Journal of
Data Mining in Genomics & Proteomics

Motsinger-Reif and Rotroff, J Data Mining Genomics Proteomics 2015, 6:3
DOI: 10.4172/2153-0602.1000179

Review Article Open Access

Leveraging Lymphoblastoid Cell Lines for Drug Response Modeling

Alison A Motsinger-Reif[1,2] and Daniel M Rotroff[1,2*]

[1]*Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA*
[2]*Department of Statistics, North Carolina State University, Raleigh, NC, USA*

Abstract

Lymphoblastoid cell lines (LCL) are becoming popular tools for modeling drug response. LCLs, and other *in vitro* assays, offer the ability to test many drugs, doses, and biological samples relatively quickly and inexpensively. In addition, a unique advantage to LCLs is that they are available from a large cohort of individuals, providing the capability to test for genetic variability on a scale not readily available in other *in vitro* systems. Since oftentimes the genotype data is publically available, the experimental costs can be limited to the cost of the drug response phenotyping. Here we describe several advantages and limitations of LCLs. In addition we review several important aspects of LCL experimental design and statistical analysis. Lastly, we present an example of LCLs being successfully used to identify candidate single nucleotide polymorphisms and genes for variability in response to a chemotherapeutic used to treat chronic myeloid leukemia.

Introduction

In vitro models allow for the distinct advantage of interrogating molecular perturbations in human cells while controlling for a variety of experimental conditions that are otherwise not possible in human clinical trials. Significant heterogeneity exists in the response to many drugs, with some individuals responding well to a prescribed therapy, and other individuals with the same disorder showing minimal or no response to the same therapy. This heterogeneity is thought to exist, at least in part, to genetic differences of individuals. Furthermore, it is difficult to address this problem during the drug development phase due to the fact that not all populations of individuals that will end up taking the drug are adequately represented during clinical trials [1-3]. Motivated by the variability of individual responses to medications, the field of pharmacogenomics and precision medicine aims to use genetic information from individuals to identify biomarkers of disease and to develop personalized therapies to improve treatment response and limit adverse outcomes [4]. *In vitro*, cell-based models, serve an important role in pharmacogenomics for early stage research questions (hypothesis generation) and late-stage research questions (functional validation). Epstein-Barr virus (EBV)-transformed, Lymphoblastoid cell lines (LCL), have proven particularly useful in pharmacogenomics, especially in the realm of cancer pharmacogenomics, due to the volume of publically available genetic information on cell lines derived from large cohorts of individuals. This resource provides the genetic variability of a clinical trial cohort, with the study design flexibility of an *in vitro* study. Here, we review advantages and disadvantages to this cell line model, and provide a brief overview of data resources, analysis techniques, and an example of a successful implementation using the LCL model.

Advantages and Disadvantages of the LCL Model

Advantages

In comparison to other model systems, there are a significant number of advantages provided by LCL models which have contributed to its widespread use. A table with key advantages to this model system can be found in (Table 1). The wealth of publically available resources makes this model system extremely cost effective. Examples of these publically available data for established cell lines include the International HapMap Project [5], Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees [6], and the Human Variation Panel Populations [7]. These resources include genome-wide, single nucleotide polymorphism (SNP) data, copy-number variation data, family structure, baseline gene-expression data, and a growing volume of next-generation sequencing data. Additionally, biobanking efforts are increasing the number of locally available genotyped resources available [8,9]. The public availability of these data reduces the cost of genetic mapping to only the cost of the drug response phenotyping. Welsh and colleagues provide a more detailed review of these and other resources in their 2009 review [10]. While many of the resources have evolved since this earlier review, the resources themselves are still commonly used.

An additional advantage provided by this model system is that there are relatively few confounding factors compared to those that must be accounted for in clinical trials. Clinical trials are rarely designed for genetic analysis, and this can make the downstream genetic and drug response analysis very complicated due to confounding issues of study design, treatment regimens, and others. Although the potential for confounding certainly exists with *in vitro* model systems (e.g., batch effects, cell proliferation rates, passage number, etc.) these can be readily corrected through proper statistical analysis and technical execution [11,12].

