Published in final edited form as:

Curr Mol Med. 2014 ; 14(7): 833–840.

Cell Lines Models of Drug Response: Successes and Lessons from this Pharmacogenomic Model

J. Jack^{1,2,§}, D. Rotroff^{1,2,§}, and A. Motsinger-Reif^{*,1,2}

¹Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA

²Department of Statistics, North Carolina State University, Raleigh, NC, USA

Abstract

A new standard for medicine is emerging that aims to improve individual drug responses through studying associations with genetic variations. This field, pharmacogenomics, is undergoing a rapid expansion due to a variety of technological advancements that are enabling higher throughput with reductions in cost. Here we review the advantages, limitations, and opportunities for using lymphoblastoid cell lines (LCL) as a model system for human pharmacogenomic studies.

There are a wide range of publicly available resources with genome-wide data available for LCLs from both related and unrelated populations, removing the cost of genotyping the data for drug response studies. Furthermore, in contrast to human clinical trials or *in vivo* model systems, with high-throughput *in vitro* screening technologies, pharmacogenomics studies can easily be scaled to accommodate large sample sizes.

An important component to leveraging genome-wide data in LCL models is association mapping. Several methods are discussed herein, and include multivariate concentration response modeling, issues with multiple testing, and successful examples of the ‘triangle model’ to identify candidate variants. Once candidate gene variants have been determined, their biological roles can be elucidated using pathway analyses and functionally confirmed using siRNA knockdown experiments.

The wealth of genomics data being produced using related and unrelated populations is creating many exciting opportunities leading to new insights into the genetic contribution and heritability of drug response.

Keywords

Cytotoxics; genetic association mapping; genome wide association analysis; lymphoblastoid cell lines; pharmacogenomics

Introduction

The identification and characterization of genetic variant associations with pharmacologic phenotypes are the ambitious goals of pharmacogenomics. The ultimate goal of such genetic association studies is clinical translation – building predictive models based on an individual's genotype(s) that maximize treatment effectiveness or reduce adverse reactions. While *in vitro* human experiments are ideal for such studies, this is not always practically possible, especially for highly toxic drugs. One important model system that has emerged in pharmacogenomics is the use of Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines (LCLs) to measure and model drug response [1-3]. There have been many successes using this model for drug discovery, functional validation, and translational research.

As the number of studies using this model has grown, there are a number of lessons learned that have contributed to understanding the genetic etiology of the drug response phenotypes measured. Example successes with the LCL model experiments have been reviewed in detail [1, 3], and will be briefly discussed in the current paper. While highlighting key successes and limitations, this review will uniquely focus on the statistical methodologies used for association mapping with the cell line models, with an emphasis on insights across studies that may motivate new analysis approaches in future studies. Advantages and disadvantages are explained, along with a brief survey of successes. We review the statistical approaches used with a discussion of what the results have revealed about the etiology of the dose response traits. Finally, future directions are proposed.

Advantages and Disadvantages of the LCL Model

Advantages

There are a number of advantages that have motivated the expanded use of the LCL model. First, by leveraging a number of established cell line resources the model system is extremely cost effective. Examples of these resources include the Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees [4], the International HapMap Project [5], and the Human Variation Panel Populations [6] which provide established cell lines with extensive publicly-available genetic data. These resources provide genome-wide single nucleotide polymorphism (SNP) data, baseline (sans drug exposure) gene-expression data, and increasing amounts of next-generation sequencing data for the cell lines [6]. The data available from these resources allows for genetic and genomic mapping for only the cost of the drug response phenotyping. Welsh and colleagues [1] reviewed some of the most commonly used resources, including availability. The 1000 Genomes Project [6] and The Cancer Genome Atlas [7] are two of the largest resources that have emerged since the 2009 review.

Second, compared to pharmacogenomics mapping in clinical trials, there are relatively few confounding issues in the *in vitro* model system. Often clinical trials are not designed for genetic analysis, and confounding issues related to study design, complex treatment regimes, etc. limit the potential for clinical trial studies to dissect the genetic etiology of single drugs. While there are potential experimental confounders such as growth rate of the cells and

batch effects [8], careful technical execution and statistical analysis can readily control for such concerns [9].

Third, as compared to more traditional pharmacogenomics studies relying on clinical trials, there are fewer limitations to the potential number of samples and study designs that can be evaluated with the *in vitro* system. Unlike clinical trials that have a limited number of samples (typically in the hundreds at most) with mostly unrelated individuals, the LCL model enables the interrogation of as many cell lines as experimentally and economically feasible, and enables family-based designs for evaluating heritability and performing association mapping [1, 3].

Fourth, the LCL model is amenable to robotic automation, allowing for efficient screening of large numbers of drugs and/or chemicals. While the first papers using the model were restricted to one or two drugs included per experiment, the scale of phenotyping is rapidly changing. A recent application in the toxicogenetics field evaluated nearly 200 chemicals with nearly 1000 cell lines [10], showing the type of throughput that is now possible with such assays.

