN, W. K. BLEIBEL, E. O. KISTNER, W. ZHANG *et al.*, 2007a A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proc Natl Acad Sci U S A **104:** 9758-9763.

HUANG, R. S., S. DUAN, E. O. KISTNER, C. M. HARTFORD and M. E. DOLAN, 2008a Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. Mol Cancer Ther **7:** 3038-3046.

HUANG, R. S., S. DUAN, E. O. KISTNER, W. ZHANG, W. K. BLEIBEL *et al.*, 2008b Identification of genetic variants and gene expression relationships associated with pharmacogenes in humans. Pharmacogenet Genomics **18:** 545-549.

HUANG, R. S., S. DUAN, S. J. SHUKLA, E. O. KISTNER, T. A. CLARK *et al.*, 2007b Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. Am J Hum Genet **81:** 427-437.

HUANG, R. S., E. O. KISTNER, W. K. BLEIBEL, S. J. SHUKLA and M. E. DOLAN, 2007c Effect of population and gender on chemotherapeutic agent-induced cytotoxicity. Mol Cancer Ther **6:** 31-36.

JAMROZIAK, K., Z. SZEMRAJ, O. GRZYBOWSKA-IZYDORCZYK, J. SZEMRAJ, M. BIENIASZ *et al.*, 2009 CD38 gene polymorphisms contribute to genetic susceptibility to B-cell chronic lymphocytic leukemia: evidence from two case-control studies in Polish Caucasians. Cancer Epidemiol Biomarkers Prev **18:** 945-953.

JEN, K. Y., and V. G. CHEUNG, 2003 Transcriptional response of lymphoblastoid cells to ionizing radiation. Genome Res **13:** 2092-2100.

JURISIC, V., N. COLOVIC, N. KRAGULJAC, H. D. ATKINSON and M. COLOVIC, 2008 Analysis of CD23 antigen expression in B-chronic lymphocytic leukaemia and its correlation with clinical parameters. Med Oncol **25:** 315-322.

KWAN, T., D. BENOVOY, C. DIAS, S. GURD, D. SERRE *et al.*, 2007 Heritability of alternative splicing in the human genome. Genome Res **17:** 1210-1218.

LAO, O., T. T. LU, M. NOTHNAGEL, O. JUNGE, S. FREITAG-WOLF *et al.*, 2008 Correlation between Genetic and Geographic Structure in Europe. Current Biology **18:** 1241-1248.

LETTRE, G., 2009 Genetic regulation of adult stature. Curr Opin Pediatr **21:** 515-522.

LI, L., B. FRIDLEY, K. KALARI, G. JENKINS, A. BATZLER *et al.*, 2008 Gemcitabine and cytosine arabinoside cytotoxicity: association with lymphoblastoid cell expression. Cancer Res **68:** 7050-7058.

LI, L., B. L. FRIDLEY, K. KALARI, G. JENKINS, A. BATZLER *et al.*, 2009 Gemcitabine and Arabinosylcytosin Pharmacogenomics: Genome-Wide Association and Drug Response Biomarkers. PLoS One **4:** e7765.

LOEUILLET, C., S. DEUTSCH, A. CIUFFI, D. ROBYR, P. TAFFE *et al.*, 2008 In vitro whole-genome analysis identifies a susceptibility locus for HIV-1. PLoS Biol **6:** e32.

MEUCCI, M. A., S. MARSH, J. W. WATTERS and H. L. MCLEOD, 2005 CEPH individuals are representative of the European American population: implications for pharmacogenetics. Pharmacogenomics **6:** 59-63.

MONKS, S. A., A. LEONARDSON, H. ZHU, P. CUNDIFF, P. PIETRUSIAK *et al.*, 2004 Genetic inheritance of gene expression in human cell lines. Am J Hum Genet **75:** 1094-1105.

MORLEY, M., C. MOLONY, T. WEBER, J. DEVLIN, K. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430:** 743-747.

OBER, C., M. ABNEY and M. S. MCPEEK, 2001 The genetic dissection of complex traits in a founder population. Am J Hum Genet **69:** 1068-1079.

POKROVSKAJA, K., B. EHLIN-HENRIKSSON, J. BARTKOVA, J. BARTEK, R. SCUDERI *et al.*, 1996 Phenotype-related differences in the expression of D-type cyclins in human B cell-derived lines. Cell Growth Differ **7:** 1723-1732.

PROVINCE, M., T. RICE, I. BORECKI, C. GU, A. KRAJA *et al.*, 2003 Multivariate and multilocus variance components method, based on structural relationships to assess quantitative trait linkage via SEGPATH. Genet. Epidemiol. **24:** 128-138.

RAYCHAUDHURI, S., E. REMMERS, A. LEE, R. HACKETT, C. GUIDUCCI *et al.*, 2008 Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet **40:** 1216-1223.

SATOH, M., T. YASUDA, T. HIGAKI, M. GOTO, S. TANUMA *et al.*, 2003 Innate apoptosis of human B lymphoblasts transformed by Epstein-Barr virus: modulation by cellular immortalization and senescence. Cell Struct Funct **28:** 61-70.

SCHORK, N. J., J. P. GARDNER, L. ZHANG, D. FALLIN, B. THIEL *et al.*, 2002 Genomic association/linkage of sodium lithium countertransport in CEPH pedigrees. Hypertension **40:** 619-628.

SCHWARZMEIER, J. D., R. HUBMANN, M. DUCHLER, U. JAGER and M. SHEHATA, 2005 Regulation of CD23 expression by Notch2 in B-cell chronic lymphocytic leukemia. Leuk Lymphoma **46:** 157-165.

SHUKLA, S. J., and M. E. DOLAN, 2005 Use of CEPH and non-CEPH lymphoblast cell lines in pharmacogenetic studies. Pharmacogenomics **6:** 303-310.

SHUKLA, S. J., S. DUAN, J. A. BADNER, X. WU and M. E. DOLAN, 2008 Susceptibility loci involved in cisplatin-induced cytotoxicity and apoptosis. Pharmacogenet Genomics **18:** 253-262.

SHUKLA, S. J., S. DUAN, X. WU, J. A. BADNER, K. KASZA *et al.*, 2009 Whole-genome approach implicates CD44 in cellular resistance to carboplatin. Hum Genomics **3:** 128-142.

SMIRNOV, D. A., M. MORLEY, E. SHIN, R. S. SPIELMAN and V. G. CHEUNG, 2009 Genetic analysis of radiation-induced changes in human gene expression. Nature **459:** 587-591.

SMITH, E. M., X. WANG, J. LITTRELL, J. ECKERT, R. COLE *et al.*, 2006 Comparison of linkage disequilibrium patterns between the HapMap CEPH samples and a family-based cohort of Northern European descent. Genomics **88:** 407-414.

SZVETKO, A. L., A. JONES, J. MACKENZIE, L. TAJOURI, P. A. CSURHES *et al.*, 2009 An investigation of the C77G and C772T variations within the human protein tyrosine phosphatase receptor type C gene for association with multiple sclerosis in an Australian population. Brain Res **1255:** 148-152.

TAYLOR, D. K., E. ITO, M. THORN, K. SUNDAR, T. TEDDER *et al.*, 2006 Loss of tolerance of anti-dsDNA B cells in mice overexpressing CD19. Mol Immunol **43:** 1776-1790.

THE INTERNATIONAL HAPMAP, C., 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

WANG, L., A. L. OBERG, Y. W. ASMANN, H. SICOTTE, S. K. MCDONNELL *et al.*, 2009a Genome-wide transcriptional profiling reveals microRNA-correlated genes and biological processes in human lymphoblastoid cell lines. PLoS One **4:** e5878.

WANG, Q. L., B. H. LI, B. LIU, Y. B. LIU, Y. P. LIU *et al.*, 2009b Polymorphisms of the ICAM-1 exon 6 (E469K) are associated with differentiation of colorectal cancer. J Exp Clin Cancer Res **28:** 139.

WATTERS, J. W., A. KRAJA, M. A. MEUCCI, M. A. PROVINCE and H. L. MCLEOD, 2004 Genome-wide discovery of loci influencing chemotherapy cytotoxicity. Proc Natl Acad Sci U S A **101:** 11809-11814.

WEI, Q., M. R. SPITZ, J. GU, L. CHENG, X. XU *et al.*, 1996 DNA repair capacity correlates with mutagen sensitivity in lymphoblastoid cell lines. Cancer Epidemiol Biomarkers Prev **5:** 199-204.

ZHANG, W., S. DUAN, W. K. BLEIBEL, S. A. WISEL, R. S. HUANG *et al.*, 2009 Identification of common genetic variants that account for transcript isoform variation between human populations. Hum Genet **125:** 81-93.

ZHU, X. P., X. Q. YANG, Z. FU, H. G. YU, G. L. LIAN *et al.*, 2004 [Expression of CD 86 on monocyte and B cell surface, the level of T(H)-derived cytokine and their correlation in children with asthma]. Zhonghua Er Ke Za Zhi **42:** 83-86.

