OXFORD

Human Molecular Genetics, 2024, 1–11
https://doi.org/10.1093/hmg/ddae138
Original Article

Deciphering single-cell gene expression variability and its role in drug response

Sizhe Liu[1] and Liang Chen[2,*]

[1]Thomas Lord Department of Computer Science, University of Southern California, 941 Bloom Walk, Los Angeles, CA 90089, United States
[2]Department of Quantitative and Computational Biology, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, United States

*Corresponding author. Department of Quantitative and Computational Biology, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, United States. Email: liang.chen@usc.edu

Abstract

The effectiveness of drug treatments is profoundly influenced by individual responses, which are shaped by gene expression variability, particularly within pharmacogenes. Leveraging single-cell RNA sequencing (scRNA-seq) data, our study explores the extent of expression variability among pharmacogenes in a wide array of cell types across eight different human tissues, shedding light on their impact on drug responses. Our findings broaden the established link between variability in pharmacogene expression and drug efficacy to encompass variability at the cellular level. Moreover, we unveil a promising approach to enhance drug efficacy prediction. This is achieved by leveraging a combination of cross-cell and cross-individual pharmacogene expression variation measurements. Our study opens avenues for more precise forecasting of drug performance, facilitating tailored and more effective treatments in the future.

Keywords: pharmacogene; drug efficacy; single-cell RNA-seq; expression variation; machine learning

Introduction

Precision medicine, a revolutionary approach that acknowledges individual variability in drug response, has gained considerable attention in recent years owing to its potential to improve the effectiveness of drug treatments. This approach recognizes the inherent diversity in individual responses to drug therapies, a variability deeply rooted in genetic differences [1, 2]. In the pursuit of understanding the distinctive variations in drug response among individuals, research efforts are directed toward pharmacogenes [3], genes within an individual's genome that profoundly influence their response to medications. These genes encode proteins involved in drug action, toxicity, transport, or metabolism, all of which play a pivotal role in determining drug efficacy and safety [4–6]. For instance, CYP2D6 is responsible for the metabolism of about 20% of commonly prescribed drugs across various medical fields, including psychiatry, pain management, and cardiology [7]. Individuals can be poor, intermediate, extensive, or ultra-rapid metabolizers based on their CYP2D6 genotypes [8]. Additionally, P-glycoprotein (ABCB1) plays a key role in drug transport, particularly in expelling anticancer drugs from cells, thereby contributing to multidrug resistance in cancer [9]. Variants in the ABCB1 gene can lead to differences in drug absorption and bioavailability. For example, certain variants might reduce the effectiveness of drugs by increasing their efflux from cells, leading to lower intracellular concentrations [10].

Genetic polymorphisms within pharmacogenes have been recognized as a significant contributor to the variability in drug response [11]. The exploration of genetic variants extends beyond coding regions to encompass regulatory elements such as promoters, enhancers, and microRNA binding regions [12]. Notably, there is a particular focus on the expression Quantitative Trait Loci (eQTLs) that influence the expression levels of pharmacogenes [13, 14]. Genetic variations in these pharmacogenes can give rise to differences in drug metabolism, absorption, distribution, and target interactions, which in turn can result in varying therapeutic outcomes and the potential for adverse drug reactions.

It is not surprising that studying the expression variation of pharmacogenes directly, in addition to their related genetic variants, can still give us essential information for predicting drug responses. Simonovsky et al. [15] developed the local coefficient of variation (LCV) as an analytical tool to probe the relationship between gene expression variability and drug efficacy, utilizing bulk RNA-seq data. Their findings reveal that drugs targeting genes with high across-individual variability in expression often exhibit reduced effectiveness within the broader population. This study underscores the importance of considering gene expression variability in medication design.

Expanding upon these foundational insights, we extend our analysis to leverage single-cell RNA sequencing (scRNA-seq) data. Recent advancements in scRNA-seq technologies and their related analysis tools have opened new horizons in understanding gene expression variability at an unprecedented resolution [16–19]. Our primary aim is to explore the variability in the expression of pharmacogenes, genes directly involved in drug response, at the cellular level. More precisely, we aim to dissect and decipher

Received: May 3, 2024. **Revised:** August 14, 2024. **Accepted:** September 4, 2024

the LCV of these crucial pharmacogenes across a myriad of cell types within eight distinct human tissues.

Our analysis has unveiled a plethora of interesting discoveries. First, we have uncovered high expression variation among pharmacogenes, not only between different individuals but also between different cells of the same individual. Such variation is usually consistently high for cells across different cell types of the same tissue or cells across different tissues of the same cell type. Additionally, we have investigated the correlation between the LCV of pharmacogenes and the efficacy of associated drugs. Our results align with previous findings, demonstrating a negative correlation between cross-individual expression variability of pharmacogenes and drug efficacy. Finally, we explore the potential of integrating cross-cell and cross-individual LCV data to predict drug efficacy, highlighting that the expression variability of pharmacogenes may be a pivotal contributor to the observed variability in drug response, even within a given tissue microenvironment. In essence, our research illuminates the complexity between gene expression heterogeneity and drug response, bringing us one step closer to the era of truly personalized medicine.

Results

Pharmacogenes are generally more variable than non-pharmacogenes across cells

Pharmacogenes have been reported to exhibit higher cross-individual expression variability than other protein-coding genes [15]. To test whether pharmacogenes' expression variability is also high across different cells of the same individuals, we obtained the snRNA-seq data across 15 944 cells from eight tissues and sixteen donors [20]. We utilized the local coefficient of variation (LCV) [15] as a metric for assessing expression variability. To obtain the cross-cell LCV of each gene, we calculated the LCV for each cell type by averaging the LCV values from the sixteen donors. The overall tissue-level LCV for that gene was then obtained by averaging these cell-type-level LCVs. As shown in Fig. 1A, except for skin tissues, pharmacogenes consistently demonstrate significantly higher cross-cell variability compared to non-pharmacogenes (P values < 0.05, T-tests).

Our own analysis of the population-level GTEx data [21] further corroborated that pharmacogenes typically exhibit increased expression variability across different individuals when compared to non-pharmacogenes (as shown in Fig. 1B, P value < 0.001, T-tests). It is worth noting that the esophagus tissue in the bulk data corresponds to esophagus mucosa and esophagus muscularis in the snRNA-seq data.

To distinguish different types of pharmacogenes and explore potential differences in their LCV patterns compared to non-pharmacogenes, we categorized pharmacogenes into three main functional groups: “regulation,” “transport,” and “metabolism.” In our analysis of cross-cell LCVs, we found significant differences between all three groups of pharmacogenes and non-pharmacogenes in six out of eight tissues studied, excluding breast and skin tissues. Specifically, the “regulation” group exhibited additional significant differences between pharmacogenes and non-pharmacogenes in breast tissue. Across all tissues, our analysis of cross-individual LCVs consistently showed significant differences between all three groups of pharmacogenes and non-pharmacogenes. Detailed plots illustrating these findings can be found in Supplementary Figs 3 and 4.