Fifth, there are a number of options for immediate follow up of associations in the LCL model. While replication studies in clinical trial pharmacogenomics are often impractical or impossible, functional testing of gene associations is a common step in an LCL experimental plan. For example, knock-down experiments can be used to follow up the initial associations found in the mapping analyses [11-13]. Other studies use the results of the mapping experiments in LCL to inform candidate gene analysis in clinical outcomes, as reviewed in detail in Wheeler *et al.* [3].

Disadvantages

While the advantages are clear, like all model systems there are limitations that must be considered. First, and arguably the biggest limitation of *in vitro* models is the difficulty in translating the relevance of *in vitro* phenotypes to *in vivo* outcomes. This is especially pertinent for highly toxic drugs, such as chemotherapies, where the therapeutic index is very narrow. Whether sensitivity *in vitro* relates to efficacy, toxicity, or both is an open, unanswered question. However, such issues with *in vivo* extrapolation are true of any model system [1]. Additionally, there may be inherent differences in the genomic characteristics of the cell lines and either germ line or somatic mutations of cancer types that may need to be kept in mind in the selection of appropriate cell lines for such experiments [14].

Second, as mentioned above, there are various experimental and technical confounders that have been discovered. Such issues are discussed in detail in Choy *et al.* [8]. In particular, growth rate has emerged as a consistent confounder across a number of studies [8, 15, 16].

Third, there are a limited number of tissues that are amenable to the EBV transformation process [1]. Given the issue of *in vitro* to *in vivo* extrapolation mentioned as limitation one, investigators would ideally like to perform the cell line experiments in lines most relevant to the tissues types involved in the pharmacologic response of interest. For example, hepatocytes are often the most relevant tissue type to identify variant gene function in drug

metabolizing enzymes; however, such tissue is extremely difficult and expensive to obtain. The cell line model experiments are most commonly performed in lymphoblastoid cells, with cell lines from diseased tissue (particularly cancer cell lines) also used in a number of studies. As stem cell technology advances, it is likely that the number of tissue types that are available for such assays will increase dramatically.

Fourth, not all cell lines express all pathways of pharmacological importance, which again limits the ability to extrapolate the relevance of some of the findings *in vivo*. It is important to consider the limitations of specific cell lines in context of the drug(s) to be evaluated. For example, like many immortalized cell lines, LCLs do not readily express the cytochrome P450 (CYP450) family of genes, which are metabolic enzymes that are important for activating many drugs [17]. Genetic associations for variants in these genes and drug metabolism cannot be interrogated in LCLs, but could be interrogated with some hepatic cell lines. Also, drugs that interact only in very specialized cells like neurons or pancreatic beta cells may not be readily assayed *in vitro*. However, a large number of drugs, like statins and cytotoxic agents, act through pathways that are universally represented in all cell types, making them readily amenable to such assays. For example, while LCLs don't express the CYP450 family of genes, they do express approximately half of the known genome [17]. Many of the concerns and considerations in actually developing an appropriate cell line system for a particular pharmacologically relevant phenotype are reviewed in Welsh *et al.* [1].

Methods Used

Laboratory Methods

Just as the choice of cell line used must be appropriate for the drug being studied, a range of options exist for assays measuring pharmacologically relevant outcomes in the cell line system. Careful selection of endophenotypes is a crucially important component for the successful use of the cell line model system. The mechanism of drug action is typically the driving factor in selecting endophenotype(s) for evaluation. For example, measures of cellular apoptosis across a range of drug doses can be an appropriate phenotype for cancer drugs that induce cell death or growth restriction. Measuring the conversion of a parent drug to its metabolite is another example that has been implemented successfully. The range of endophenotypes used is extensively reviewed in Welsh *et al.* [1].

While there are a wide range of assays available to measure endophenotypes, potential experimental artifacts and confounders must be considered and controlled whenever possible. For example, growth rate is consistently shown to be a confounding issue in cytotoxicity assays [8]. If confounders are present, careful statistical analysis must be performed to correct for potential issues with confounding [9].

One clear trend that can be observed with these experiments is the increasing scale of the number of endophenotypes evaluated. Early studies focused on assays using one or two drugs; however, as the scale of genotyping has increased, the use of robotics has increased the number of drugs and individual cell lines that are readily assayed to dozens [18] or even hundreds of drugs/chemicals and hundreds of cell lines [10]. As discussed below, this

increase in scale motivates very careful quality control and more complex statistical analysis.