CHAPTER FOUR: CONCLUSION AND DISCUSSION

INTRODUCTION

The primary utility of cell culture model systems for the study of genetic variation lies in the control over experimental conditions and the experimental tractability. Their primary weakness lies in the fact that cell culture model systems are somewhat artificial and divorced from natural conditions; even in those using mircoorganims or those in which the genetic variation is the same as variation observed in natural populations. While the phenotypes measured for cell culture model systems in the lab may be analogous or similar to those with fitness effects outside the lab, one cannot argue that they are the same. Therefore, cell culture model systems are only a useful resource for research when the question and application allow the utility of the system to overcome the weakness. Unfortunately, many applications of cell culture model systems do not maximize their utility. I have examined the use of two cell culture model systems, laboratory selection on yeast and human-derived LCL pedigrees, for the study of natural genetic variation.

LABORATORY SELECTION ON YEAST AS A MODEL SYSTEM

In this work, I have assessed the use of laboratory selection on *S. cerevisiae* as a model system for natural genetic variation. Although laboratory selections have been used experimentally to test aspects of evolutionary theory, their use as a resource to overcome the limitations of natural isolates in the study of natural genetic variation has not been explored. The selected lines described in Chapter Two had adaptive phenotypes with a complex genetic basis similar to that observed in natural isolates of *S. cerevisiae*,

confirming that laboratory selection could be used to generate genetic variation with the characteristics of natural variation. This alone, however, does not demonstrate that laboratory selection on yeast is a useful model system for natural genetic variation. Evaluating laboratory selection on yeast as a model system is dependent on both the genetic variation in the derived lines and the particular advantages of yeast as an experimental organism.

There are many advantages to the use of microorganisms and yeast, in particular, for the study of genetic variation; but there are a few advantages specific to yeast strains developed in the laboratory that make them particularly attractive complements to the use of natural isolates to study natural genetic variation. The specific advantages of laboratory selection on yeast focus around two points: the ability to observe the entire evolutionary history of the derived lines and to study many replicates simultaneously.

The ability to observe the entire evolutionary history of a line is based on two factors. First, samples can be taken and stored throughout the process. These samples can be revived later to allow direct comparison between different time points in the evolutionary history. Second, the ancestral population can be defined, allowing both the response to selection to be quantified and the genetic underpinnings of that response to be investigated. Although the lines studied in this work represent only the ancestral sample and the derived lines without any intermediate samples, simple knowledge of the selection procedure permitted additional conclusions (i.e., that major effect mutations fixed in the population during the initial selection) to be drawn with confidence that

would not have been possible had the evolutionary history been unknown, as in the case of natural isolates.

The ability to define the ancestral population and to control the cell culture environment is critical for the use of multiple replicates and conditions in laboratory selections. Defining the ancestral population ensures that the initial genotype for all replicates is the same. Not only does this control an important variable during selection, but it may also simplify the identification of adaptive polymorphisms due to the reduced background noise of neutral polymorphisms (GRESHAM *et al.* 2008; GRESHAM *et al.* 2006). Similarly, the ability to control the cell culture environment during selection allows the variable of interest to be manipulated while keeping all other environmental variables either constant or documented between lines. Genotype and environmental control are important advantages for laboratory selections that allow both rigorous experimentation and the use of many replications.

The ability to usefully exploit multiple replicates is also dependent on the development of methods to efficiently and accurately phenotype samples. I developed a high-throughput, quantitative method for phenotyping yeast to identify the response to selection. The ability to compare phenotypes by accurately and efficiently quantify the response to selection alone, however, is not sufficient for genetic analysis of replicates. The ability to efficiently compare the replicate genotypes is also needed.

Next-generation sequencing and microarray technology make it possible to identify polymorphisms throughout the entire genome in yeast. The associated costs of this approach, however, become impractical, as the number of replicate samples gets

larger. Because the high-throughput, quantitative growth assay described in this work could easily be used for multiple growth conditions in liquid media (e.g., carbon source, amino acid depletion, toxic challenge, etc.), the distribution of phenotypic value for characters that were not under selection can be used to obtain information on genotypic similarities between lines. Differences in these characters between lines may result from genetic variants that were selectively neutral, but affect the character in question. The characters may also be correlated characters – traits that are not directly under selection, but change due to pleiotropic effects from adaptive variants under selection. Phenotypic similarity for these characters would suggest genotypic similarity. The distance between the phenotypic values of lines for these characters (plotted in *n*-dimensional space, where *n* is the number of characters assayed) is expected to be proportional to the genotypic distance between lines. Two lines possessing the same adaptive, genetic variants (i.e., the same genotype) would not only have identical phenotypic values for the character under selection, but would have identical phenotypic values for all correlated characters measured. The high-throughput, quantitative growth assay provides the opportunity to estimate genotypic similarities between lines in order to prioritize comparisons for more labor intensive and costly analyses, such as crosses and genotyping.

Replicate samples would also be particularly useful for questions of frequency and probability in adaptive evolution. For example, two independent studies examining sporulation efficiency in two different sets of *S. cerevisiae* strains both identified causal polymorphisms at the *RME1* locus, but not at any of the other four, combined loci that were identified (DEUTSCHBAUER and DAVIS 2005; GERKE *et al.* 2009; GERKE *et al.*

2006). That they did not identify the same polymorphism (DEUTSCHBAUER and DAVIS 2005; GERKE *et al.* 2009; GERKE *et al.* 2006) suggests that the genetic variants arose, were selected for, and fixed independently in the strains, instead of sharing a common ancestor with the polymorphisms. This, in turn, suggests that the *RME1* locus is the most likely site of large effect, beneficial mutations affecting sporulation efficiency.

The hypothesis that *RME1* is the most likely site of large effect, beneficial mutations affecting sporulation efficiency could be tested by placing a large number of replicate samples under selection for sporulation (e.g., by flow assisted cell sorting or ether treatment). Sporulation efficiency in the derived lines can be readily quantified by flow cytometry (GERKE *et al.* 2009; GERKE *et al.* 2006). Because the experiment is directed at a specific locus, genetic variation at the *RME1* locus can be determined by direct sequencing. Stored samples during the selection process could be used to identify when variants arose. If the hypothesis that the *RME1* locus is the most likely site of large effect, beneficial mutations affecting sporulation efficiency is correct, variants at the *RME1* locus would be expected, on average, to fix early in the selection process. The effect of variants on phenotype can be established by allelic replacement via homologous recombination. The same samples could be used to apply similar analyses to the other sporulation efficiency QTL that have been identified (DEUTSCHBAUER and DAVIS 2005; GERKE *et al.* 2009; GERKE *et al.* 2006). They could also be used to study the possible distribution of variants with effects on sporulation efficiency or the influence of the genetics of sporulation efficiency on correlated characters. Laboratory selections on yeast

have the capacity to be applied to specific hypotheses in evolutionary biology and general discovery.

These questions require the ability to generate both many replicate samples and genetic variation like that observed in nature. The first is a well-known advantage of microorganisms as a model system. The latter is demonstrated in this thesis work. Yet, the capacity to generate many sample lines in a yeast laboratory selection experiment has been underexploited due, perhaps, to a new focus on obtaining genome-wide sequence data, as opposed to other methods for genotypic comparison. The constraints imposed by genome-wide sequencing may also be alleviated by building on the knowledge gained in prior studies of both laboratory and natural isolates, as described above.

LYMPHOBLASTOID CELL LINES AS A MODEL SYSTEM

In this work, I have assessed the components of phenotypic variation in LCLs, in order to evaluate their use as a model system for human genetic variation. In addition, I examined the use of localized protein expression level as quantifiable alternative to transcript abundance studies that might better represent functional variation underlying phenotypes on the scale of human subjects.

Contrary to other work suggesting that LCLs are unreliable (CHOY *et al.* 2008), I found that the majority of phenotypic variation between LCLs is due to intrinsic difference between lines. The proportion of this variation that can be explained by additive genetic variation is not large ($0.03 \leq h^2 \leq 0.30$).

There are two general explanations for the observation that additive genetic variance does not explain a majority of the differences between lines, which explains

most of the phenotypic variance in the LCLs for the phenotypes assayed. The remaining line variance is composed of non-additive genetic variance (e.g., dominance and epistasis) and non-genetic variance (e.g., sampling and immortalization). While it is generally believed that additive genetic variance dominates the genetic contribution to variance, even in human traits (HILL *et al.* 2008), this observation may indicate that the genetic component of phenotypic variation between the lines may be dominated by non-additive genetic effects. Because it is not possible to define the non-genetic component of the variation between lines, the non-additive genetic hypothesis cannot be considered as anything more than a possibility. Indeed, parsimony favors the hypothesis that non-genetic effects dominate the remaining line variance.

The large, publicly available LCL collections (CEPH and the International HapMap Project) completely lack the types of samples necessary to estimate the non-genetic components of the line variance. The collections themselves can hardly be faulted for this oversight. These collections were established to investigate the distribution of DNA polymorphisms in human genomes in extant populations. The DNA sequence at a given base pair is not a continuous variable making the estimation of variance components, such as those with which I have been concerned, irrelevant. Therefore, these collections could not be expected to prioritize samples to address these issues.

As large LCL collections continue to be collected, now with an interest in using the LCLs for both phenotyping and genotyping, it would useful to comment on the types of samples that should be included in such collections in order to allow all the

components of phenotypic variance to be understood, in order to identify useful phenotypes and inform experimental design.