We conducted a more detailed investigation into the correlation between cross-cell and cross-individual LCVs for pharmacogenes for the eight different tissues. As shown in Fig. 1C, all tissues exhibit a significant positive correlation between the two types of variability measurements. This suggests that there might be shared underlying mechanisms or factors contributing to the expression variability of pharmacogenes within these particular tissue contexts, both at the intra-individual and inter-individual levels. The positive correlation suggests that factors affecting variability within individual cells also contribute to variability across different individuals. Conversely, a lack of correlation would imply that high cross-individual variability is driven by population-specific factors like genetic diversity or environmental influences, which might not manifest uniformly within individual cells.

Variability of pharmacogenes at the cell-type level

Analyzing variability patterns of pharmacogenes across cells of the same cell type and tissue, our study found that pharmacogenes still demonstrate higher variability than non-pharmacogenes. As an illustration, in the esophagus mucosa tissue (Fig. 2A), six out of seven cell types exhibit a higher cross-cell LCV for pharmacogenes compared to non-pharmacogenes. This contrast in variability is particularly evident in three epithelial cell types, namely basal, squamous, and suprabasal cells, as well as in endothelial vascular cells (P value < 0.001, T-tests). Moreover, our exploration extends to other tissues (Supplementary Figs 1 and 2). In six out of seven cell types of the heart and all eight cell types of the prostate, pharmacogenes exhibit significantly higher LCV values than non-pharmacogenes.

The extended scope of our analysis further solidifies our findings. Figure 2B provides a comprehensive visual representation of our observations, presenting the T-test p-values that compare pharmacogene and non-pharmacogene expression variability across all cell types found in the eight distinct tissues. More than 75% (19 out of 25) of cell types show pronounced pharmacogene variability vs. non-pharmacogenes in at least one tissue.

For each pharmacogene, we calculated the range in LCV across different cell types within identical tissues. As depicted in Fig. 2C, pharmacogenes demonstrated a comparable or even narrower range (observed in skin and breast tissues) in their LCVs compared to non-pharmacogenes. Notably, the LCVs of pharmacogenes exhibit similarity between pharmacogenes and non-pharmacogenes in the skin tissue (Fig. 1A). For the breast tissue, pharmacogenes show elevated LCVs in contrast to non-pharmacogenes (Fig. 1A). This heightened LCV is consistently observed across different cell types within the breast, resulting in a reduced overall range (Fig. 2C).

Consistent variability patterns of pharmacogenes in the same cell types across different tissues

We analyzed the LCV distribution for specific cell types across different tissues. Our selection criteria included only those cell types found in more than three tissue types from our dataset, including “Endothelial cell (vascular),” “Fibroblast,” and “Adipocyte.” For each pharmacogene, we computed the range of LCV across different tissues of the same cell type. Our results yielded statistically significant differences in the distribution of LCV ranges between pharmacogenes and non-pharmacogenes (Fig. 3, P-value < 0.05 for Endothelial cells and Adipocytes, T-tests). Notably, pharmacogenes displayed a lower range of LCV compared to non-pharmacogenes across various tissues. This pattern suggests a consistently high expression variation for pharmacogenes across varied tissue environments within specific cell types. Thus, compared to non-pharmacogenes, pharmacogenes tend to exhibit variation across different cells of the same cell type