Cell Line Resources

One of the key efficiencies of the cell line model approach is the availability of both cell lines and genetic data which were used in previous studies. For example, Coriell has immortalized LCLs available for purchase from a diverse set of individuals that have been used in the field of genetics for a variety of studies (<http://ccr.coriell.org>). Importantly, these cell lines have been genotyped at a genome-wide scale through projects such as HapMap [5], and 1000 Genomes [6]. These samples can be purchased and used for any drug response phenotyping assay, and the genetic data is freely available for academic use.

While some studies use private cohorts for the cell line assays, there are a few major resources that are often used. As the HapMap database is widely used. HapMap provides baseline gene expression data for the HapMap samples from a range of platforms through www.ncbi.nlm.nih.gov/geo/. Genetic variation data is available for the HapMap samples (with over ten million SNPs) through www.hapmap.org. For cancer cell lines, the NCI60 database has a number of resources available. SNP, expression, and copy number variation data on cell lines from different tissues is available through dtp.nci.nih.gov.

These resources include cell lines from families, as well as unrelated individuals from global populations. Association mapping can be performed in both family-based studies, or in unrelated samples (though blending the two study designs can be a concern). As always, there are a number of advantages and disadvantages to the two different study designs, as reviewed in Evangelou *et al.* [19].

Family based samples can be used to estimate the overall heritability of drug response [15]. They are robust to certain types of confounding which is commonly a concern in population-based (unrelated) studies. Additionally, family based samples can be used to identify certain types of genetic association, such as maternity and imprinting effects that cannot be identified in population-based studies. The primary disadvantage of the family based studies, as compared to population-based, is the level of resolution of the genetic mapping. With a larger number of meiotic recombination occurring in a population-based sample as compared to a similarly sized family-based study, the linkage disequilibrium patterns in a population based sample allow for more narrow regions of associations to be identified. Also, larger numbers of individuals are typically available using a population-based study, resulting in higher sample sizes.

An important consideration in the selection of cell lines for population-based studies of drug response is the ethnic background of the individuals used to generate the cell lines. There are many potential issues with confounding due to population heterogeneity and admixture that must be considered and properly handled in the cell line model experiments, just as in any association analysis with non-family based studies [15]. This potential confounding is called population stratification, and refers to the concerns due to systematic differences in allele frequencies between subpopulations in a population, possibly due to different ancestry. When allele frequency differences are combined with group level differences in dose

response, spurious/false positive findings become a concern. It has been shown repeatedly that a broad range of drug response phenotypes exhibit ethnic differences in drug sensitivity [20, 21]. There are several approaches for taking heterogeneity into account, as discussed in the review of statistical methods typically used.

Choice of Genotyping Approach

As with any genetic mapping study, numerous options are available for assaying genetic variation in cell lines studies [22]. Any number of variants can theoretically be evaluated – from testing one specific single nucleotide polymorphism (SNP), to evaluating millions of SNPs across the entire genome, depending on the scope of the genetic hypotheses being tested. There are many examples of each extreme, with genome wide approaches becoming increasingly popular [1, 2]. The majority of studies have relied on SNP data, but with newer technologies becoming available, other types of DNA-level variation are being interrogated for association with drug response. Copy number variants, epigenetic variation, and rare variants discovered through next generation sequencing technologies can all be evaluated in the cell line model. In addition to DNA level variation, gene expression data is also commonly evaluated in the cell line assays [1, 2]. Baseline (non-drug exposed) microarray assays results are available for many of the cell lines, and these assay results are often used to interpret and/or reinforce the DNA level variation association results.

Statistical Analysis Approaches

While the details of the best statistical approaches used vary according to the particular goals of the study, there are a number of trends that have emerged across studies. Typical approaches used are reviewed here.

Quality Control

As the Choy *et al.* [8] study highlighted, potential issues with experimental artifacts need to be evaluated and assessed using careful quality control (QC) procedures of the laboratory data. Biological and technical replication should be an important component of the experimental design, and replicate samples should be used to evaluate the reproducibility and reliability of the experimental data [9]. Other metrics often used for QC include outlier detection and characteristics of the dose response relationship (e.g., monotonicity), etc. [9].

QC measures must also be interrogated for the genotype data, as well as the phenotype data. Details of the QC depend on the genotyping platform and should follow best practices for the particular platform. Some common QC filters applied to genotyping data include percent missing data by SNP, percent missing data by individual, violation of Hardy-Weinberg equilibrium, and minor allele frequency. Details of best practices for genetic quality control can be found in Motsinger-Reif *et al.* [9].

Association Analysis

The main goal of any of the LCL gene mapping experiments is to correlate/associate genotype with phenotype. There are a number of methodologies currently used, and selecting the appropriate method depends heavily on the study design and underlying genotypic and phenotypic assumptions. For example, incorporating batch effects and

population stratification, summarizing response profiles into univariate measures (e.g., curve fit parameters), and making additive vs dominant genotypic assumptions may have considerable effects on the association results [9]. It is important to consider these modeling assumptions early in the study design process to maximize confidence and improve accuracy of the statistical associations.