First, there may be variation in the immortalization process that causes LCLs to vary from each other by chance. With current collections, this variation is a component of variation between LCLs. The effect of variation in immortalization process could be defined by generating multiple LCLs by independent immortalization of aliquots from a single B cell sample from a single individual. The ability to generate these samples would be dependent on the efficiency of the immortalization process. These samples could easily be added to any new LCL collection, allowing the contribution of immortalization variation to the differences between LCLs to be defined for each phenotype of interest.

Second, there may be variation in the sampling process from subjects that causes LCLs to vary from each other by chance. Some variation between independent samples is expected by chance. With current collections, this variation is a component of variation between LCLs. Although variation between independent samples from the same individual is expected to be small, it is not defined. Quantification of variation between independent samples from a single individual would require the development of LCLs from multiple, independent B cell samples from a single individual. Including this class of sample in new LCL collections in combination with replicate immortalization samples would allow key, non-genetic components of line variance to be defined.

Third, the age of the subject at the time of the sampling may affect phenotypes measured in LCLs. The analysis in Chapter 3 shows that the distribution of phenotypic values does not change with age. Although this suggests that the LCL phenotypes of an

individual remain constant as they age, it does not rigorously confirm that conclusion. Longitudinal collection of LCLs from the same subject over time, especially across the developmental stages represented in existing collections (e.g., esterus, puberty, menopause), would make it possible to explicitly define age dependent variability in phenotypic values. Because the relationship of age to phenotype may be different, this class of samples should be included in new LCL collections in order that this relationship can be defined for the phenotypes of interest for a particular research question. While it may difficult to identify subjects who are willing to be sampled multiple times over a period of years, only a small number of individuals, representing a tiny fraction of total individuals in a collection, would be needed to describe the relationship between age and phenotype.

Finally, immortalization converts B cells with normal physiology into continuously growing LCLs . This process changes the growth phenotype. It is reasonable to assume that this process causes changes in other phenotypes. The effect of these changes could be estimated by measuring the same phenotypes in isolated B cells before immortalization and the resulting LCLs after immortalization. This knowledge would help us understand how to connect the cell culture results in LCLs to the actual physiology of B cells in the human body. Because the B cells described would be primary cell culture, not immortalized cell culture, it would not be possible to add these cells to a publicly accessible cell repository. Furthermore, the phenotypes of interest would have to be identified before collection. Interest in new phenotypes would require the collection of new samples, negating a major advantage of cell culture relative to human subjects (i.e.,

eliminating the need for new sampling for each new research question). This class of sample is both the least practical of the four listed and does not directly define components of variation between LCLs. It could reasonably be omitted from any future LCL collections.

How well LCLs represent normal B cell physiology could be addressed by comparing transcriptional profiles between LCLs and B cells across multiple samples from one individual and multiple individuals. There is evidence that transcript abundance QTL identified in LCLs replicate in primary tissue samples (BULLAUGHEY *et al.* 2009). Assaying transcript abundances using next generation sequencing (RNAseq) would allow these comparisons to be made efficiently. Efficient comparisons might be used in the construction of new collections to screen for the most representative lines amongst the sampling/immortalization replicates described above.

As LCLs have transitioned from being a renewable nucleic acid source to being a genetic model system, variables in the collection of LCLs that were not important before have become relevant as they may affect phenotypes measured in LCLs. Because current collections do not contain the samples necessary to evaluate these effects, new LCL collections would be needed to add the necessary samples.

If the goal of genetic research with LCLs is simply to identify genotype-phenotype associations, then current collections may be sufficient. If, however, the goal is to understand the genetic underpinnings of complex traits, the samples described above needed to define the non-genetic sources of variation are necessary to understand the

genetics of complex traits - the relative contributions of additive, dominance, and epistatic genetic effects to overall genetic control of phenotype.

The conclusion that existing new LCL collections are needed if LCLs are to be used a genetic model system for the basic study of complex traits in humans, highlights a more general issue regarding the utility of LCLs as a model system for human genetic variation, including practical genotype-phenotype association discovery, such as in pharmacogenomics.

It was hoped that the ability to control the cell culture environment would, by reducing the non-genetic components of phenotypic variation, make it easier to identify the genetic variants controlling the phenotypes of interest more completely. Genotype-phenotpe association studies using LCLs, however, have shown similar results, including small additive genetic effect sizes for identified QTL and low predictive power to those using human subjects. Because phenotypes measured in human subjects have the distinct advantage of being the biomedically relevant phenotypes of interest - LCL phenotypes are, at best, a proxy for these phenotypes, can situations in which LCLs are preferable to human subjects be identified, even without an improvement in genotype-phenotype association performance?

Pharmacogenomics is an obvious area of research for which cell culture is preferable, because many of the experiments, such as evaluating genetic factors that determine the response to chemotherapeutics, cannot be ethically conducted with human subjects. LCLs, however, are derived from a single cell type (B cells), which may not represent the appropriate cell type for the chemical in question. Induced pluripotent stem

(iPS) cells may be a better resource for pharmacogenomics, as they can be differentiated into the appropriate cell type for the chemical in question.

Gene expression, primarily transcript abundance (BERGEN *et al.* 2007; CHEUNG *et al.* 2003; CHEUNG and EWENS 2006; CHEUNG *et al.* 2005; CHOY *et al.* 2008; CORREA and CHEUNG 2004; DEUTSCH 2005; DUAN *et al.* 2007; DUAN *et al.* 2009; FORD *et al.* 2001; HUANG *et al.* 2007; HUANG *et al.* 2008; JEN and CHEUNG 2003; LI *et al.* 2008; LI *et al.* 2009; MONKS *et al.* 2004; MORLEY *et al.* 2004; PRICE *et al.* 2008; SMIRNOV *et al.* 2009; SPIELMAN *et al.* 2007; STRANGER *et al.* 2007; WANG *et al.* 2009; ZHANG *et al.* 2009), is also commonly studied in LCLs. Unlike pharmacogenomic studies, the fact that LCLs represent a single cell type is not necessarily a weakness. The reliability of expression levels are not influenced by the possible inclusion of multiple cell types, such as in primary tissue samples. Still, iPS cell lines may be preferable, as they would permit the study of expression differences between cell types and individuals.

Transcript abundance studies measure thousands of phenotypes. The small effect sizes and multiple hypothesis penalties are accommodated with a low false discovery rate because the number of phenotypes assayed would allow an expected, high false negative rate to be tolerated. Protein expression cannot be quantified in such a high-throughput manner. As effect sizes do not appear to be significantly increased over transcript abundance phenotypes, it is not clear that protein expression enjoys a significantly lower false negative rate than transcript abundance. Similarly, the smaller number of phenotypes assayed does not allow protein expression studies to have the same flexibility with the false discovery rate that is enjoyed by transcript abundance studies. At this time,

localized protein expression levels do not appear to be inherently superior phenotypes for developing an understanding of complex human genetic variation. The use of both transcript abundance and protein expression to study the genetic basis of phenotypic variation in humans would benefit from the addition of samples to define the non-genetic components of phenotypic variation in LCLs.

LCLs may be preferable to human subjects due to cost and labor. Genotype-phenotype association studies using human subjects require sample sizes of thousands of individuals. These samples are not reusable. The phenotypes that can be studied are limited by the study design. Addressing phenotypes outside those included in the original sample collection requires recruiting new subjects. These factors have made genotype-phenotype association studies the exclusive domain of large, collaborative research groups.

LCL collections, however, are an accessible resource for individual research groups. Although obtaining LCLs from repositories are not without cost, the expense and administrative issues are substantially less daunting than obtaining comparable numbers of human subjects. While weaknesses in current collections have been discussed previously and may require new collections to be developed, these collections will allow independent research groups to participate in the study of human genetic variation while building a community knowledge base from research on a shared set of samples.

Finally, LCLs carry some hope of testing genotype-phenotype association hypotheses by direct experimentation. These types of experiments are, in general, technically or ethically impossible, especially at the small effect sizes being discovered,

in human subjects. Already, siRNA technology has been used to confirm a genotype-phenotype association in LCLs (SHUKLA *et al.* 2009). These experiments may further benefit from future, technological developments. The use of cell culture to test genotype-phenotype associations discovered in human subjects may, again, benefit from the use of iPS cell lines, which would allow disease-relevant cell types to be tested, instead of using B cell derived LCLs as a proxy.

CONCLUSION

Laboratory selection on yeast can be used to produce genetic variation with the same characteristics as natural genetic variation: quantitative phenotypic variation controlled by a small number of loci of large effect. Many question of interest in evolutionary biology and quantitative genetics involve rare events, small differences in variables, or the interaction of multiple variables. These questions would benefit from the ability to systematically generate and phenotype samples. The demonstration that laboratory selection on yeast is relevant to natural genetic variation and the high-throughput, quantitative growth curve method developed here make this systematic approach possible.

The measurement of phenotypes in LCLs is both highly repeatable within individual lines and variable between lines, important characteristics for any model system of genetic variation. The high repeatability of phenotypes measured here (cell surface expression of protein) is expected to extend to other phenotypes and other cell types. If the advantages of cell culture for the study of human genetic variation are to be fully exploited, the non-genetic components of phenotypic variation must be well

defined. This will require new collection of cell lines, which provides the opportunity to consider other cell types, like iPS cells, as a resource.