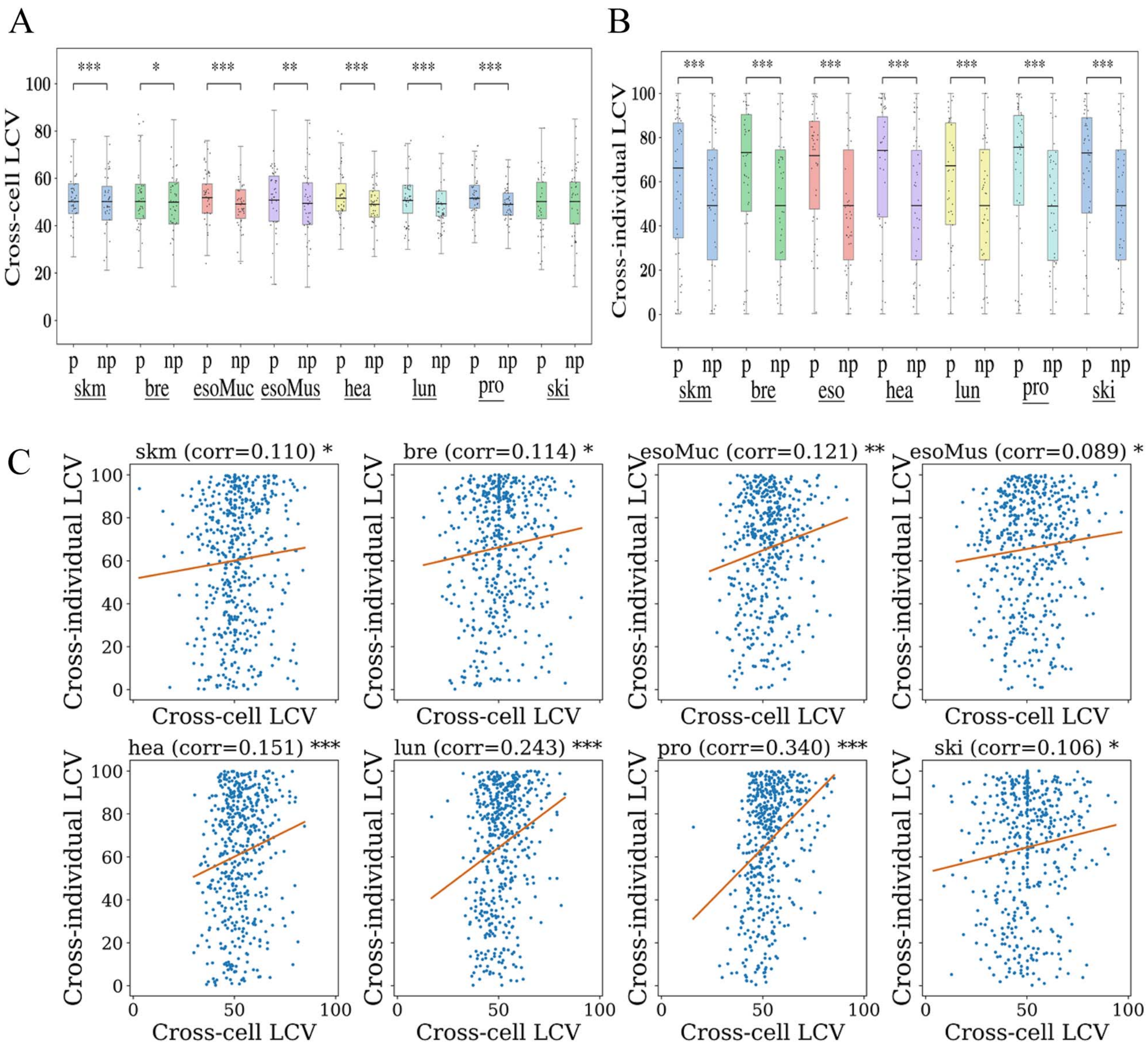

Figure 1. Expression variability of pharmacogenes in different tissues. (A) Comparison of pharmacogenes(p) and non-pharmacogenes(np) for their cross-cell expression variability. (B) Comparison of pharmacogenes(p) and non-pharmacogenes(np) for their cross-individual expression variability. (C) Comparison of cross-cell variability and cross-individual variability of pharmacogenes. The tissue types included are skeletal muscle (skm), breast (bre), esophagus mucosa (esoMuc), esophagus muscularis (esoMus), heart (hea), lung (lun), prostate (pro), and skin (ski). For cross-individual LCV consideration, esophagus (eso) tissue is considered without further considering the sub-locations. ***: *P*-value < 0.001, **: *P*-value < 0.01, *: *P*-value < 0.05; T-tests for A and B and spearman correlation tests for C.

(see Fig. 2A and B). Moreover, they consistently display such high variation across different cell types within the same tissue environment (Fig. 2C) and across different tissues of the same cell type (Fig. 3).

Distribution of pharmacogenes' LCV values across different cell types and different tissues

Figure 4A illustrates how pharmacogenes distribute across tissue cell types based on their peak LCV (largest local coefficient of variation), with LCV values averaged across different individuals. Let N_{ij} denote the count of pharmacogenes that exhibit their maximal LCV within a cell type *i* of tissue *j*. The average of N_{ij} across all possible combinations of *i* and *j* is 14.78. Cell types within the skin (including epithelial cells, sebaceous cells, and unknown types) and the heart (including endothelial cells, immune, fibroblasts, and adipocytes) show N_{ij} values exceeding the average of 14.78. Conversely, all cell types in tissues like the lung and prostate have N_{ij} values below this average. This pattern highlights the diversity in LCV distribution across various cell types and tissues.

The heatmaps in Fig. 4B and C display the correlation between LCV values among different cell types and tissues (averaged across multiple individuals) using hierarchical clustering for pharmacogenes and non-pharmacogenes, respectively. Figure 4B demonstrates that, while most cell types exhibit a low positive correlation with each other, a notably high correlation is observed between heart and prostate myocytes. Additionally, myocytes and epithelial cells exhibit a stronger correlation with each other compared to other cell types. These findings provide valuable insights into the relationships between different cell types and tissues regarding the variability of pharmacogenes. Conversely, Fig. 4C reveals that the correlations for non-pharmacogenes are predominantly weak, showing no significant trends.

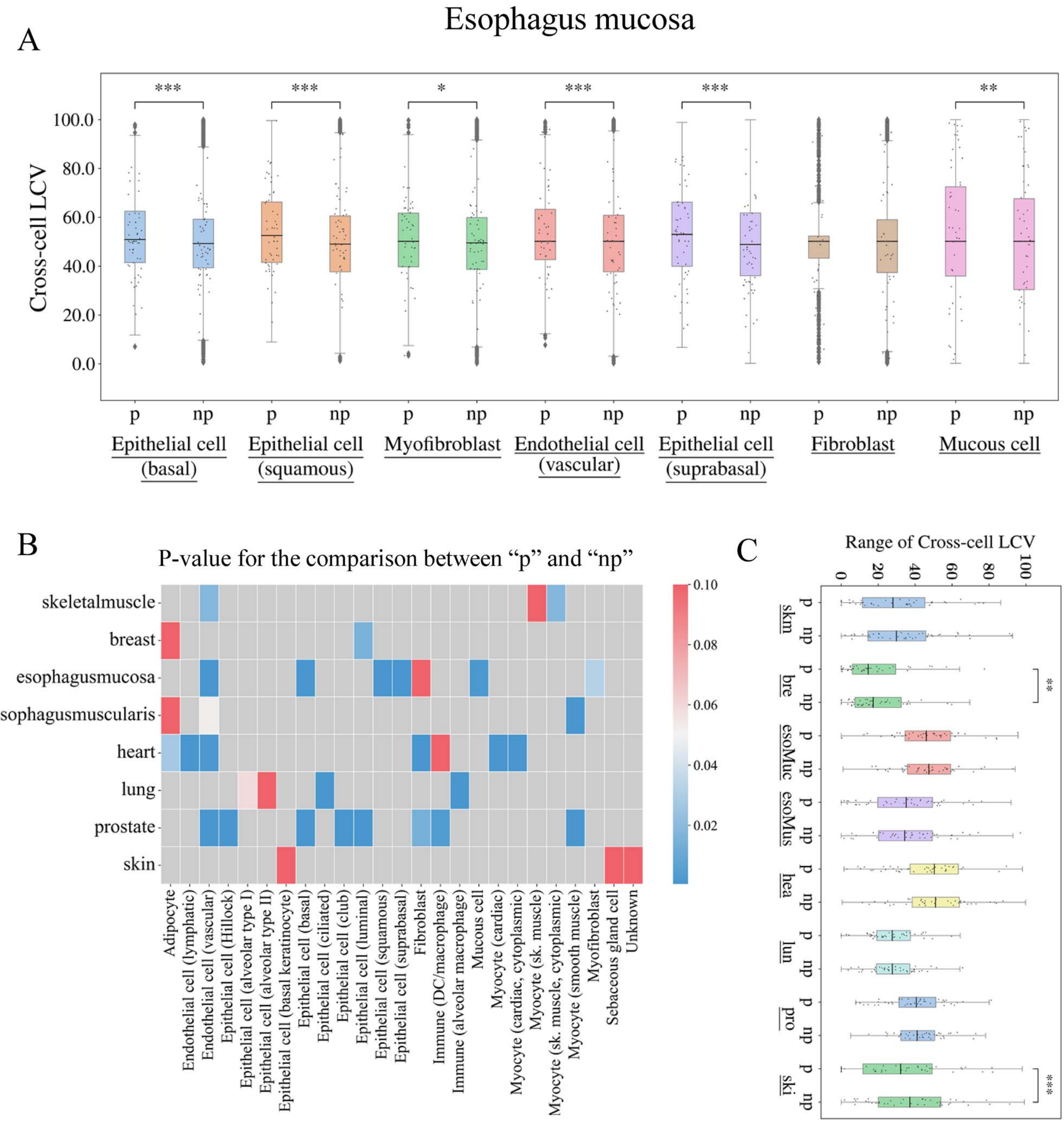

Figure 2. Expression variability of pharmacogenes ("p") and non-pharmacogenes ("np") at the cellular level. (A) Comparison of pharmacogenes and non-pharmacogenes in different cell types of the esophagus mucosa tissue. (B) Comparing pharmacogenes and non-pharmacogenes across all cell types in the eight considered tissues. *P*-values from T-tests are shown in the heatmap. Cell types not present for a specific tissue in our dataset are shaded in grey. (C) Comparing the range of LCV across different cell types of the same tissue. T-tests are performed between the ranges of pharmacogenes and those of non-pharmacogenes. ***: *P*-value < 0.001, **: *P*-value < 0.01, *: *P*-value < 0.05.