The interpretation of the assay result and processing of the phenotypic data may have a dramatic effect on the association results. While the variation of the concentration response drives the association results, phenotypes are often complex and “different” responses are not always well defined. Most commonly, the concentration response profile data is fit to a non-linear curve, and a parameter from this curve is used to summarize the entire response profile—e.g., the half-inhibitory concentration, or IC50 [9]. Selecting a single curve parameter as a univariate anchor reduces the complexity of the phenotypic response in a way that may undermine the association results [23-25]. For example, how do we know that a nonlinear (hill) function is the optimal model for a particular set of concentration response data? Furthermore, does the IC50 (or another univariate curve fit parameter like a minimum or Hill slope) sufficiently capture the variation between individuals? Do these assumptions hold true across genotype categories for a single drug? Do these assumptions hold true for all drugs or chemicals in a high throughput experiment?

Recent simulation studies have shown that modeling the full concentration response profile in a true multivariate analysis approach can be more powerful than methods involving summary measures [23]. These multivariate models have shown promising results with experimental data as well [18].

Model covariates can be used to correct for additional confounding elements in the experimental design – in both the phenotype collection and the genotypes used for association analysis. Additional experimental information collected along with the phenotype (temperature, growth rate, date, and time) may be used to correct batch effects and separate the phenotypic signal from the noise. The genotypic data can be analyzed for population stratification, typically *via* principal component analysis, which can be used to correct for admixture and heterogeneity that can drive false positive associations. Such genotype confounding issues are not unique to the cell line model, and they are reviewed in [20]. The relative importance of these cofactors should be evaluated in variable selection modeling approaches to determine if they should be used as covariates in association analysis.

Typically, variant-by-variant analysis is performed to test for associations. The majority of studies test for the association of a single SNP with differential dose response. Multiple linear regression and Multivariate analysis of covariance (MANCOVA) are the two most commonly used approaches in studies with unrelated/non-familial study designs [2, 3, 18]. As previously mentioned, covariates such as batch effects, growth rate, and summary measures of population ancestry are included in a model that also incorporates a term or terms for the genotype categories for a single SNP. The specifics of how the SNP (or other variant type) is entered into this association analysis is an important analytical choice which reflects biological assumptions.

There are generally two approaches for genetic data encoding: allelic and genotypic tests/encoding. Using an allelic encoding (such as 0,1,2 for a biallelic SNP) makes an additive genetic model assumption, where an investigator is assuming that the mean dose response (however quantified) changes consistently (on a linear, log, log2, etc. scale) with the number of minor (less frequent) alleles. There are a couple of different genotypic encoding approaches that are typically used [26]. The first is to make a dominance assumption, by treating the heterozygote and one of the homozygote genotypes as a single category. This dichotomization of the SNP genotypes forces heterozygotes to have the same risk or mean phenotype as one of the homozygotes. A more general genetic modeling approach simply uses the maximum number of distinct genotype classes and avoids dominance assumptions using dummy variable encoding. The statistical consequences of these choices depend on the true biological model. If the biological assumptions tested are valid, methods with a dominance assumption have higher power to detect genotype-phenotype associations, but if they are incorrect will limit power. A more detailed discussion of these concerns can be found in Lunetta *et al.* [26]. Typically, whatever set of covariates and encoding choices are made is repeated for the total number of variants tested. This is often the case for genome-wide association data, such that millions of these single variant tests are performed. Because the true biological model varies across the genome, researchers typically apply allelic and or dummy encoded assumptions consistently across all variables.

Due to the extremely large number of comparisons made in a typical GWAS, it is important to reduce the probability of encountering type I errors (false positives) or type II errors (false negatives). This is known as a correction for multiple comparisons or multiple testing. The standard for evidence of significance in GWAS involving one million SNPs is generally considered to be $p < 5 \times 10^{-8}$ [27], which comes from assuming a 5% chance that a positive association occurs randomly, divided by the total number of effective associations tested (the effective number of variants in the genome, accounting for the correlation between them). Typically in such studies, p-values less than 5×10^{-8} are considered to be significant. This is based on a Bonferroni correction for the effective number of tests performed in a typical GWAS. This is considered a reasonable approach to multiple testing since it aims to reduce the number of false positives but at a cost of potentially increasing the number of false negatives. In fact, lower powered studies (smaller sample sizes) can have inflated type II error rates under Bonferroni correction. Alternatively, depending on the context for the association analysis, it may be more desirable to focus on reducing the number of false negatives: in this case, the False Discovery Rate method can be applied [27]. Finally, while traditional permutation testing is essentially a computationally intractable solution for multiple testing when dealing with millions of SNPs, there have been efforts to adapt the method for GWAS data [28].