Third, LCLs are amenable to robotic automation creating a high-throughput assay system, which when combined with its relatively low cost, provides an opportunity for efficiently screening large number of drugs/chemicals. A recent study by Abdo et al. tested 179 chemicals in LCLs from 1086 individuals to interrogate genetic associations with environmental chemical toxicity [13]. In addition to the large numbers of chemicals/drugs that can be tested, LCLs are also available from large populations of individuals. Clinical trials usually consist of hundreds of mostly unrelated individuals, LCLs allow for as many cell lines as

***Corresponding author:** Daniel Rotroff, NCSU Statistics Department, 2311 Stinson Drive, Raleigh, NC 27695-7566, USA, Tel: (919) 413-0147; E-mail: daniel.rotroff@ncsu.edu

Received August 07, 2015; **Accepted** September 03, 2015; **Published** September 09, 2015

Citation: Motsinger-Reif AA, Rotroff DM (2015) Leveraging Lymphoblastoid Cell Lines for Drug Response Modeling. J Data Mining Genomics Proteomics 6: 179. doi:10.4172/2153-0602.1000179

Key Advantages	Key Disadvantages
Wealth of publically available resources for LCL genotypes, which reduces the cost of genetic mapping to only the cost of the drug response phenotyping.	Difficulty of translating *in vitro* results to *in vivo* biological/clinical relevance
Fewer confounding factors compared to human clinical trials (e.g. concomitant medications, treatment regimes)	Extrapolating *in vitro* test concentrations to *in vivo* doses or exposures
Amenable to robotic automation which helps maintain quality control and reduces cost of testing	Cell lines amenable to immortalization may not be from the target tissue of interest for certain drugs
Amenable to high-throughput testing, which allows large numbers of chemicals, concentrations, cell lines, and replicates to be tested rapidly.	Not all biological pathways are represented in LCLs, which may limit its utility for certain studies
Availability of genetic information allows for heritability estimation of phenotypes without needing to design complex human studies	Many metabolizing enzymes (e.g. CYP450s) are not represented in LCLs and other cell lines, which under some circumstances can make testing pro-drugs and other therapeutics challenging.
LCLs can be easily manipulated to validate study findings using knock-down or forced gene expression methods.	

Table 1: Key advantages and disadvantages to the LCL Model System.

experimentally and economically feasible while allowing for heritability estimations and association mapping from family-based study designs [10,14].

Lastly, where replication of genetic associations from clinical trials is generally infeasible due to cost and practicality, the ease of genetic manipulation via knock-down or forced gene expression in LCLs provides a readily available system for functional validation of any findings in the initial gene mapping exercise. Several studies have used LCL models to functionally validate findings using knock-down experiments [15-17] and the advantages offered by these techniques are described in greater detail below.

Disadvantages

As with any model system, LCLs have several key limitations. Translating the biological relevance of *in vitro* phenotypes to *in vivo* outcomes remains a considerable challenge [10]. This limitation can be particularly significant when the drug is highly toxic, with a narrow therapeutic index. In many cases, it is still not clear whether *in vitro* sensitivity to drugs, such as chemotherapeutics, relates to drug efficacy, toxicity, or both. A related challenge is extrapolating *in vitro* test concentrations to relevant *in vivo* doses or exposures in humans. This can be especially difficult when very little pharmacokinetic information is available; however, models using combinations of *in vitro* and *in silico* assays have been previously developed to help address this limitation [18,19].

LCLs are created from peripheral B lymphocytes that have undergone transformation by EBV [20]. However, this cell type may not represent the tissue of interest for some drugs, and unfortunately, not all tissue types are amenable to EBV transformation. Hepatocytes are generally considered a highly relevant tissue for pharmacogenetic studies due to genetic variants impacting drug metabolizing enzymes, but hepatocytes are difficult and expensive to obtain. With continued advancement in stem cell technology, it is possible that the number of tissue types available for conducting *in vitro* genetic association mapping will expand.