Drug efficacy is negatively correlated with the variability LCV of pharmacogenes

Previous research conducted by Simonovsky et al. [15] highlighted a negative correlation between the variability of pharmacogenes across individuals (i.e. cross-individual variability) and drug efficacy using bulk RNA-seq data. In our study, we aimed to delve deeper into this correlation for cross-cell variability by focusing on cell-level LCV obtained from single-nucleus RNA sequencing (snRNA-seq).

To investigate the relationship between pharmacogenes' cross-cell variability and drug efficacy, we computed the weighted average LCV of each drug's target genes (details in Materials and methods). Specifically, we extracted a drug's target genes using the DGIdb database [22]. For each pharmacogene gene, we calculated the LCV across all cell types of a tissue and for each donor individually, then averaged these donor-specific cell-type level LCVs. The maximum averaged LCV across different cell types of a particular tissue was chosen as its representative value for subsequent analyses. Next, we used the interaction score for each gene-drug pair as the weight to compute the weighted average LCV. Similarly, we calculated the cross-individual LCV of pharmacogenes based on GTEx tissue bulk RNA-seq data and

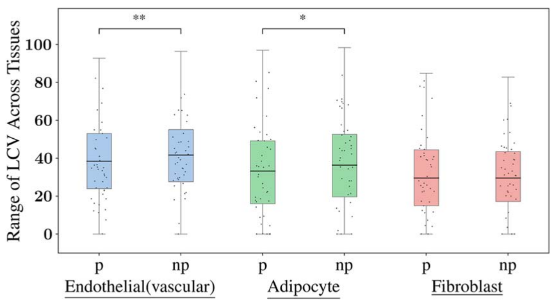

Figure 3. Comparing the range of LCV across different tissue types of the same cell. T-tests are performed to compare pharmacogenes (“p”) and non-pharmacogenes (“np”). ***: *P*-value < 0.001, **: *P*-value < 0.01, *: *P*-value < 0.05.

obtained the weighted average for each gene-drug pair. Consequently, for every drug and every considered tissue, there exists a corresponding cross-cell pharmacogene LCV (C_k) and a cross-individual pharmacogene LCV (I_k).

The analysis was conducted separately for each tissue. Our findings reveal a consistent negative correlation between drug efficacy and the LCV of pharmacogenes. As demonstrated in Fig. 5A, six out of eight tissues exhibit a negative correlation between drug efficacy and cross-cell pharmacogene LCVs. Meanwhile, a negative correlation was observed for all considered seven tissues between drug efficacy and cross-individual LCVs (Fig. 5B). In essence, drugs targeting genes with higher LCV values tend to exhibit lower efficacy. The statistical significance of this negative correlation was confirmed for the cross-cell variability in the skeletal muscle and lung tissue (correlation = −0.149 or −0.256, $P = 0.04$ or 0.002, one-tailed Spearman’s tests) and for the cross-individual variability in the esophagus and heart tissues (correlation = −0.182 or −0.287, $P = 0.0005$ or 0.02, one-tailed Spearman’s tests). These results underscore the significance of gene variability in understanding drug efficacy at both the individual and population levels across various tissues.

Enhanced drug efficacy prediction through joint consideration of cross-cell and cross-individual LCV

As drug efficacy is influenced by both cross-cell LCV and cross-individual LCV, we embarked on an exploration to assess whether the combination of these two variables could enhance our predictive capabilities. To accomplish this, we first formulated multiple linear regression models utilizing various combinations of LCV features (see Materials and methods for details). We selected linear regression for its straightforward interpretability of coefficients and simplicity. Given our relatively small sample size of drugs, linear regression is less likely to overfit compared to more complex models.

When we exclusively utilized tissue-level cross-individual LCV features to predict drug efficacy (model 1), the resulting adjusted R-squared value was a mere 0.043. Alternatively, focusing solely on tissue-level cross-cell LCV features (model 2) yielded a somewhat improved adjusted R-squared of 0.074. However, it was when we jointly considered both cross-individual and cross-cell LCV features (model 3) that our model demonstrated substantial improvement, achieving an adjusted R-squared of 0.121. Notably, several predictors ($P < 0.05$) emerged as significant contributors to this enhanced prediction, encompassing cross-individual LCV features in the esophagus ($P = 0.02$) and heart ($P = 0.007$) tissues, along with the cross-cell LCV feature in the lung ($P = 0.003$).

Furthermore, we explored an approach that integrates tissue cell-type-level LCVs (T_k, computed across cells belonging to the same cell type within a specific tissue) with cross-individual LCVs (model 4). Remarkably, this approach exhibited superior predictive power, resulting in an adjusted R-squared of 0.214. Among the significant predictors ($P < 0.05$) were cross-cell LCV features for breast epithelial cells (luminal), prostate epithelial cells (Hillock), prostate fibroblasts, and lung epithelial cells (alveolar type II). None of the cross-individual LCV terms stood out.

To further investigate the relationship between drug efficacy and the LCVs in target tissues, we focused on two distinct sets of drugs targeting the heart (including Amiodarone, Digoxin, Diltiazem, Disopyramide, Dofetilide, Dronedarone, Flecainide, Lidocaine, Propafenone, and Sotalol) and lung (including Aminophylline, Arformoterol, Montelukast, Pseudoephedrine, Salbutamol (Albuterol), Theophylline, Tiotropium, Zafirlukast, Zileuton, and Levofloxacin), respectively. For each drug set, we developed simple linear regression models using tissue-level cross-cell LCV (C_k) and cross-individual LCV (I_k) calculated from each tissue. For both drug sets, the models trained on the LCVs of the target tissues consistently ranked among the top three performers. Specifically, for heart-targeting drugs, the top three models performed best with features from the heart (adjusted $R^2 = 0.379$), breast (adjusted $R^2 = 0.261$), and skin (adjusted $R^2 = 0.143$). Similarly, for lung-targeting drugs, the top three models performed best with features from the breast (adjusted $R^2 = 0.229$), heart (adjusted $R^2 = 0.211$), and lung (adjusted $R^2 = 0.117$).

To validate our findings, we repeated the same analysis by randomly selecting the same number of drugs as in the tissue-specific drug sets. We repeated this process 1000 times and calculated the mean adjusted R^2. Notably, the results from the random selection demonstrated significantly lower performance, with a mean adjusted R^2 of 0.094 for models trained on heart features and 0.064 for models trained on lung features.