A number of studies try to leverage information from multiple data types to inform analysis – often call the “Triangle model/approach” in these studies. The Triangle model is an example of an approach commonly used to confirm findings in genome wide analysis studies [1]. The first leg of this model identifies which genomic markers (e.g. SNPs, methylation sites, etc.) are significantly associated with the phenotypic outcome being investigated. Next, the genomic markers are linked to transcript expression (i.e. eQTL

analysis). Lastly the target genes identified in the second leg are analyzed for their association with the phenotypic target being investigated. An advantage of this approach is that true associations between genomic markers and phenotypic outcomes are likely to be eQTLs. Furthermore, by tagging genomic markers to eQTLs, information from multiple platforms can be utilized and compared, and pseudo-replication across data types can be leveraged [29]. However, caution must be used when implementing this model because it does not take into account gene-environment interactions, which may play a significant role in the phenotypic response. In addition, the technological platform being used may not represent the appropriate mechanism *via* which an eQTL impacts a phenotypic outcome. Despite these limitations, the triangle model has been used successfully to identify novel genetic variants capable of predicting chemotherapeutic sensitivity for etoposide [30], daunorubicin [31], cisplatin [32], carboplatin [33], cytarabine arabinoside [34].

Validation

Because of the large numbers of targets tested in genome wide studies, and limitations of the model system itself, the issue of false positives is non-trivial. Once the triangle model (or other method) has been used to identify putative genetic variants, these variants must undergo additional testing to confirm their clinical relevancy. A commonly implemented method for validating findings from genome wide studies is to use gene knock down experiments using small interfering RNA (siRNA). By silencing the genetic variant of interest one can test the functional effect of that target. This technology has been explored extensively for functional follow-up studies, but is also being actively investigated for therapeutic treatments and is reviewed by Shan [35]. Although, not as common, forced gene expression can also be used to validate the role of a specific genetic target to a phenotype of interest by increasing expression of a target gene. Both knockdown and forced gene expression experiments can be performed in the *in vitro* LCL model or using *in vivo* mammalian model systems (e.g. murine models).

Successes of the LCL Model

Despite the limitations discussed, there are a number of empirical successes that support the use of the LCL model system. The model system has been used for discovery of new genetic associations and the functional testing of previously proposed associations. Successes before 2009 are thoroughly reviewed in Welsh *et al.* [1], and more recent successes are reviewed in Cox *et al.* [2]. A few of these examples are highlighted below.

Li *et al.* demonstrated how the triangle model can be used to successfully identify target genes and siRNA knockdowns can subsequently be used to validate that the target genes do alter LCL sensitivity to chemotherapeutics, Gemcitabine and AraC [36]. This study identified SNPs in IQGAP2 (rs3797418) and TGM3 (rs6082527) that were associated with variation in both Gemcitabine and AraC potency. The siRNA knockdown of these two genes altered the sensitivity to both chemotherapeutics, confirming the role of these two genes at least in the *in vitro* system [36].

Other studies have used the triangle model to identify SNPs associated with sensitivity to platinating chemotherapeutics [37]. A study by O'Donnell *et al.* provided insight into the

increased sensitivity to these agents seen clinically in Asian populations [38]. Furthermore, platinating therapies, such as carboplatin and cisplatin, are commonly used to treat testicular cancer, endometrial cancer, bladder cancer, as well as many others, illustrating the potential utility of this cell line for diseases other than hematologic malignancies.

Additional studies with promising translational and clinical impacts have been performed using LCLs in other areas including, but not limited to, radiation sensitivity [39], statin sensitivity [40, 41], selective serotonin reuptake inhibitor sensitivity [42], and responsiveness to immunosuppression therapy [42].

Lessons Learned

After about a decade of experience using the cell line models for pharmacogenomics, several important lessons have emerged that must be considered as the field of pharmacogenomics more frequently applies this model system.

First, it is clear that the genetic etiology of drug response *in vitro* is similar to the etiology seen in common, complex trait analysis. The *in vitro* drug response traits have a broad range of heritabilities, similar to *in vivo* disease traits [15]. This is an important insight, since many *in vivo* pharmacogenomics studies do not have the family-based data needed and are underpowered for the estimation of heritability of the drug response phenotypes. The cell line model data supports that many drug response phenotypes are indeed heritable, and are further supported by the clinical validation of many of these cell line findings.

Also, like in disease mapping, it is clear that heritability does not guarantee “easy” gene mapping. While there are a number of clear successes searching for single variants that are associated with response in the cell line system, even the largest “hits” explain very little variation in drug response [15, 18]. This is very similar to the results being seen in disease and complex trait genetics [43]. While in many ways this is not surprising, there has been a tradition of assuming that drug response traits may have more simple genetic etiologies, as initial successes with single variant associations in drug metabolizing enzymes may have provided an overestimate of the expected effect sizes of associations in pharmacogenomics [15, 18].