Finally, cell line models do not always express all pharmacologically relevant molecular pathways. It is important to consider the limitations of each model cell line in the context of the drug(s) being investigated. Similar to most cell lines, LCLs do not readily express cytochrome P450 (CYP450) genes. CYP450s are enzymes are responsible for metabolizing many drugs and xenobiotics. For example, prodrugs such as tamoxifen, require metabolic activation by CYP450s (i.e. CYP2D6 and CYP3A4) in order to perform their therapeutic action of binding, and ultimately antagonizing the estrogen receptor [21]. Therefore, it may be preferable to test an active metabolite, such as 4-hydroxytamoxifen, in LCLs rather than the parent chemical. In spite of not expressing CYP450s, LCLs do express approximately half of the known genome, making them a useful model system for generating and testing many different hypotheses [22].

Genotyping Approach

There is a wide range of options available for investigating many types of genetic variation in cell lines, such as LCLs. These methods range from testing candidate lists of targets (gene expression, SNPs, etc.) to testing the entire genetic landscape [23]. Although, it is becoming increasingly common to conduct genome-wide approaches, the best method will likely depend on the research questions being asked. Many studies conducted so far have relied on SNP data, but other types of genetic variation (e.g., epigenetic variation, copy number variation, gene expression) are being implemented more frequently [24,25]. Lastly, the International Hap Map project, and the other resources listed above, is making data spanning a wide range of technologies publically available [5-7].

Approaches for Statistical Analysis

Careful planning and deliberate approaches are necessary to ensure proper quality control (QC) to prevent potential issues related to experimental artifacts, and are highlighted in Choy et al. [11]. It is critically important to have sufficient biological and technical replication to evaluate the reliability and reproducibility of the experimental results [12]. Additionally, metrics for outlier detection, dose response relationships are often used in the QC process [12]. For genetic association studies, the more broadly used genome-wide association studies (GWAS) QC approaches can also be applicable for LCLs, and commonly include filtering data by percent missing data by individual, percent missing data by SNP, Hardy-Weinberg equilibrium violation, minor allele frequencies, and others. Additional details on best practices for genetic quality control can be found in Motsinger-Reif et al. [12].

The motivation behind most gene mapping experiments using LCLs is to associate genotypes (or some other measure of biological variation) with a given phenotype (e.g., cytotoxicity). Appropriate methods will depend heavily on the study design and underlying assumptions (e.g., population stratification, batch effects). The most common approach, particularly in toxicogenomics studies, is to fit a non-linear dose response model, typically a Hill model, and use a parameter from the model to summarize the response profile [13]. These parameters include the maximum efficacy as a percentage of positive control (EMAX) or 50% inhibitory/effective concentration, IC50 or EC50, respectively [26]. There are several important assumptions that must be recognized when implementing this approach. 1) Is the non-linear model chosen the most appropriate to describe the underlying biology, 2) does the IC50 (or other univariate measure) sufficiently capture the variation between experimental conditions? Recently, simulation studies have indicated that modeling the full concentration response profile using a multivariate analysis of covariance (MANCOVA) or multiple regression approach can be more powerful than using univariate summary measures [26-28].

Typically, the phenotypic response is then compared to each variant tested. For testing allelic of genotypic coding an additive or dominance model is often assumed, meaning that the number of minor alleles changes with the phenotype linearly as in the former model, or the heterozygote and one of the homozygotes are treated equally in the latter model [29]. A more detailed account regarding different genetic models can be found in the article published by Lunetta [29]. Confounding variables can be incorporated into these models as covariates. However, a challenge in GWAS is that the number of variants being tested is often in the millions, and this large number of comparisons can lead to many type I errors (false positives). There are several commonly used approaches for correcting the problem of multiple comparisons. The Bonferroni correction is implemented by dividing the *p* value threshold by the number of statistical tests to obtain the new, corrected threshold for significance. This approach is considered conservative in its ability to reduce type I error, but can lead to inflated type II error (false negatives) in studies with smaller sample sizes. An alternative to the Bonferroni correction is to use a False Discovery Rate (FDR) approach, which aims to address inflated type I error while limiting inflated type II error [30]. A traditional threshold in GWAS is to use a *p*-value threshold of 5×10^{-8}, which assumes a 5% chance that a positive association occurs randomly if 1 million SNPs are being tested [31].