The costs of large genotype-phenotype studies with human subjects have increasingly excluded the independent laboratory from the leading edge of human genetic variation. Cell culture methods, as a complement to human subject studies, are accessible to individual researchers and may, once again, allow the creativity and flexibility of the independent laboratory to impact the direction of human genetic variation research.

Cell culture, in both microorganisms and humans, provides an important complementary approach to the study of natural genetic variation that addresses many of the challenges inherent to this field of research.

LITERATURE CITED

BERGEN, A. W., A. BACCARELLI, T. K. MCDANIEL, K. KUHN, R. PFEIFFER *et al.*, 2007 Cis sequence effects on gene expression. BMC Genomics **8:** 296.

BULLAUGHEY, K., C. I. CHAVARRIA, G. COOP and Y. GILAD, 2009 Expression quantitative trait loci detected in cell-lines are often present in primary tissues. Hum Mol Genet.

CHEUNG, V., L. CONLIN, T. WEBER, M. ARCARO, K. JEN *et al.*, 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet **33:** 422-425.

CHEUNG, V. G., and W. J. EWENS, 2006 Heterozygous carriers of Nijmegen Breakage Syndrome have a distinct gene expression phenotype. Genome Res **16:** 973-979.

CHEUNG, V. G., R. S. SPIELMAN, K. G. EWENS, T. M. WEBER, M. MORLEY *et al.*, 2005 Mapping determinants of human gene expression by regional and genome-wide association. Nature **437:** 1365-1369.

CHOY, E., R. YELENSKY, S. BONAKDAR, R. PLENGE, R. SAXENA *et al.*, 2008 Genetic Analysis of Human Traits In Vitro: Drug Response and Gene Expression in Lymphoblastoid Cell Lines. PLoS Genet **4:** e1000287.

CORREA, C. R., and V. G. CHEUNG, 2004 Genetic variation in radiation-induced expression phenotypes. Am J Hum Genet **75:** 885-890.

DEUTSCH, S., 2005 Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. Human Molecular Genetics **14:** 3741-3749.

DEUTSCHBAUER, A. M., and R. W. DAVIS, 2005 Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat Genet **37:** 1333-1340.

DUAN, S., W. BLEIBEL, R. HUANG, S. SHUKLA, X. WU *et al.*, 2007 Mapping Genes that Contribute to Daunorubicin-Induced Cytotoxicity. Cancer Research **67:** 5425-5433.

DUAN, S., R. S. HUANG, W. ZHANG, S. MI, W. K. BLEIBEL *et al.*, 2009 Expression and alternative splicing of folate pathway genes in HapMap lymphoblastoid cell lines. Pharmacogenomics **10:** 549-563.

FORD, B. N., D. WILKINSON, E. M. THORLEIFSON and B. L. TRACY, 2001 Gene expression responses in lymphoblastoid cells after radiation exposure. Radiat Res **156:** 668-671.

GERKE, J., K. LORENZ and B. COHEN, 2009 Genetic interactions between transcription factors cause natural variation in yeast. Science **323:** 498-501.

GERKE, J. P., C. T. CHEN and B. A. COHEN, 2006 Natural isolates of Saccharomyces cerevisiae display complex genetic variation in sporulation efficiency. Genetics **174:** 985-997.

GRESHAM, D., M. M. DESAI, C. M. TUCKER, H. T. JENQ, D. A. PAI *et al.*, 2008 The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet **4:** e1000303.

GRESHAM, D., D. M. RUDERFER, S. C. PRATT, J. SCHACHERER, M. J. DUNHAM *et al.*, 2006 Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science **311:** 1932-1936.

HILL, W., M. GODDARD, P. VISSCHER and T. MACKAY, 2008 Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLoS Genetics **4:** e1000008.

HUANG, R. S., S. DUAN, W. K. BLEIBEL, E. O. KISTNER, W. ZHANG *et al.*, 2007 A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proc Natl Acad Sci U S A **104:** 9758-9763.

HUANG, R. S., S. DUAN, E. O. KISTNER, W. ZHANG, W. K. BLEIBEL *et al.*, 2008 Identification of genetic variants and gene expression relationships associated with pharmacogenes in humans. Pharmacogenet Genomics **18:** 545-549.

JEN, K. Y., and V. G. CHEUNG, 2003 Transcriptional response of lymphoblastoid cells to ionizing radiation. Genome Res **13:** 2092-2100.

LI, L., B. FRIDLEY, K. KALARI, G. JENKINS, A. BATZLER *et al.*, 2008 Gemcitabine and cytosine arabinoside cytotoxicity: association with lymphoblastoid cell expression. Cancer Res **68:** 7050-7058.

LI, L., B. L. FRIDLEY, K. KALARI, G. JENKINS, A. BATZLER *et al.*, 2009 Gemcitabine and Arabinosylcytosin Pharmacogenomics: Genome-Wide Association and Drug Response Biomarkers. PLoS One **4:** e7765.

MONKS, S. A., A. LEONARDSON, H. ZHU, P. CUNDIFF, P. PIETRUSIAK *et al.*, 2004 Genetic inheritance of gene expression in human cell lines. Am J Hum Genet **75:** 1094-1105.

MORLEY, M., C. MOLONY, T. WEBER, J. DEVLIN, K. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430:** 743-747.

PRICE, A. L., N. PATTERSON, D. C. HANCKS, S. MYERS, D. REICH *et al.*, 2008 Effects of cis and trans genetic ancestry on gene expression in African Americans. PLoS Genet **4:** e1000294.

SHUKLA, S. J., S. DUAN, X. WU, J. A. BADNER, K. KASZA *et al.*, 2009 Whole-genome approach implicates CD44 in cellular resistance to carboplatin. Hum Genomics **3:** 128-142.

SMIRNOV, D. A., M. MORLEY, E. SHIN, R. S. SPIELMAN and V. G. CHEUNG, 2009 Genetic analysis of radiation-induced changes in human gene expression. Nature **459:** 587-591.

SPIELMAN, R. S., L. A. BASTONE, J. T. BURDICK, M. MORLEY, W. J. EWENS *et al.*, 2007 Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet **39:** 226-231.

STRANGER, B. E., A. C. NICA, M. S. FORREST, A. DIMAS, C. P. BIRD *et al.*, 2007 Population genomics of human gene expression. Nat Genet **39:** 1217-1224.

WANG, L., A. L. OBERG, Y. W. ASMANN, H. SICOTTE, S. K. MCDONNELL *et al.*, 2009 Genome-wide transcriptional profiling reveals microRNA-correlated genes and biological processes in human lymphoblastoid cell lines. PLoS One **4:** e5878.

ZHANG, W., S. DUAN, W. K. BLEIBEL, S. A. WISEL, R. S. HUANG *et al.*, 2009 Identification of common genetic variants that account for transcript isoform variation between human populations. Hum Genet **125:** 81-93.

Supplemental Table 1: Lymphoblatoid Cell Line Panel (1 of 4)