These two observations indicate the same problems with “missing heritability” that is receiving so much attention in human genetic mapping in general. Results from the cell line model mapping experiments suggest that while there may be a few variants that explain a significant amount of variation in drug response traits [44], overall they do not adequately address the gap between the estimated heritability of a drug variation and the actual percent of drug response explained from the measured genetic variation. The potential sources and explanations of this missing heritability are likely the same as those discussed for any other complex traits – rare variants with strong penetrance, gene-gene interactions, a lack of complete genomic coverage in current assays, limited power based on sample size, etc. Additionally, the use of summary measures for phenotypes *in vitro*, e.g., IC50, and the noise and uncertainty in the model fitting process may not properly capture the variation across individuals, which could lead to problems in dissecting the etiology of traits with such high noise.

Opportunities

Many of the new opportunities in the cell line system mirror those in human genetics in general.

Cutting edge technologies, such as next-generation sequencing (NGS), are capable of characterizing multiple types of genetic variations such as SNPs and copy-number variations. Other technologies explore epigenetic variations (i.e. methylation patterns) that may alter regulation of certain genes. As additional advancements in technology are realized, genetic information will be obtained for larger study sizes and with increased genomic resolution. Cell line experiment studies will be able to take advantage of the rapid development in statistical methods to perform rare variant association, pathway analysis, gene-gene and gene-environment interaction modeling, gene score modeling, etc. [43]. The expansion of these technologies and statistical methods development will allow for increased sensitivity and identification of novel targets with potential clinical applications. Furthermore, as databases expand and additional data from large scale genome wide studies become available, annotation and computational efficiency will improve, driving the integration of multiple technology platforms.

Many of the major efforts in the field of genetics to genotype individuals using next generation sequencing, such as the 1000 Genome Project [6], use populations with established cell lines available for purchase, which will let pharmacogenomics directly take advantage of these genotyping efforts. Since the results of the previous studies suggest that the etiology of these drug response phenotypes are complex, it is likely that many of the drugs and chemicals tested show differential response based on rare variants with large, or accumulating effects. Additionally, as the cost of NGS decreases, it may be feasible to perform NGS genotyping on more private cohorts of cell lines.

One way that these technologies can be integrated is through pathway analysis. Pathway analyses are also used with individual technology platforms to help identify the biological processes impacted by the genes of interest and for dimensionality reduction for predictive modelling. Although there are many different types of pathway analyses routinely performed, one common and relatively simple approach is to identify which genes are significantly perturbed in an experiment and determine if certain sets of these genes are enriched for any particular biological pathway, usually obtained from pathway databases, such as KEGG (<http://www.genome.jp/kegg/>), GO (<http://www.geneontology.org/>), and MSigDB (<http://www.broadinstitute.org/gsea/msigdb/index.jsp>), but many other pathways databases also exist, including many pharmacogenomics-specific resources like PharmGKB [45].

Another important area of opportunity is the possibility to interrogate gene-gene interactions. While all pharmacogenomics studies are inherently gene-environment interaction types of analyses, the sample sizes in clinical trials is often under-powered for the analyses of interactions. There have been some examples of interaction analysis in clinical trials data [46], to date the cell line models have not been interrogated for gene-gene interactions, or epistasis. Just as epistatic interactions are known to be important in the

etiology of other complex traits, and given the similarity in the dose response phenotypes in dose response, this is an important future direction to be assessed using the cell line models [47].

The benefits of the cell line models described herein—such as, potential for larger sample sizes due to cost effectiveness—creates an opportunity to explore previously intractable problems of pharmacogenomics if relying on clinical trial sorts of study designs. Mirroring trends in common disease mapping, there is increasing appreciation for the fact that very large samples will be needed for association mapping (based on the effect sizes seen) and for the newer approaches discussed above. The ready availability of new samples and the potential for ready biological and functional follow up is an increasing advantage of the model system.

Finally, the LCLs are being used to increase the scope of toxicogenomics, investigating heritability underlying differential responses across populations [10]. While there will be some differences in the details of experimental design based on toxicological instead of pharmacological goals (e.g., some differences in specific endophenotypes, most relevant sources of cell lines, etc.), many of the same lessons in regards to the complex etiology of the genetic response to chemical exposure will likely be very similar.

Conclusion

The cell line models discussed here offer a cost-effective solution for interrogating many issues surrounding heritability and disease in pharmacogenomics. Moreover, with the wealth of already publicly available data from cell line models of pharmacogenomics response, it is likely researches are only scratching the surface on the benefits offered by these models.

Acknowledgments

This work was supported by NIH NCI RO1CA161608 from the National Cancer Institute.