In some circumstances, it is necessary to determine how well the biomarker or biomarkers (e.g., SNP) will predict the drug response phenotype. To develop a valid predictive model it is inadequate to assume that an observed association between a genetic marker and a phenotype will yield meaningful predictive capabilities. More comprehensive information on developing predictive models using biological data can be found in a book by Streyerberg [32] and an article by Moons et al. [33]. Briefly, it is important that the model developed from the biomarker, or set of biomarkers, is tested for its predictive performance using a dataset other than the one from which the model was constructed (training data). Cross-validation is one method of doing this, where a subset of the entire dataset is withheld from the model construction in order to provide a test data set for testing the predictive capability of the model. In order for a model to be clinically relevant, it must have the ability to accurately detect true-positives (sensitivity), and also accurately distinguish false-negatives and true-negatives (specificity) [34]. If a model is not sufficiently sensitive then it may fail to detect individuals that may be at risk for an adverse effect. On the other hand, if a model is not sufficiently specific it may lead to individuals inappropriately being included for a therapy or for unnecessary follow-up testing [34].

Due to the large numbers of variants commonly tested in these cell line models, it is critical that candidate biomarkers and therapeutic targets are tested in a follow-up assay to confirm their association with the phenotype. A routinely used approach is to knock down the gene of interest using small interfering RNA (siRNA). Through silencing the candidate gene, it is possible to elucidate the functional effect that target has on the outcome of interest. This method is reviewed in detail by Shan [35]. Another alternative, forced gene expression, is used less often and requires upregulating of the expression of the candidate gene to investigate the effect on the outcome of interest. Either forced gene expression or siRNA knock downs can be readily performed in the *in vitro* LCL model or *in vivo* mammalian model systems, such as a murine model.

Examples of Success

The substantial advantages of the LCL model have led to many exciting successes. Thorough reviews on these successes can be found in Welsh et al. [10] and Cox et al [36]. Here we briefly highlight one recent example of a successful use of LCL models for pharmacogenomic drug response by Tong et al [37]. Homoharringtonine (HHT) is a naturally occurring chemical used in China to treat acute myeloid leukemia (AML) and chronic myeloid leukemia (CML) for many years, and a synthetic version of HHT has been recently approved by the US FDA for CML [37,38]. However, not everyone responds to HHT, so Tong et al. conducted a GWAS to determine what SNPs were associated with cytotoxicity to HHT in LCLs. The cytotoxicity dose response was modeled and the area under the curve (AUC) was used as the phenotype and associated with SNPs in a GWAS. Although no SNPS reached genome wide significance ($p<1\text{x}10^{-8}$), 18 SNPs had $p<1\text{x}10^{-5}$ and 281 SNPs had $p<1\text{x}10^{-4}$. Subsequently, 15 regions with ≥ 2 SNPs with $p<1\text{x}10^{-4}$ were identified and up to 3 genes within these regions were tested for functional validation. mRNA gene expression was also used to narrow this candidate gene list down to 7 genes for functional validation involving siRNA knockdown followed by MTS cytotoxicity. Interestingly, *CCDC88A*, *CTBP2*, and *SOCS4* knockdowns demonstrated significant resistance to HHT sensitivity. *CCDC88A* impacts AKT signaling which is involved in DNA replication and cell proliferation, and has been shown to be relevant in AML [37,39]. In addition, *CTBP2* and *SOCS4* are also involved in signaling pathways that have been implicated in various cancers [37]. Although, additional validation is needed to establish the clinical relevancy of these target genes, this study highlights the tremendous potential that LCLs can provide when used in conjunction with proper validation.