Line	Family	Generation	Sex	Age (Years)	Plate	Median CD4 (High)	Median CD4 (Low)	Median CD19	Median CD23	Median CD38	Median CD40	Median CD45RA	Median CD86	Median ICAM-1	Median TLR-9
6980	13291	3	F	13	C	51.45	14.50	52.15	93.60	19.95	154.50	30.20	80.00	38.70	67.30
6981	13291	3	F	15	C	54.95	15.95	57.25	110.50	19.30	179.50	24.50	74.45	36.60	69.30
6982	1333	3	M	16	B	53.75	14.70	55.25	113.50	21.85	142.00	44.40	97.05	69.50	83.30
6983	1331	3	M	11	C	81.30	21.80	80.80	160.00	27.50	259.50	33.30	109.00	57.35	93.35
6984	13293	1	M	63	A	49.10	14.95	57.75	152.50	20.40	140.50	23.90	83.50	52.55	93.05
6985	1341	1	F	69	A	43.80	13.30	47.60	99.20	18.05	83.55	24.50	96.20	40.15	91.00
6986	13291	1	M	74	A	34.80	10.41	46.15	147.00	15.40	117.00	39.05	65.65	43.95	114.00
6987	1333	2	F	39	A	57.60	18.20	64.20	126.00	30.90	191.50	43.30	94.20	58.15	84.45
6988	1331	3	F	13	A	33.30	12.35	45.80	110.00	16.10	85.70	23.40	86.20	43.00	65.10
6989	13293	1	F	62	B	101.10	25.60	100.10	135.50	35.75	262.50	60.80	137.00	67.25	113.00
6990	1331	2	F	41	A	59.85	17.45	59.60	131.00	22.25	160.00	36.45	92.70	40.30	84.00
6991	1341	2	F	42	C	63.40	16.40	59.20	97.40	20.90	168.00	46.65	79.10	39.30	74.75
6992	1331	3	M	8	C	52.95	12.06	55.10	87.80	17.50	176.00	38.40	74.70	30.70	67.10
6993	1341	1	M	74	C	70.30	18.70	68.70	94.40	23.60	262.50	45.85	133.00	37.30	87.25
6994	1340	1	M	68	A	73.65	25.90	90.35	145.00	29.40	221.00	47.85	126.00	65.85	100.90
6995	13291	2	M	45	C	53.55	15.40	54.80	95.80	18.65	198.00	34.50	70.90	44.05	67.75
6997	13291	2	F	40	C	89.10	25.20	81.00	137.00	28.80	206.00	37.85	93.20	68.90	138.50
6998	13292	3	M	13	C	63.55	17.60	60.40	85.95	21.45	205.00	35.85	82.90	35.05	83.15
6999	1331	3	F	15	B	54.20	14.65	56.20	99.10	16.90	179.50	36.50	82.70	35.10	67.05
7000	1340	1	F	66	B	72.95	20.20	72.35	137.50	24.60	184.50	41.10	98.60	43.00	93.60
7001	13292	3	M	10	B	51.80	13.70	57.15	127.00	21.00	148.00	37.50	86.50	41.75	77.80
7002	1333	1	F	63	A	54.90	18.30	67.05	91.45	28.20	182.50	43.85	129.00	42.00	75.95
7003	13292	3	F	19	B	66.15	20.05	74.85	121.00	25.40	205.50	40.85	104.00	73.30	90.60
7004	1333	3	M	19	C	69.70	17.95	67.40	105.50	23.95	205.50	46.80	99.80	53.25	81.85
7005	1331	3	F	16	B	77.90	19.30	85.30	130.50	24.55	228.50	29.30	123.00	49.40	94.70
7006	1341	3	F	10	C	57.40	16.65	59.65	86.50	19.45	167.50	57.75	109.50	30.65	74.50
7007	1331	1	M	95	B	55.10	14.60	55.50	95.75	19.35	173.00	30.40	96.35	53.85	72.10
7008	1340	3	M	23	B	68.80	20.25	72.35	119.50	24.60	167.50	47.20	116.00	42.30	85.15
7009	1333	3	M	12	A	58.50	18.30	69.10	115.00	25.75	194.50	35.20	107.00	52.80	78.25
7010	1341	3	F	8	A	46.05	14.30	56.20	113.00	20.45	115.00	49.20	78.65	44.25	97.45
7011	1333	3	F	14	B	79.20	23.70	86.20	151.00	38.45	245.00	52.80	120.50	78.55	98.10
7012	1341	3	F	18	C	64.85	18.95	65.65	115.50	20.60	208.00	67.80	106.50	45.20	78.45
7014	13292	2	F	39	B	55.75	15.70	58.35	104.50	17.85	116.50	52.60	79.55	33.35	71.85
7016	1331	1	M	71	C	68.45	17.85	68.75	148.50	22.85	238.50	53.15	87.55	51.05	81.45
7017	1333	1	M	61	A	49.40	14.75	54.85	99.95	18.75	143.50	35.85	87.60	33.50	68.80
7018	13291	3	M	18	C	49.05	14.40	51.10	114.00	19.10	176.00	18.80	73.90	48.40	62.90
7019	1340	2	F	45	C	59.05	16.95	63.15	108.00	20.65	185.50	42.90	84.00	36.45	72.60
7020	1341	3	M	7	B	73.10	21.50	76.75	123.00	26.65	240.00	34.50	130.50	48.35	97.15
7021	1341	3	M	14	A	45.90	13.90	53.35	122.00	20.65	137.00	40.50	65.40	46.45	56.90
7022	1340	1	M	63	B	79.45	24.90	79.15	147.50	27.10	227.00	43.75	124.00	55.55	104.00
7023	1331	3	F	21	B	67.15	18.70	68.30	113.50	21.70	232.00	48.30	97.30	43.40	81.35
7025	13293	3	F	13	C	61.10	16.55	61.40	131.00	22.20	201.00	28.10	81.70	46.95	73.70
7026	1333	3	M	6	B	63.70	16.40	63.20	124.00	21.05	207.00	57.10	77.90	55.15	79.10
7027	1340	3	M	11	B	68.00	19.55	68.60	110.00	22.35	126.00	37.90	107.00	45.65	89.85
7028	13293	3	M	16	B	77.65	23.40	78.30	146.00	27.20	199.00	36.40	120.50	51.20	85.35
7029	1340	2	M	47	C	67.15	19.20	71.35	115.00	27.50	214.50	45.75	97.35	54.20	84.00
7030	1331	3	M	9	C	67.40	17.80	65.50	135.00	23.05	203.50	29.10	99.55	42.10	81.40
7031	13294	1	F	63	C	55.55	14.70	54.85	73.25	21.40	186.00	41.45	74.80	31.80	66.70
7032	13292	3	M	11	A	43.25	15.15	55.50	96.95	22.50	126.50	44.60	94.70	37.00	69.10
7033	1331	3	M	20	C	65.00	18.85	64.50	113.00	23.40	191.50	43.30	105.00	52.35	82.20
7034	1341	1	M	71	C	82.30	23.20	82.40	165.50	27.40	261.00	40.40	128.00	88.35	99.95
7036	13291	3	M	17	C	56.75	15.30	57.10	105.00	22.55	173.50	27.20	74.45	49.30	71.65
7037	13294	1	F	65	A	57.40	17.40	60.70	115.00	28.40	170.50	41.75	74.25	48.05	74.05
7038	1333	2	M	41	C	82.10	21.40	71.65	129.50	32.45	196.00	49.90	101.60	51.20	94.20
7039	13293	3	M	15	A	60.30	18.40	68.60	127.00	28.60	198.00	37.55	114.50	72.60	76.45
7040	1340	3	M	19	A	68.90	22.40	77.90	104.00	21.90	191.50	38.65	121.00	44.15	82.65
7041	13293	2	F	38	B	41.75	14.80	52.85	118.50	19.50	154.50	44.30	74.25	41.60	71.85
7042	13292	2	M	44	B	91.05	31.80	104.00	120.00	28.15	285.00	46.45	137.00	69.50	107.00
7044	1341	3	F	20	B	66.60	16.90	64.90	79.00	19.15	202.50	37.60	116.00	46.85	82.35
7045	13291	1	F	67	C	53.85	13.00	49.75	126.00	23.00	147.00	29.60	75.50	39.50	70.25
7046	13293	2	M	39	B	77.60	22.05	86.50	127.00	23.00	267.00	40.00	103.60	51.80	85.20
7047	13291	3	M	11	C	62.75	18.30	61.35	103.50	21.10	176.00	34.90	82.70	33.90	75.50
7048	1341	2	M	43	C	58.35	16.80	59.40	125.00	21.50	170.00	37.05	81.95	48.75	72.05
7049	1333	1	M	68	A	65.10	22.40	78.20	122.00	33.30	236.50	50.95	106.00	75.85	74.05
7050	1331	1	F	62	B	69.90	18.20	72.20	127.00	25.50	206.00	28.50	100.25	51.10	87.95
7051	13294	1	M	67	B	79.05	23.95	83.20	109.00	25.25	203.00	37.50	116.00	47.20	92.15
7052	1333	3	M	17	A	45.50	15.40	55.90	105.00	20.90	142.50	29.60	97.60	64.35	69.60
7053	1340	3	F	24	C	83.60	24.60	84.20	131.50	28.00	235.00	44.10	105.00	53.15	108.50
7055	1341	1	F	70	B	59.80	15.00	57.60	90.65	21.65	174.00	49.20	91.10	37.85	71.70
7056	1340	1	F	65	C	61.30	16.55	59.40	103.00	20.90	176.50	55.40	78.95	34.50	72.50
7057	1331	2	M	48	A	35.00	10.18	40.85	137.00	22.25	115.00	26.70	121.00	83.00	109.00
7058	13291	3	F	7	C	57.30	16.05	57.05	93.90	20.40	188.00	22.80	85.70	37.50	72.45

Supplemental Table 1: Lymphoblatoid Cell Line Panel (2 of 4)