References

1. Welsh M, Mangravite L, Medina MW, et al. Pharmacogenomic discovery using cell-based models. *Pharmacol Rev.* 2009; 61(4):413–29. [PubMed: 20038569]
2. Cox NJ, Gamazon ER, Wheeler HE, Dolan ME. Clinical translation of cell-based pharmacogenomic discovery. *Clin Pharmacol Ther.* 2012; 92(4):425–7. [PubMed: 22910437]
3. Wheeler HE, Dolan ME. Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation. *Pharmacogenomics.* 2012; 13(1):55–70. [PubMed: 22176622]
4. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics.* 1990; 6(3):575–7. [PubMed: 2184120]
5. International HapMap C. The International HapMap Project. *Nature.* 2003; 426(6968):789–96. [PubMed: 14685227]
6. Abecasis GR, Altshuler D, et al. Genomes Project C. A map of human genome variation from population-scale sequencing. *Nature.* 2010; 467(7319):1061–73. [PubMed: 20981092]
7. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* 2008; 455(7216):1061–8. [PubMed: 18772890]

8. Choy E, Yelensky R, Bonakdar S, et al. Genetic analysis of human traits *in vitro*: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* 2008; 4(11):e1000287. [PubMed: 19043577]
9. Motsinger-Reif, A.; Havener, T.; Hardison, N.; Peters, E.; Beam, A.; Everitt, L.; McLeod, H., editors. Joint Statistical Meetings. 2011. Ex-Vivo Modeling for Heritability Assessment and Genetic Mapping in Pharmacogenomics.
10. Abdo, NXM.; Kosyk, O.; Huang, R.; Sakamuru, S.; Brown, C.; Jack, J.; Gallins, P.; Zhou, Y.; Motsinger-Reif, A.; Austin, C.; Tice, R.; Wright, F.; Rusyn, I., editors. The 1000 Genomes Toxicity Screening Project: Utilizing the Power of Human Genome Variation for Population-Scale In Vitro Testing Society of Toxicology. 2013.
11. Shukla SJ, Duan S, Wu X, Badner JA, Kasza K, Dolan ME. Whole-genome approach implicates CD44 in cellular resistance to carboplatin. *Hum Genomics.* 2009; 3(2):128–42. [PubMed: 19164090]
12. Li L, Fridley B, Kalari K, et al. Gemcitabine and cytosine arabinoside cytotoxicity: association with lymphoblastoid cell expression. *Cancer Res.* 2008; 68(17):7050–8. [PubMed: 18757419]
13. Huang CH, Chen ML, Tsai YL, Tsai MT, Chen CH. Elevated adrenomedullin mRNA in lymphoblastoid cells from schizophrenic patients. *Neuroreport.* 2004; 15(9):1443–6. [PubMed: 15194870]
14. Domcke S, Sinha R, Levine DA, Sander C, Schultz N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat Communications.* 2013; 4:2126.
15. Peters EJ, Motsinger-Reif A, Havener TM, et al. Pharmacogenomic characterization of US FDA-approved cytotoxic drugs. *Pharmacogenomics.* 2011; 12(10):1407–15. [PubMed: 22008047]
16. Watson VG, Motsinger-Reif A, Hardison NE, et al. Identification and replication of loci involved in camptothecin-induced cytotoxicity using CEPH pedigrees. *PLoS One.* 2011; 6(5):e17561. [PubMed: 21573211]
17. Jen KY, Cheung VG. Transcriptional response of lymphoblastoid cells to ionizing radiation. *Genome Res.* 2003; 13(9):2092–100. [PubMed: 12915489]
18. Brown CC, Havener TM, Medina MW, et al. Genome-wide Association and Pharmacological Proling of 29 Anticancer Agents Using Lymphoblastoid Cell Lines. *Pharmacogenomics.* 2014; 15(2):137–46. [PubMed: 24444404]
19. Evangelou E, Trikalinos TA, Salanti G, Ioannidis JP. Family-based versus unrelated case-control designs for genetic associations. *PLoS Genet.* 2006; 2(8):e123. [PubMed: 16895437]
20. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet.* 2008; 17(R2):R143–50. [PubMed: 18852203]
21. Wheeler HE, Gamazon ER, Stark AL, et al. Genome-wide meta-analysis identifies variants associated with platinating agent susceptibility across populations. *Pharmacogenomics J.* 2013; 13(1):35–43. [PubMed: 21844884]
22. Amos W, Driscoll E, Hoffman JI. Candidate genes versus genome-wide associations: which are better for detecting genetic susceptibility to infectious disease? *Proc Biol Sci.* 2011; 278(1709): 1183–8. [PubMed: 20926441]
23. Brown CC, Havener TM, Medina MW, Krauss RM, McLeod HL, Motsinger-Reif AA. Multivariate methods and software for association mapping in dose-response genome-wide association studies. *BioData Min.* 2012; 5(1):21. [PubMed: 23234571]
24. Beam A, Motsinger-Reif AA. Beyond IC50s: Towards Robust Statistical Methods for *in vitro* Association Studies. *J Pharmacogenom Pharmacoproteomics.* 2013; 5:120.
25. Brown C, Havener TM, Everitt L, McLeod H, Motsinger-Reif AA. A comparison of association methods for cytotoxicity mapping in pharmacogenomics. *Front Genet.* 2011; 2:86. [PubMed: 22303380]
26. Lunetta KL. Genetic association studies. *Circulation.* 2008; 118(1):96–101. [PubMed: 18591452]
27. Moskvina V, Schmidt KM. On multiple-testing correction in genome-wide association studies. *Genet Epidemiol.* 2008; 32(6):567–73. [PubMed: 18425821]
28. Che R, Jack JR, Motsinger-Reif A, Brown CC. An Adaptive Permutation Approach for Genome-Wide Association Study: Evaluation and Recommendations for Use. *BioData Min.* 2014; 7:9. [PubMed: 24976866]

29. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 2010; 6(4):e1000888. [PubMed: 20369019]
30. Huang RS, Duan S, Bleibel WK, et al. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc Natl Acad Sci U S A.* 2007; 104(23):9758–63. [PubMed: 17537913]
31. Huang RS, Duan S, Kistner EO, et al. Genetic variants contributing to daunorubicin-induced cytotoxicity. *Cancer Res.* 2008; 68(9):3161–8. [PubMed: 18451141]
32. Huang RS, Duan S, Shukla SJ, et al. Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am J Hum Genet.* 2007; 81(3):427–37. [PubMed: 17701890]
33. Huang RS, Duan S, Kistner EO, Hartford CM, Dolan ME. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. *Mol Cancer Ther.* 2008; 7(9):3038–46. [PubMed: 18765826]
34. Hartford CM, Duan S, Delaney SM, et al. Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. *Blood.* 2009; 113(10):2145–53. [PubMed: 19109566]
35. Shan G. RNA interference as a gene knockdown technique. *Int J Biochem Cell Biol.* 2010; 42(8):1243–51. [PubMed: 19442757]
36. Li L, Fridley BL, Kalari K, et al. Gemcitabine and arabinosylcytosin pharmacogenomics: genome-wide association and drug response biomarkers. *PLoS One.* 2009; 4(11):e7765. [PubMed: 19898621]
37. O'Donnell PH, Gamazon E, Zhang W, et al. Population differences in platinum toxicity as a means to identify novel genetic susceptibility variants. *Pharmacogenet Genomics.* 2010; 20(5):327–37. [PubMed: 20393316]
38. Millward MJ, Boyer MJ, Lehnert M, et al. Docetaxel and carboplatin is an active regimen in advanced non-small-cell lung cancer: a phase II study in Caucasian and Asian patients. *Ann Oncol.* 2003; 14(3):449–54. [PubMed: 12598352]
39. Smirnov DA, Morley M, Shin E, Spielman RS, Cheung VG. Genetic analysis of radiation-induced changes in human gene expression. *Nature.* 2009; 459(7246):587–91. [PubMed: 19349959]
40. Simon JA, Lin F, Hulley SB, et al. Phenotypic predictors of response to simvastatin therapy among African-Americans and Caucasians: the Cholesterol and Pharmacogenetics (CAP) Study. *Am J Cardiol.* 2006; 97(6):843–50. [PubMed: 16516587]
41. Mangravite LM, Medina MW, Cui J, et al. Combined influence of LDLR and HMGCR sequence variation on lipid-lowering response to simvastatin. *Arterioscler Thromb Vasc Biol.* 2010; 30(7):1485–92. [PubMed: 20413733]
42. Morag A, Pasmanik-Chor M, Oron-Karni V, Rehavi M, Stingl JC, Gurwitz D. Genome-wide expression profiling of human lymphoblastoid cell lines identifies CHL1 as a putative SSRI antidepressant response biomarker. *Pharmacogenomics.* 2011; 12(2):171–84. [PubMed: 21332311]
43. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009; 461(7265):747–53. [PubMed: 19812666]
44. Brown CC, Havener TM, Medina MW, et al. A genome-wide association analysis of temozolomide response using lymphoblastoid cell lines shows a clinically relevant association with MGMT. *Pharmacogenet Genomics.* 2012; 22(11):796–802. [PubMed: 23047291]
45. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res.* 2002; 30(1):163–5. [PubMed: 11752281]
46. Darbar D, Motsinger AA, Ritchie MD, Gainer JV, Roden DM. Polymorphism modulates symptomatic response to antiarrhythmic drug therapy in patients with lone atrial fibrillation. *Heart Rhythm.* 2007; 4(6):743–9. [PubMed: 17556195]
47. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered.* 2003; 56(1-3):73–82. [PubMed: 14614241]