Opportunities and Conclusion

The rapid advancements of cutting edge technologies offer exciting new opportunities for leveraging the benefits from LCL models. One example is next-generation sequencing, which is capable of measuring multiple types of genetic variations (e.g. SNPs, copy-number aberrations) in the same assay. Additional research into epigenetic modification and proteomics may generate new and exciting hypothesis for drug response variation and other research areas. The use of established LCL populations can provide opportunities for combining studies using multi-omics and integrative analyses, which in turn may provide new insight into disease etiology and drug response variation. One way to integrate these technologies is through pathway-based analyses. A variety of methods for conducting pathway analyses is available and reviewed by Khatri et al. [40]. Furthermore, many publically available databases for conducting pathway analysis are readily accessible, and include KEGG (http://www.genome.jp/kegg/), Gene Ontology (http://www.geneontology.org/), and MSigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp) to name only a few.

The LCL model offers an exciting and cost-effective solution for investigating many research questions regarding heritability of disease, drug response, disease etiology, precision medicine, and many others. Given the abundance of publically available data to augment the use of this cell line, it is likely that the use of LCLs will provide many new and impactful discoveries in the near future.

References

1. Murthy VH, Krumholz HM, Gross CP (2004) Participation in cancer clinical trials: Race-, sex-, and age-based disparities. JAMA 291: 2720–2726.
2. Baquet CR, Commiskey P, Daniel Mullins C, Mishra SI (2006) Recruitment and participation in clinical trials: Socio-demographic, rural/urban, and health care access predictors. Cancer Detect Prev 30: 24–33.
3. Jagsi R, Motomura AR, Amarnath S, Jankovic A, Sheets N, et.al. (2009) Under-representation of women in high-impact published clinical cancer research. Cancer 115: 3293–3301.
4. Insel PA, Amara SG, Blaschke TF (2015) Introduction to the Theme 'Precision Medicine and Prediction in Pharmacology.' Annu Rev Pharmacol Toxicol 55: 11–14.
5. Gibbs RA (2003) The International HapMap project. Nature 426: 789–796.
6. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, et.al. (1990) Centre d'etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. Genomics 6: 575–577
7. Consortium, 1000 Genomes Project and others (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
8. Birdwell KA, Grady B, Choi L, Xu H, Bian A, et.al. (2012) The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement in kidney transplant recipients. Pharmacogenet. Genomics 22: 32–42.
9. McCarty CA, Wilke RA (2010) Biobanking and pharmacogenomics. Pharmacogenomics 11: 637–641.
10. Welsh M, Mangravite L, Medina MW, Tantisira K, Zhang W, et.al. (2009) Pharmacogenomic discovery using cell-based models. Pharmacol Rev 61: 413–429.
11. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, et.al. (2008) Genetic analysis of human traits in vitro: Drug response and gene expression in lymphoblastoid cell lines. PLoS Genet 4: e1000287.
12. Motsinger-Reif AA, Chad Brown, Tammy Havener, Nicholas Hardison, Eric Peters, et.al. (2011) Ex-vivo modeling for heritability assessment and genetic mapping in pharmacogenomics. In Proceedings of the Joint Statistical Meeting 306–318.
13. Abdo N, Xia M, Brown CC, Kosyk O, Huang R, et.al. (2015) Population-Based in Vitro Hazard and Concentration–Response Assessment of Chemicals: The 1000 Genomes High-Throughput Screening Study. Environ Health Perspect 123: 458–466.
14. Wheeler HE, Dolan ME (2012) Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation. Pharmacogenomics 13: 55–70.
15. Shukla SJ, Duan S, Wu X, Badner JA, Kasza K, et.al. (2009) Whole-genome approach implicates CD44 in cellular resistance to carboplatin. Hum Genomics 3: 128–142.