Line	Family	Generation	Sex	Age (Years)	Plate	Median CD4 (High)	Median CD4 (Low)	Median CD19	Median CD23	Median CD38	Median CD40	Median CD45RA	Median CD86	Median ICAM-1	Median TLR-9
7059	1331	3	F	18	C	47.75	14.10	51.20	79.75	17.35	164.50	41.85	66.55	30.50	62.05
7060	13293	3	M	11	B	67.50	19.85	73.10	123.00	28.70	204.00	34.70	122.50	39.75	85.95
7061	13293	3	M	8	A	51.00	15.95	55.35	117.00	25.50	125.50	34.35	91.55	39.75	100.85
7062	1340	3	F	27	C	71.00	19.60	67.70	107.50	23.40	189.00	44.55	95.15	34.10	85.70
7340	1331	1	F	83	A	28.30	11.65	33.20	129.00	14.55	134.00	27.10	79.55	50.10	78.80
7341	1333	1	F	61	C	50.25	14.10	50.10	136.50	23.00	180.50	44.55	85.30	55.55	64.25
7342	1340	3	M	16	B	70.20	23.00	84.65	130.00	26.25	205.50	66.35	105.00	51.00	111.50
7343	1341	3	F	22	A	42.45	14.20	52.60	95.90	19.20	94.85	54.85	78.50	31.50	71.60
7344	1341	3	F	17	A	42.65	13.40	47.60	112.50	16.75	112.50	25.55	74.40	38.00	76.80
7345	1345	1	F	69	C	57.10	16.55	57.45	98.35	21.35	205.00	48.70	75.90	41.85	68.25
7346	1345	1	F	85	B	65.50	18.00	62.10	124.00	21.95	157.00	40.25	95.65	51.50	80.25
7347	1345	1	M	86	A	29.30	28.95	94.40	131.00	37.15	261.00	49.90	140.00	98.45	93.60
7348	1345	2	F	45	C	78.75	21.85	84.00	150.50	26.75	221.00	46.20	109.00	48.75	92.80
7349	1345	2	M	49	C	59.10	16.05	55.80	115.00	18.70	139.50	28.85	77.40	40.25	69.25
7350	1345	3	F	25	B	55.35	15.20	59.30	101.75	16.10	104.90	38.95	78.35	31.50	69.20
7351	1345	3	M	23	B	62.75	16.80	63.30	112.50	20.50	159.50	41.15	88.05	34.30	75.85
7352	1345	3	M	22	C	64.50	18.50	65.40	118.00	23.50	185.50	39.75	90.75	55.75	76.05
7353	1345	3	M	20	B	57.10	15.80	57.10	81.05	18.95	152.50	35.40	73.65	30.90	69.70
7354	1345	3	F	19	A	50.90	16.00	54.55	102.00	19.05	121.00	26.45	82.25	34.50	69.25
7355	1345	3	M	16	B	79.40	25.20	91.30	172.00	29.65	228.50	47.15	126.00	62.85	102.35
7356	1345	3	M	13	C	NA	NA	NA	NA	NA	NA	NA	NA	NA	79.10
7357	1345	1	M	69	B	71.40	20.85	79.00	117.50	24.50	159.50	34.50	128.50	40.85	86.35
7431	13292	3	M	9	C	64.10	17.35	62.65	124.50	25.20	190.50	40.00	83.75	44.10	81.15
7432	13293	3	M	6	A	46.30	14.55	55.40	94.20	21.65	91.25	42.65	86.95	47.55	87.15
7433	13291	3	M	NA	C	58.80	16.95	60.95	121.00	19.60	215.00	23.90	82.85	46.65	72.55
7435	13294	1	M	77	A	47.70	13.40	51.00	100.40	29.70	106.50	41.45	99.25	51.35	89.45
7436	13292	3	F	18	A	67.40	22.15	74.70	115.00	28.35	155.50	31.60	129.00	54.55	90.80
7437	13293	3	F	12	A	56.05	16.20	57.25	80.90	33.50	95.20	42.80	94.60	35.20	74.10
7678	1333	3	M	12	B	53.95	14.70	56.85	95.55	18.80	161.00	40.25	71.95	40.30	69.15
7679	1333	3	M	6	C	61.90	15.90	61.55	106.00	19.60	222.50	43.50	87.60	48.15	78.05
7694	13292	3	F	17	B	46.75	14.60	53.45	105.50	19.00	120.00	33.05	77.40	37.65	75.15
10832	1413	2	M	49	B	99.00	29.90	107.00	161.00	36.20	293.00	79.20	130.00	70.50	125.00
10833	1413	2	F	48	A	71.10	24.85	80.80	116.00	36.10	272.00	46.60	99.10	63.20	109.65
10834	1416	2	F	40	C	52.50	14.60	54.45	106.50	18.90	181.50	34.50	70.20	37.85	65.55
10835	1416	2	M	39	B	56.30	14.20	58.35	85.35	29.20	143.50	30.90	74.70	38.20	67.20
10838	1420	2	M	60	C	69.75	19.10	72.05	111.50	29.65	214.50	51.75	96.80	52.30	86.90
10839	1420	2	F	55	A	65.20	19.05	68.60	101.25	30.95	163.00	56.30	91.00	46.00	83.20
10842	1423	2	M	43	C	50.75	14.20	52.65	131.00	16.85	120.50	34.65	66.30	35.25	64.75
10843	1423	2	F	49	A	54.65	19.25	67.70	112.00	25.90	125.00	37.15	126.00	51.05	103.60
10844	1424	2	F	50	B	65.60	17.55	68.25	113.00	24.95	157.50	35.50	95.20	43.35	90.55
10845	1424	2	M	49	B	60.25	16.10	62.75	97.50	23.40	158.50	34.80	92.20	39.00	74.10
10848	1332	2	M	48	B	57.10	15.30	58.45	89.15	30.15	150.50	25.95	94.80	45.40	71.15
10849	1332	2	F	48	B	58.65	16.35	63.70	83.80	30.40	211.00	40.70	101.25	53.30	82.30
10850	1344	2	F	48	C	54.30	14.60	55.60	114.00	19.65	163.50	32.80	77.40	40.05	70.45
10851	1344	2	M	52	A	56.20	16.90	59.90	119.00	25.05	133.00	47.40	89.20	44.65	82.60
10852	1346	2	F	48	A	49.25	14.25	52.90	90.70	18.20	137.50	37.25	85.60	42.05	65.65
10853	1349	2	M	45	C	60.05	16.65	61.40	116.00	28.00	205.00	34.00	93.80	45.10	77.70
10854	1349	2	F	42	A	60.10	19.00	62.65	93.40	21.85	162.50	42.00	94.65	33.80	78.00
10857	1346	2	M	47	A	51.70	18.70	66.55	127.50	25.05	165.50	45.35	88.55	48.15	77.75
10858	1347	2	M	42	A	39.45	14.25	51.55	113.50	28.45	107.50	41.90	97.55	41.70	102.05
10859	1347	2	F	41	A	42.75	14.05	54.15	97.00	30.05	79.45	39.50	84.20	44.35	86.10
10860	1362	2	M	50	B	68.00	19.30	76.25	137.00	23.85	197.50	49.35	94.45	50.70	86.25
10861	1362	2	F	49	A	88.30	32.10	107.00	197.00	36.10	277.50	42.40	153.00	80.00	112.00
11833	1349	3	F	24	B	54.50	14.50	54.55	98.55	17.90	177.00	53.45	74.35	31.20	67.05
11835	1349	3	F	18	A	41.80	12.30	47.20	107.50	21.15	92.50	33.70	78.20	54.40	78.15
11836	1349	3	F	17	C	63.05	16.95	72.65	125.00	29.15	211.50	38.45	95.10	46.30	83.30
11837	1349	3	F	9	A	42.60	13.50	46.50	86.75	22.70	116.50	30.50	70.85	35.60	71.95
11838	1349	3	M	5	A	60.60	21.00	77.70	120.50	NA	152.50	45.45	93.40	41.85	85.35
11839	1349	1	M	67	B	NA	NA	151.00	NA	NA	NA	NA	NA	61.90	111.00
11840	1349	1	F	67	C	56.15	15.60	58.50	132.00	33.25	151.50	36.80	94.70	34.90	72.85
11841	1349	3	M	23	C	58.25	16.15	58.30	102.50	20.40	174.00	29.75	76.40	30.60	72.30
11842	1349	3	F	21	A	31.70	10.85	40.55	90.35	20.20	71.80	29.55	100.50	36.75	70.75
11843	1349	1	M	77	C	61.55	17.30	60.90	116.50	21.10	152.00	39.80	79.95	38.05	77.00
11870	1347	3	F	21	C	73.70	23.00	85.90	139.00	30.60	245.50	45.10	100.00	53.60	91.50
11871	1347	3	M	19	B	47.85	13.20	47.70	87.25	20.15	138.00	48.85	83.95	34.80	74.40
11872	1347	3	M	18	B	49.05	13.65	50.25	91.20	18.15	123.50	42.00	73.60	41.25	68.95
11873	1347	3	M	16	C	51.95	14.80	55.50	102.50	28.00	166.00	36.70	77.50	43.75	67.30
11874	1347	3	F	16	B	55.55	16.80	60.00	117.00	22.05	153.00	36.65	79.40	54.60	82.95
11875	1347	3	F	13	B	49.30	13.70	50.75	82.75	22.65	107.00	38.15	74.80	39.50	66.95
11876	1347	3	M	11	A	81.60	23.20	96.75	173.00	39.40	247.00	47.55	112.00	81.00	98.00
11877	1347	3	M	8	B	74.20	20.30	77.00	115.00	26.00	249.00	43.85	96.20	67.90	89.35
11878	1347	3	M	6	C	48.35	13.55	48.55	76.10	21.40	141.00	37.55	63.50	38.00	56.50

Supplemental Table 1: Lymphoblatoid Cell Line Panel (3 of 4)