Linear regression assumes a linear relationship between LCV features and drug efficacy. This assumption may oversimplify complex biological relationships, potentially leading to an incomplete representation of underlying patterns in the data. To address this, we subsequently employed a random forest machine learning model, utilizing the cell-type-level LCVs ($\mathbf{T_k}$) and cross-individual LCVs ($\mathbf{I_k}$) from model 4, to explore the potential nonlinear relationship between pharmacogene expression variability and drug efficacy. Fig. 6A presents scatter plots that illustrate the relationships between drug relative efficacy and the top five features based on the highest node purity (i.e. how well a node separates samples of the same class from those of different classes). These include four cross-cell LCV features for heart endothelial cells (vascular), esophagus mucosa fibroblasts, lung epithelial cells (alveolar type I), lung epithelial cells (alveolar type II), and cross-individual LCV for the heart. A LOWESS line was incorporated to more accurately capture and illustrate the underlying negative trends in the data. Complementing this, Fig. 6B presents an incMSE (‘increase in mean squared error’) that measures the improvement in prediction accuracy achieved by a feature) plot, highlighting the relative importance of the top three features: all are cross-cell LCV features for esophagus mucosa fibroblasts, lung epithelial cells (alveolar type I and II).

Discussion

Individual responses to drug treatments are intricately tied to the variability in gene expression, especially within

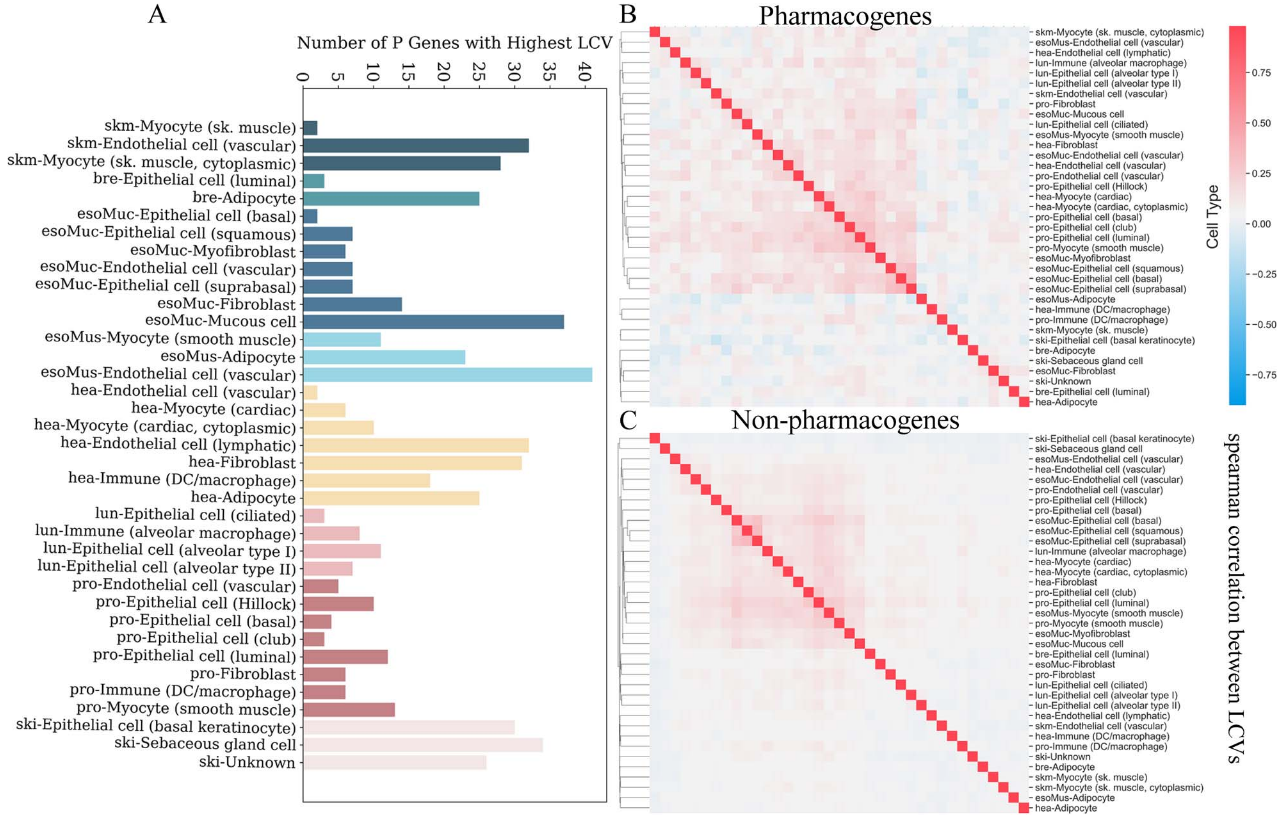

Figure 4. Cell-type expression variation of pharmacogenes. (A) Cell types exhibiting the maximum LCV of a pharmacogene. (B) Spearman correlations between LCV values in different cell types for pharmacogenes. (C) Spearman correlations between LCV values in different cell types for non-pharmacogenes. Tissue abbreviations: Skeletal muscle (skm), breast (bre), esophagus mucosa (esoMuc), esophagus muscularis (esoMus), heart (hea), lung (lun), prostate (pro), and skin (ski).

pharmacogenes, which play crucial roles in drug responses. Our study utilized single-cell RNA sequencing (scRNA-seq) data to delve into the expression variability of pharmacogenes across various cell types in eight human tissues. scRNA-seq allows for the capture of expression patterns at the individual cell level, enabling the identification of cell-type-specific gene expression that bulk RNA-seq averages out. This detailed resolution is crucial for understanding the heterogeneity within tissues and the specific roles of different cell types in biological processes. Our findings not only confirm the well-established link between pharmacogene expression variability and drug efficacy but also offer insights into how the cellular-level variability can be leveraged for improved predictions.

The discrepancy in LCV between the GTEx bulk RNA-seq data and scRNA-seq data is notable, with the former showing greater variability, indicated by a larger interquartile range (Fig. 1B vs. Figure 1A). This difference can be attributed to several factors. Firstly, the GTEx bulk dataset includes samples from a diverse population of approximately 1000 donors, reflecting a broad spectrum of genetic backgrounds that likely contribute to increased gene expression variability. Secondly, bulk RNA-seq captures gene expression across all cell types within a tissue. Tissue heterogeneity can introduce additional variability. In contrast, scRNA-seq provides cell-type-specific data, enabling independent calculation of LCV for each cell type. Aggregating these cell-type-specific LCVs yields a more precise measurement of overall tissue LCV.