16. Li L, Fridley B, Kalari K, Jenkins G, Batzler A, et.al. (2008) Gemcitabine and cytosine arabinoside cytotoxicity: association with lymphoblastoid cell expression. Cancer Res. 68: 7050–7058.
17. Huang CH, Chen ML, Tsai YL, Tsai MT, Chen CH (2004) Elevated adrenomedullin mRNA in lymphoblastoid cells from schizophrenic patients. Neuroreport 15: 1443–1446.
18. Rotroff DM, Wetmore BA, Dix DJ, Ferguson SS, Clewell HJ, et.al. (2010) Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. Toxicol Sci 117: 348–358.
19. Wetmore BA, Wambaugh JF, Ferguson SS, Sochaski MA, Rotroff DM, et.al. (2012) Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. Toxicol Sci 125: 157–174.
20. Hui-Yuen J, McAllister S, Koganti S, Hill E, Bhaduri-McIntosh S (2011) Establishment of Epstein-Barr Virus Growth-transformed Lymphoblastoid Cell Lines. J Vis Exp 8: 3321.
21. Desta Z, Ward BA, Soukhova NV, Flockhart DA (2004) Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system in vitro: prominent roles for CYP3A and CYP2D6. J Pharmacol Exp Ther 310: 1062–1075.
22. Jen KY, Cheung VG (2003) Transcriptional Response of Lymphoblastoid Cells to Ionizing Radiation. Genome Res. 13: 2092–2100.
23. Amos W, Driscoll E, Hoffman JI (2010) Candidate genes versus genome-wide associations: Which are better for detecting genetic susceptibility to infectious disease? Proc Biol Sci 22: 278.
24. Zhang X, Zhang W (2014) Pharmacoepigenetics explores the epigenetic contribution to anti-cancer agent-induced cytotoxicities. Stem Cell Epigenetics 1: e359.
25. Jeon JP, Shim SM, Nam HY, Baik SY, Kim JW, et.al. (2007) Copy number increase of 1p36.33 and mitochondrial genome amplification in Epstein–Barr virus-transformed lymphoblastoid cell lines. Cancer Genet Cytogenet 173: 122–130.
26. Beam A, Motsinger-Reif A (2014) Beyond IC50s: Towards Robust Statistical Methods for in vitro Association Studies. J Pharmacogenomics Pharmacoproteomics 5: 1000121.
27. Brown CC, Havener TM, Medina MW, Krauss RM, McLeod HL, et.al. (2012) Multivariate methods and software for association mapping in dose-response genome-wide association studies. BioData Min 5: 21.
28. Brown C, Havener TM, Everitt L, McLeod H, Motsinger-Reif AA (2011) A comparison of association methods for cytotoxicity mapping in pharmacogenomics. Front Genet 2: 86.
29. Lunetta KL (2008) Genetic association studies. Circulation 118: 96–101.
30. Moskvina V, Schmidt KM (2008) On multiple-testing correction in genome-wide association studies. Genet Epidemiol 32: 567–573.
31. Monteiro ANA, Coetzee GA, Freedman ML, De Biasi M, Casey G, et.al. (2010) Principles for the post-GWAS functional characterisation of risk loci. Nat Preced.
32. Steyerberg EW (2008) Clinical prediction models: A practical approach to development, validation, and updating.
33. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: What, why, and how? BMJ 338: b375.
34. Pencina MJ, D' Agostino RB Sr, D' Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Stat Med 27: 157–172.
35. Shan G (2010) RNA interference as a gene knockdown technique. Int J Biochem Cell Biol 42: 1243–1251.
36. Cox NJ, Gamazon ER, Wheeler HE, Dolan ME (2012) Clinical Translation of Cell-Based Pharmacogenomic Discovery. Clin Pharmacol Ther 92: 425–427.
37. Tong Y, Niu N, Jenkins G, Batzler A, Li L, et.al. (2015) Identification of genetic variants or genes that are associated with Homoharringtonine (HHT) response through a genome-wide association study in human lymphoblastoid cell lines (LCLs). Front Genet 5: 465.
38. Kantarjian HM, O'Brien S, Cortes J (2013) Homoharringtonine/omacetaxine mepesuccinate: The long and winding road to food and drug administration approval. Clin Lymphoma Myeloma Leuk 13: 530–533.
39. Park S, Chapuis N, Tamburini J, Bardet V, Cornillet-Lefebvre P, et.al. (2010) Role of the PI3K/AKT and mTOR signaling pathways in acute myeloid leukemia. Haematologica 95: 819–828.
40. Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 8: e1002375.