Line	Family	Generation	Sex	Age (Years)	Plate	Median CD4 (High)	Median CD4 (Low)	Median CD19	Median CD23	Median CD38	Median CD40	Median CD45RA	Median CD86	Median ICAM-1	Median TLR-9
11879	1347	1	M	66	A	30.80	10.20	42.65	80.10	21.00	56.10	41.55	81.55	34.60	87.25
11881	1347	1	M	62	C	48.50	13.55	50.50	78.90	21.30	149.50	35.75	65.60	42.05	61.20
11882	1347	1	F	61	B	77.75	21.00	79.00	128.00	26.70	155.50	49.50	105.00	49.35	92.00
11909	1423	3	M	19	B	57.80	15.70	60.65	92.45	18.45	164.00	39.45	85.90	31.85	71.05
11910	1423	3	F	17	B	54.30	14.80	59.75	99.10	20.45	99.45	56.80	96.30	39.10	99.15
11911	1423	3	F	15	B	61.65	18.10	61.90	93.50	19.90	174.50	54.40	93.25	32.60	76.80
11912	1423	3	M	12	C	70.65	19.05	70.95	97.85	23.65	220.50	37.10	102.00	52.55	80.90
11913	1423	3	M	11	C	62.30	16.90	63.35	118.50	22.95	200.50	50.60	88.05	45.95	79.50
11914	1423	3	F	9	C	51.55	14.50	50.65	94.70	17.70	154.50	39.90	67.25	36.70	61.85
11915	1423	3	F	8	A	42.40	13.70	51.20	150.00	17.85	96.55	25.85	114.00	60.20	84.15
11916	1423	3	M	5	C	57.75	16.80	58.85	102.50	19.40	189.50	47.85	88.85	34.60	75.00
11917	1423	1	M	66	A	29.35	8.25	35.70	167.50	13.15	53.15	37.50	73.40	54.00	92.30
11918	1423	1	F	64	A	35.90	13.20	49.20	137.00	21.80	116.00	40.85	113.50	40.40	68.00
11919	1423	1	M	67	B	51.35	15.30	55.55	127.00	20.90	131.00	36.30	91.25	55.50	78.15
11920	1423	1	F	66	A	42.25	12.75	49.35	128.00	22.40	77.65	31.10	94.05	65.70	75.80
11921	1423	3	F	7	B	29.30	8.52	37.30	112.00	11.50	77.15	30.00	76.75	44.80	85.25
11922	1424	3	M	30	C	50.80	13.70	50.10	77.50	19.60	135.50	26.30	76.90	26.80	60.60
11923	1424	3	M	29	A	42.40	12.75	48.90	118.50	25.05	98.35	31.25	92.50	43.60	67.95
11924	1424	3	F	27	A	46.45	13.25	45.35	176.00	23.20	82.35	35.75	76.55	51.00	103.00
11925	1424	3	M	25	B	65.15	18.50	65.95	89.80	21.40	177.00	46.70	95.95	30.70	82.75
11926	1424	3	F	21	B	63.10	18.30	61.80	112.00	20.80	208.00	53.80	89.35	35.50	76.75
11927	1424	3	M	19	A	44.70	13.00	48.85	83.80	20.65	146.50	30.80	70.15	39.85	65.15
11928	1424	3	M	15	A	52.35	16.70	60.05	124.00	22.40	144.00	34.90	85.65	30.30	85.55
11929	1424	3	M	13	B	56.75	16.30	58.60	120.50	17.55	192.00	33.50	74.95	41.70	70.90
11930	1424	1	M	NA	B	57.70	16.90	62.60	134.50	23.25	161.50	32.75	85.25	56.70	77.30
11931	1424	1	F	78	C	60.45	17.10	63.80	115.50	23.55	217.00	25.70	90.40	39.65	73.90
11932	1424	1	M	76	A	52.30	14.50	59.70	126.50	20.95	120.50	29.45	74.55	50.30	98.75
11933	1424	1	F	74	C	69.90	20.00	75.25	111.00	26.45	214.00	37.05	126.00	45.55	88.85
11982	1362	3	F	28	B	68.20	18.05	70.50	113.50	23.10	196.50	45.40	87.60	36.50	77.35
11984	1362	3	M	26	A	46.85	15.30	54.80	92.20	18.80	129.00	59.70	85.90	30.10	66.00
11985	1362	3	F	24	B	58.50	16.70	64.90	114.50	19.50	190.00	29.65	105.50	27.35	81.05
11986	1362	3	F	22	B	78.80	22.80	81.00	128.50	27.65	211.50	36.95	115.50	37.95	103.00
11987	1362	3	M	19	B	67.35	20.05	72.90	167.00	26.40	185.00	48.60	93.45	64.05	83.25
11988	1362	3	F	17	B	64.15	17.40	69.15	120.00	22.85	149.50	78.20	90.15	41.10	81.45
11989	1362	3	F	14	C	64.70	18.15	60.50	111.00	21.50	197.50	45.45	78.25	28.35	78.20
11990	1362	3	M	9	B	86.45	27.60	104.00	186.00	34.20	317.50	49.90	141.50	86.70	116.50
11991	1362	3	F	7	A	42.05	13.75	55.10	167.50	18.80	82.30	44.35	74.20	72.85	81.00
11992	1362	1	M	86	C	53.15	15.05	55.50	88.70	18.60	155.50	51.60	73.30	28.90	69.45
11993	1362	1	F	80	B	62.75	17.15	64.30	99.50	20.90	143.00	34.80	101.35	36.20	78.50
11994	1362	1	M	80	A	48.60	16.20	53.30	104.15	22.40	111.00	49.35	72.05	45.55	73.80
11995	1362	1	F	84	A	42.35	15.35	54.35	78.00	19.30	128.00	56.35	80.60	35.00	68.70
11996	1362	3	M	21	B	70.15	19.40	72.45	116.00	24.50	205.00	52.85	115.50	41.40	87.35
11997	1420	3	F	31	A	52.40	20.20	70.50	123.00	25.35	127.50	47.50	97.50	40.05	76.80
11998	1420	3	F	27	B	110.00	27.25	89.50	127.00	32.10	194.00	49.45	114.00	47.75	101.70
11999	1420	3	F	28	C	98.80	28.60	96.40	138.50	42.50	261.00	64.10	148.00	78.00	103.95
12000	1420	3	F	22	A	54.70	19.50	68.30	113.00	22.50	186.00	31.10	96.65	39.70	84.40
12001	1420	3	F	20	C	54.10	15.05	54.55	100.50	17.25	164.00	37.30	71.05	29.10	66.60
12002	1420	3	F	20	B	55.10	15.50	54.15	97.10	22.00	133.00	53.70	76.15	34.65	67.30
12003	1420	1	M	97	A	68.05	19.90	74.30	122.00	29.30	190.00	48.20	110.50	52.40	84.25
12004	1420	1	F	92	C	70.35	21.50	69.50	149.50	28.55	206.00	57.55	96.95	49.85	85.95
12005	1420	1	M	77	C	51.05	14.30	51.45	88.10	16.95	137.50	30.35	65.10	30.60	60.80
12006	1420	1	F	75	B	75.75	22.50	81.90	135.00	33.25	213.00	39.90	108.50	48.00	89.90
12007	1420	3	M	28	A	48.50	15.00	56.25	109.00	29.95	148.00	42.15	74.30	39.40	73.20
12035	1346	3	M	27	A	45.30	14.10	50.60	109.50	16.45	124.50	28.25	62.65	34.25	63.90
12036	1346	3	M	25	A	60.90	18.40	66.90	117.00	23.05	156.50	36.40	99.50	38.35	81.30
12037	1346	3	M	24	B	53.65	14.90	55.30	127.50	19.45	111.30	37.70	73.10	64.70	70.65
12038	1346	3	M	21	A	46.80	15.15	53.55	129.50	19.45	88.15	38.30	73.10	48.80	82.30
12039	1346	3	M	21	C	52.75	14.50	52.10	85.90	21.65	156.00	39.30	72.40	35.85	65.60
12040	1346	3	F	16	C	59.50	16.55	58.75	115.50	19.55	166.00	36.20	71.65	29.05	73.95
12041	1346	3	F	13	A	76.80	NA	81.60	123.00	30.00	145.50	51.35	106.00	36.40	91.75
12043	1346	1	M	74	A	34.15	13.50	49.95	150.00	31.00	83.95	37.65	83.65	51.45	78.20
12044	1346	1	F	70	C	64.60	20.80	64.90	115.00	24.85	211.00	38.90	96.50	45.60	87.00
12045	1346	1	M	74	A	51.80	17.90	69.85	82.90	27.70	131.00	38.20	96.20	48.10	84.90
12046	1346	1	F	72	B	91.95	26.50	104.00	127.50	31.20	268.00	38.90	130.00	59.40	113.00
12047	1344	3	F	27	A	32.70	18.40	62.50	125.50	21.90	205.50	34.30	87.10	55.90	68.60
12048	1344	3	F	25	A	56.65	15.90	61.00	126.00	26.55	105.60	41.85	78.15	45.45	77.90
12049	1344	3	F	22	B	58.65	16.20	59.75	94.60	21.25	172.50	41.05	79.85	43.10	75.15
12051	1344	3	M	17	C	58.70	16.60	60.20	154.00	22.60	145.00	42.85	72.70	60.80	76.70
12052	1344	3	M	16	B	62.35	19.30	66.70	128.00	24.70	170.00	48.10	93.00	51.55	79.35
12053	1344	3	F	14	C	65.65	19.45	68.75	110.00	22.30	216.00	28.45	94.35	50.35	81.35
12054	1344	3	F	11	B	52.20	14.60	51.60	124.00	22.40	152.00	33.40	69.40	36.65	65.95
12055	1344	3	M	8	B	57.30	15.00	56.90	100.40	20.35	181.00	41.10	75.30	41.95	70.25

Supplemental Table 1: Lymphoblatoid Cell Line Panel (4 of 4)