We observed significant expression variability among pharmacogenes, both between different individuals and between different cells of the same individual. Pharmacogenes consistently exhibited higher variability compared to non-pharmacogenes, a trend that was evident across various tissues. This aligns with previous findings that pharmacogenes often display increased variability in expression, contributing to the observed diversity in drug responses among individuals. At the cross-individual level, genetic differences among patients can lead to varying expression levels of the same pharmacogene, resulting in different drug responses. For instance, patients with higher expression of certain key pharmacogenes may metabolize or react to drugs differently compared to those with lower expression levels. At the cross-cell level, our findings indicate that even within a single individual, different cell types can show significant variability in pharmacogene expression. This suggests that a drug's effectiveness could be influenced by the specific cellular composition of the targeted tissue. In patients with different disease states, changes in cellular composition could further contribute to variability in drug responses.

Additionally, we found that the variability in gene expression among different cell types is closely linked to their specialized functions. For instance, epithelial and endothelial cells exhibit high gene variability due to their pivotal roles in drug transport. Epithelial cells, lining organ surfaces, play critical roles in the absorption and excretion of various substances, including drugs [23]. They must dynamically regulate gene expression to manage the influx, processing, and efflux of a wide array of compounds. Similarly, endothelial cells are central to the circulatory system's transport functions, including nutrient and oxygen delivery, waste removal, and immune surveillance [24]. Serving as the vital interface between the bloodstream and tissues, endothelial cells must

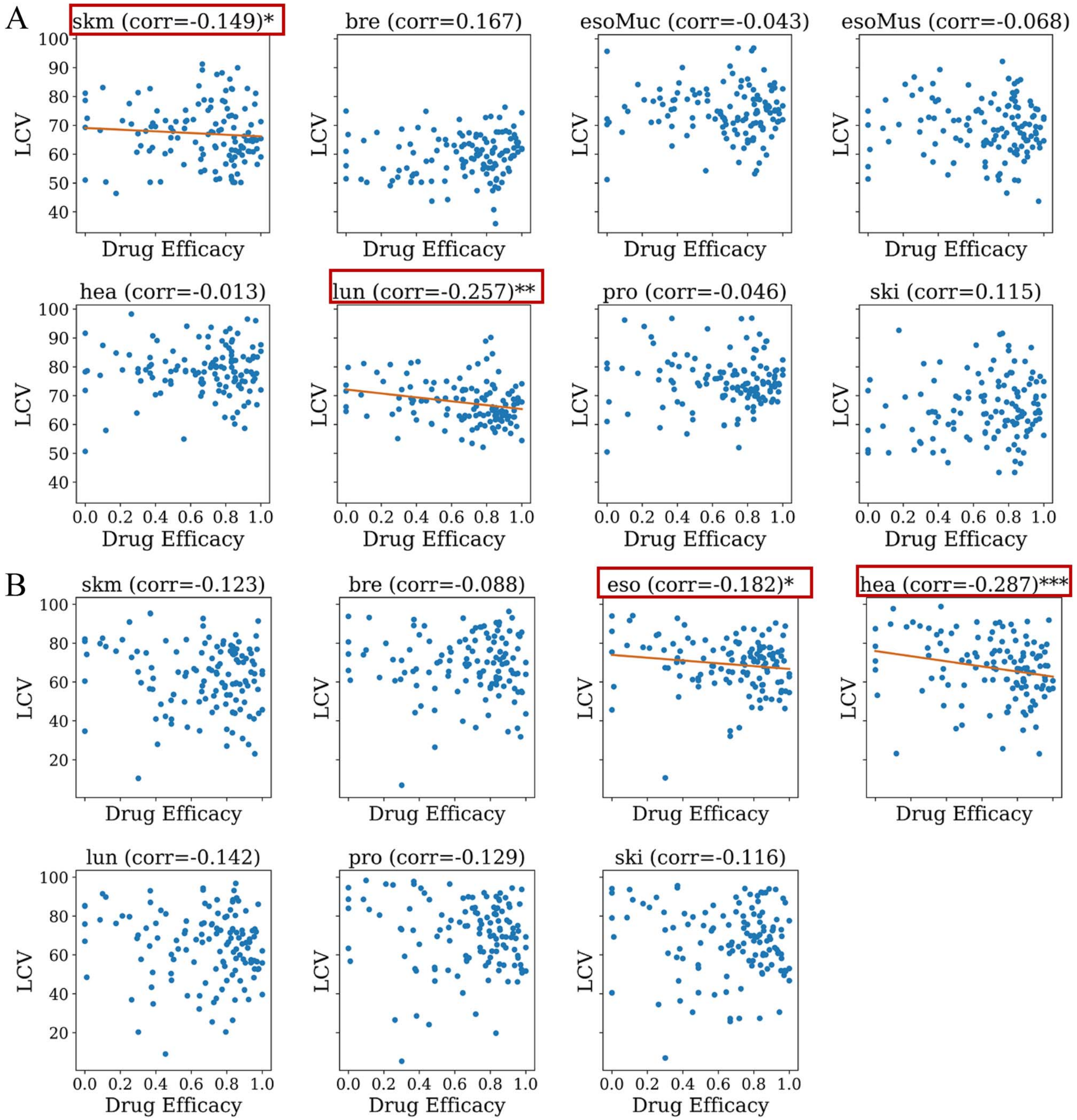

Figure 5. Negative correlation between drug relative efficacy and LCV of pharmacogenes. (A) Correlation based on individual-level cross-cell LCVs. (B) Correlation based on population-level cross-individual LCVs. One-sided spearman tests. ***: P-value < 0.001, **: P-value < 0.01, *: P-value < 0.05. Tissue abbreviations: Skeletal muscle (skm), breast (bre), esophagus mucosa (esoMuc), esophagus muscularis (esoMus), heart (hea), lung (lun), prostate (pro), and skin (ski).

adapt to diverse conditions and demands, requiring flexible gene expression patterns.

To further validate our findings, we performed randomized tests for all comparisons between pharmacogenes and non-pharmacogenes. In each test, we randomly selected an equal number of non-pharmacogenes to compare with pharmacogenes. We then repeated this process 10 000 times to evaluate the significance of the average p-value. Notably, the results were similar to those obtained without the randomized tests. Detailed plots illustrating these findings can be found in Supplementary Figs 5–9.

Our analysis unveiled a negative correlation between the variability of pharmacogenes and drug efficacy, both at the cross-cell and cross-individual levels. Drugs targeting genes with higher expression variability tended to exhibit reduced efficacy, highlighting the importance of considering gene expression heterogeneity when designing and predicting drug responses. This correlation was observed across multiple tissues, emphasizing the broad impact of pharmacogene variability on drug outcomes.