Line	Family	Generation	Sex	Age (Years)	Plate	Median CD4 (High)	Median CD4 (Low)	Median CD19	Median CD23	Median CD38	Median CD40	Median CD45RA	Median CD86	Median ICAM-1	Median TLR-9
12056	1344	1	M	79	C	67.80	19.60	70.70	150.00	24.60	223.50	50.65	87.50	41.70	85.75
12057	1344	1	F	75	B	56.15	16.25	56.85	101.50	19.35	135.00	40.40	77.30	37.30	77.45
12058	1344	1	F	81	B	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12089	1332	3	F	25	B	62.80	18.15	66.85	97.20	28.50	203.00	32.75	134.00	62.35	77.50
12091	1332	3	M	15	A	69.75	27.50	88.70	NA	41.65	122.00	42.00	115.00	45.10	76.80
12092	1332	3	F	14	C	66.90	20.05	69.80	107.00	25.95	256.50	36.20	96.00	51.50	85.70
12093	1332	3	F	13	B	64.15	16.15	63.40	97.15	24.50	156.00	67.50	87.50	46.10	75.00
12095	1332	3	M	8	A	NA	NA	NA	NA	NA	NA	NA	NA	NA	61.00
12096	1332	1	M	94	A	51.05	14.80	54.90	123.50	24.60	113.50	58.05	80.40	59.70	85.65
12097	1332	1	F	85	C	52.95	21.50	72.10	105.00	27.60	310.50	35.70	98.95	44.60	61.90
12099	1332	1	M	73	B	76.30	20.00	78.20	138.50	33.05	206.00	41.40	122.50	75.30	89.35
12101	1413	3	M	28	A	51.20	18.00	63.30	90.30	27.30	238.00	42.80	112.00	45.05	79.15
12104	1413	3	F	23	C	NA	30.10	64.10	NA	26.60	254.00	42.95	NA	NA	89.10
12106	1413	3	M	21	A	75.00	20.90	86.30	139.00	31.45	265.50	45.90	117.50	76.45	100.40
12107	1413	3	M	19	C	71.85	19.40	73.30	116.50	28.90	220.50	39.60	91.50	52.70	92.30
12108	1413	3	M	18	C	86.30	23.45	83.95	176.50	33.70	199.50	46.60	128.00	52.70	99.70
12109	1413	3	M	17	C	60.40	16.55	62.30	118.50	21.85	205.50	30.15	86.55	43.40	73.80
12110	1413	3	F	15	A	49.25	14.95	52.40	89.90	17.90	137.50	31.90	85.10	35.25	68.15
12111	1413	3	M	13	A	NA	NA	NA	NA	NA	NA	NA	NA	NA	83.40
12112	1413	3	M	12	C	64.40	20.55	66.85	132.50	24.65	225.00	35.90	97.15	43.60	85.40
12113	1413	3	M	11	A	53.55	19.10	72.00	127.00	24.90	221.00	42.40	92.90	47.15	82.40
12114	1413	3	F	10	A	50.15	15.80	57.50	93.60	27.15	164.00	42.95	91.20	43.05	96.55
12115	1413	3	M	6	A	53.40	15.10	56.75	95.65	20.45	111.00	30.50	81.60	43.20	85.10
12116	1413	1	F	68	C	61.00	15.60	66.05	103.75	18.50	150.00	35.70	94.25	48.05	71.80
12117	1413	1	F	72	B	61.30	17.70	63.50	102.00	28.40	194.50	49.50	96.05	49.20	86.20
12237	1332	3	F	9	B	79.95	22.70	85.00	110.50	30.10	221.00	47.05	110.50	57.20	96.80
12240	1416	3	M	19	B	52.90	14.95	55.55	107.00	18.40	143.50	53.25	78.45	44.95	70.00
12241	1416	1	M	16	A	NA	26.20	92.75	NA	30.00	205.00	50.75	91.90	36.40	91.90
12243	1416	3	M	13	A	47.50	16.10	53.55	78.05	22.40	155.50	38.90	88.40	37.30	65.60
12244	1416	3	F	12	C	51.40	14.40	53.00	119.00	18.30	164.00	31.95	71.50	49.00	62.65
12245	1416	3	M	10	B	54.55	14.65	57.65	125.00	18.35	153.50	36.75	87.75	50.55	69.95
12246	1416	3	F	8	B	53.30	14.55	55.15	116.00	20.30	159.50	43.25	71.85	33.15	67.35
12247	1416	3	F	5	A	49.50	15.15	56.65	101.50	18.75	144.00	26.25	68.95	37.60	70.90
12249	1416	1	M	77	C	53.20	14.35	55.55	104.00	20.45	177.00	45.35	88.20	44.95	69.05
12250	1416	1	M	66	C	103.00	27.75	104.50	147.00	35.40	214.00	55.90	128.00	55.10	94.10
12251	1416	3	F	63	C	54.20	14.75	54.95	81.60	23.25	176.50	47.30	83.65	41.15	66.90
12252	1416	1	F	5	A	41.10	11.90	50.65	121.00	22.90	81.05	32.00	85.00	43.00	119.00
12253	1416	3	F	6	C	52.75	14.60	53.15	96.20	18.45	170.00	30.85	71.85	40.65	67.00
13130	1332	3	F	NA	A	59.60	21.25	66.20	117.00	33.85	159.00	37.05	106.00	63.00	92.65

Supplemental Table 2: Pilot Lymphoblatoid Cell Line Panel

Line	Family	Generation	Sex	Age (Years)	Mean CD40	Mean CD86	Line	Family	Generation	Sex	Age (Years)	Mean CD40	Mean CD86
6985	1341	1	F	69	231.50	161.60	10850	1344	2	F	48	248.00	127.83
6991	1341	2	F	42	230.17	120.67	10851	1344	2	M	52	279.50	133.00
6993	1341	1	M	74	235.40	175.83	10860	1362	2	M	50	249.80	125.00
6994	1340	1	M	68	388.17	186.40	10861	1362	2	F	49	361.00	160.75
7000	1340	1	F	66	285.60	122.75	11982	1362	3	F	28	253.33	129.50
7006	1341	3	F	10	191.67	151.80	11984	1362	3	M	26	176.00	128.25
7008	1340	3	M	23	370.25	182.83	11985	1362	3	F	24	262.00	154.50
7010	1341	3	F	8	214.67	102.83	11986	1362	3	F	22	258.50	142.25
7012	1341	3	F	18	225.00	155.50	11987	1362	3	M	19	318.17	130.75
7019	1340	2	F	45	285.67	125.60	11988	1362	3	F	17	278.20	143.17
7020	1341	3	M	7	351.83	158.83	11989	1362	3	F	14	263.00	102.00
7021	1341	3	M	14	222.40	101.30	11990	1362	3	M	9	361.33	161.25
7022	1340	1	M	63	375.00	151.50	11991	1362	3	F	7	206.67	129.00
7027	1340	3	M	11	214.67	151.20	11992	1362	1	M	86	180.75	105.50
7029	1340	2	M	47	456.17	160.83	11993	1362	1	F	80	220.60	161.75
7034	1341	1	M	71	256.33	141.20	11994	1362	1	M	80	254.67	94.78
7040	1340	3	M	19	338.75	244.50	11995	1362	1	F	84	214.33	110.87
7044	1341	3	F	20	328.00	192.20	11996	1362	3	M	21	268.50	140.25
7048	1341	2	M	43	227.50	118.74	11997	1420	3	F	31	250.20	134.40
7053	1340	3	F	24	256.50	116.60	11998	1420	3	F	27	197.60	132.33
7055	1341	1	F	70	245.00	154.80	11999	1420	3	F	28	340.75	205.67
7056	1340	1	F	65	214.60	107.25	12000	1420	3	F	22	212.50	160.00
7062	1340	3	F	27	195.40	99.08	12001	1420	3	F	20	234.40	114.83
7342	1340	3	M	16	303.50	91.90	12002	1420	3	F	20	192.20	125.20
7343	1341	3	F	22	204.67	121.80	12003	1420	1	M	97	327.40	138.20
7344	1341	3	F	17	198.00	124.00	12004	1420	1	F	92	246.80	139.83
7345	1345	1	F	69	271.75	122.17	12005	1420	1	M	77	230.20	117.60
7346	1345	1	F	85	329.50	150.25	12006	1420	1	F	75	273.67	124.00
7347	1345	1	M	86	359.20	163.33	12007	1420	3	M	28	238.50	131.50
7348	1345	2	F	45	321.60	160.80	12047	1344	3	F	27	326.50	143.50
7349	1345	2	M	49	190.25	110.80	12048	1344	3	F	25	276.25	124.00
7350	1345	3	F	25	188.25	133.00	12049	1344	3	F	22	229.00	137.67
7351	1345	3	M	23	237.60	131.50	12051	1344	3	M	17	249.25	126.00
7352	1345	3	M	22	223.17	159.25	12052	1344	3	M	16	216.25	108.35
7353	1345	3	M	20	218.25	119.75	12053	1344	3	F	14	255.75	131.50
7354	1345	3	F	19	201.00	142.25	12054	1344	3	F	11	252.00	102.43
7355	1345	3	M	16	380.33	162.67	12055	1344	3	M	8	310.17	134.17
7357	1345	1	M	69	291.50	240.83	12056	1344	1	M	79	272.50	123.80
10838	1420	2	M	60	324.00	135.25	12057	1344	1	F	75	191.50	120.17
10839	1420	2	F	55	326.00	115.25							