To enhance our understanding and predictive capabilities, we developed regression and machine learning models that

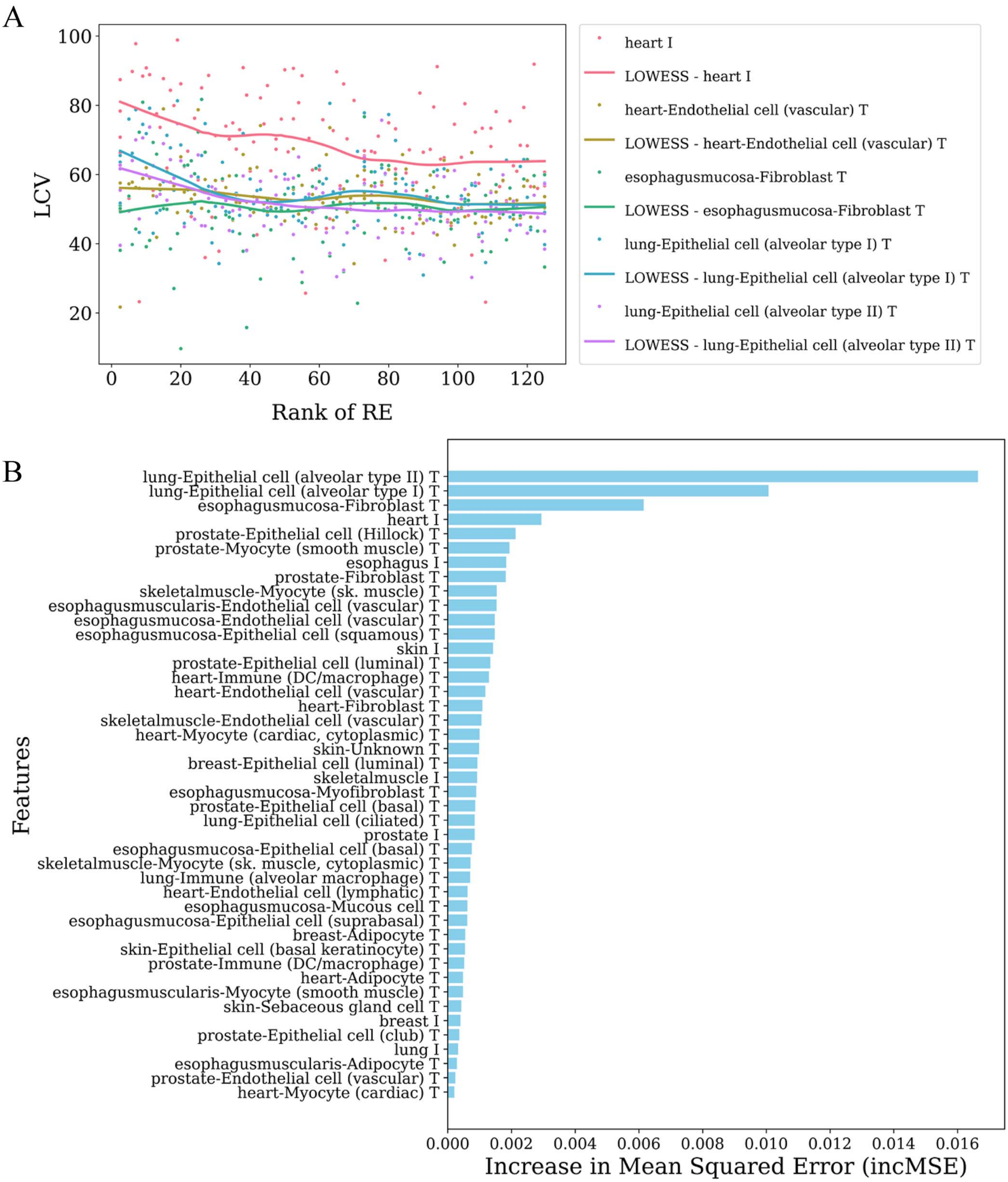

Figure 6. Top LCV features in drug efficacy prediction via a random forest model. (A) Relationship between the rank (increasing order) of drug relative efficacy (RE) and top 5 LCV features based on the node purity from the random forest analysis. LOWESS smoothing lines have been superimposed. I: Cross-individual LCVs. T: Cross-cell LCVs. (B) Importance ranking of all LCV features based on incMSE.

integrated cross-cell and cross-individual pharmacogene expression variability. These models showed promising results, particularly when combining both types of variability. Notably, the joint consideration of cross-cell and cross-individual LCV features yielded a substantial improvement in predicting drug efficacy. This suggests that a comprehensive approach, encompassing variability at both the cellular and individual levels, can provide valuable insights into drug performance. Notably, our analysis identified the cross-cell LCV features, especially those in the lung, as dominant predictors.

We trained linear regression models on tissue-specific drug sets to demonstrate that LCVs in the target tissue are predictive of drug efficacy. Models trained on features from the target tissue consistently ranked among the top three performers. However, it should be noted that they were not always the best models. This variability can be attributed to well-known off-target drug effects [25] and the inherent bias introduced by our small sample size of drugs.

Our findings suggest that incorporating single-cell gene expression variability into the early stages of drug development could enhance the design process. By identifying pharmacogenes with high variability across different cell types and tissues, researchers can pinpoint potential targets likely to produce variable patient responses. This approach could guide the development of drugs that account for such variability, leading to more consistent and effective treatments. Additionally, considering both cross-individual and cross-cell variability could improve predictions of drug efficacy and safety, ultimately supporting the creation of more personalized, context-specific therapies.

In summary, our research underscores the complexity of gene expression variability in pharmacogenes and its profound impact on drug efficacy. By elucidating these variability patterns at both cellular and tissue levels, we move closer to the era of personalized medicine. Understanding how individual genetic differences manifest in drug responses allows for more tailored and effective treatment strategies.

While our study provides valuable insights, several limitations should be acknowledged. The reliance on single-cell RNA sequencing data, while offering high resolution, may not capture the entirety of gene expression variability in complex tissues. Additionally, the dataset's focus on healthy tissues limits the extrapolation to disease contexts where drug responses may differ. Future studies could explore how these variability patterns translate into clinical settings and consider a broader range of tissues and disease conditions.

The scRNA-seq data and GTEx bulk RNA-seq data originate from different sample populations, which may introduce bias in comparisons. Specifically, the scRNA-seq data include only 16 donors. The LCV calculated from cells of this relatively small number of individuals may not generalize well to larger populations. Future studies should include a larger and more diverse set of donors to enhance the generalizability and robustness of the findings.

In addition to the limitations of scRNA-seq data mentioned earlier, further insights could be gained by more detailed examination of tissue-specific factors and exploring cross-cell LCVs within specific tissues relevant to drug targeting. Considering additional covariates, such as the genetic diversity of pharmacogenes, may offer a clearer understanding of drug efficacy.

Furthermore, the drug efficacy score in our analysis is derived from adverse event reports in the FDA Adverse Event Reporting System. However, the interpretation and reporting of adverse events can vary significantly among patients and healthcare providers, introducing variability that may affect the accuracy of the relative efficacy quantification. Future studies could mitigate this issue by adopting a more comprehensive approach to calculating drug efficacy, such as integrating multiple data sources to enhance the robustness and reliability of the measurements.

In conclusion, our study contributes to the growing body of evidence supporting the importance of gene expression variability, particularly in pharmacogenes, for understanding and predicting drug responses. By integrating cross-cell and cross-individual variability measurements, we provide a framework for more precise drug efficacy predictions. This work lays the foundation for further investigations into the complicated relationships between gene expression, cellular heterogeneity, and drug outcomes, ultimately advancing the field of precision medicine.

Materials and methods

Gene expression data from normal human tissue samples

Single-cell RNA-seq, more precisely single-nucleus RNA-seq (snRNA-Seq) data here [20] were obtained from the Genotype-Tissue Expression (GTEx) [21] V9 release (https://gtexportal.org/home/). snRNA-seq uses isolated nuclei instead of whole cells to profile gene expression. The data were collected from non-disease samples of sixteen donors and eight tissues (skeletal muscle, breast, esophagus mucosa, esophagus muscularis, heart, lung, prostate, and skin). A total of 15 944 cells were investigated. Raw read counts were normalized by GTEx using CP10k (copy per 10 k transcripts). We filtered genes expressed in less than 50 cells and removed cells with less than 1650 genes. Because snRNA-Seq data contain a large number of zero values, we also removed genes with mean expression lower than the 10th quantile of the means. In addition, bulk RNA-seq data for seven of these tissues (esophagus in the bulk data corresponds to esophagus mucosa and esophagus muscularis in the snRNA-seq data) were downloaded from the GTEx Analysis V8 release. Raw read counts were normalized using TPM (transcripts per million) by GTEx. Similarly, samples and genes with low quality were filtered according to GTEx analysis procedures.

Expression variability calculation

The inflated zero expression values in snRNA-Seq data result in a biased measure of expression variability when applying the coefficient of variation (CV) directly. Therefore, we adopted the local coefficient of variation (LCV) algorithm [15] to estimate the expression variability. This algorithm uses a ranking approach based on a sliding window, which has been validated as the least biased towards lowly expressed genes and the most robust to data incompleteness compared to other variability measures [15], including standard deviation (SD), mean absolute deviation (MAD), coefficient of variation (CV), dispersion measure (DM), and entropy variance (EV). Here, we used a 500-gene window. The LCV values range from 0 to 100. A larger LCV represents higher expression variability.

Selection of pharmacogenes

A list of 389 pharmacogenes, referred to as "PGRN pharmacogenes," was obtained from Chhibber *et al.* [26]. These genes were identified from various resources and publications related to drug responses, including PharmGKB [27], PharmaADME [28], and FDA Pharmacogenomics Biomarkers [29]. We then compared the expression variability of pharmacogenes with that of the remaining non-pharmacogenes profiled in GTEx. To extend our list of pharmacogenes for the drug efficacy study, we incorporated 312 additional genes from the DGIdb database [22] (https://www.dgidb.org/). Such additional selection focused on genes interacting with more than two drugs.

Drug-gene interaction score and drug relative efficacy

Drug-gene interactions were downloaded from DGIdb [22]. This database presents an interaction score between a drug d and a target gene g as:

$$IS_{d,g} = \frac{\text{\#of average known gene partners of all drugs}}{\text{\#of known gene partners of drug } d} \times \frac{\text{\#of average known drug partners of all genes}}{\text{\#of known drug partners of gene } g} \times \text{evidence score}$$

This drug-gene interaction score, treated at the logarithmic scale, serves as the weight for computing the overall LCV across all n target genes of the same drug d: $LCV_d = \frac{\sum_{g=1}^{n} \log(IS_{d,g}) \times LCV_g}{\sum_{g=1}^{n} \log(IS_{d,g})}$. Note that LCV_g can be cross-cell or cross-individual LCV values for gene g.

Furthermore, the relative efficacy (RE) scores for drug-disease pairs were obtained from Guney et al. [30]. The RE scores were computed using text-mining methods on reports submitted to the FDA's Adverse Event Reporting System (FAERS, https://open.fda.gov/data/faers/) and comparing the number of ineffective reports with the number of reports stating the most common complaints. RE has a range from 0 to 1, and a higher RE score indicates that a drug is more effective in treating the disease. A total of 129 drugs were considered in our study.

Computational models to predict drug relative efficacy

To predict drug relative efficacy (RE), we devised multiple regression models leveraging different combinations of LCV values for their corresponding pharmacogenes.

1) Cross-individual LCV Model: This model relies exclusively on tissue-level cross-individual LCV features. $RE = \beta_0 + \sum_{\mathbf{k}=1}^{7} \beta_k I_k$. Here, the combined LCV value of a drug for each tissue k (I_k) was calculated as the weighted average of LCV values of all pharmacogenes associated with that drug according to the formula above. For each pharmacogene, the LCV was calculated across multiple individual samples of tissue k.
2) Cross-cell LCV Model: This model exclusively employs the tissue-level cross-cell LCV features. $RE = \beta_0 + \sum_{\mathbf{k}=1}^{8} \beta_k \mathrm{C}_\mathbf{k}$. For each pharmacogene, the LCV was calculated across cells of the same cell type within tissue k and then averaged across individuals. To further obtain a tissue-level measurement, we employed three different methods to aggregate LCV values across different cell types within that tissue: maximum, mean, and median. The corresponding adjusted R^2 values obtained from these methods were 0.074, 0.050, and 0.051, respectively. As a result, we chose the maximum LCV among different cell types within tissue k (C_k) for the drug efficacy prediction.

Joint LCV Model: This model jointly considers both the cross-individual and cross-cell LCV features. $RE = \beta_0 + \sum_{\mathbf{k}=1}^{7} \beta_k I_k + \sum_{\mathbf{j}=1}^{8} \beta_{7+j} \mathrm{C}_j$.Comprehensive joint LCV Model: This model integrates tissue-level cross-individual LCVs with cell-type level LCVs: $RE = \beta_0 + \sum_{\mathbf{k}=1}^{7} \beta_k I_k + \sum_{j=1}^{37} \beta_{7+j} T_j$. In this case, the LCV for each pharmacogene was calculated across cells of the same cell type within a tissue (a total of 37 cell-type and tissue combinations) and then averaged across individuals. The weighted average across all pharmacogenes for a drug is denoted as T_j.The above four regression models provide a comprehensive framework for predicting drug relative efficacy by considering various combinations of LCV features, encompassing both individual and cell-level variability.

Moreover, to capture the potential non-linear relationship between expression variability and drug efficacy, we applied a random forest model using the cell-type-level LCVs (T_j's) and cross-individual LCVs (I_k's) identified in Model 4. We ranked the impact of various LCV features on drug efficacy based on node purity and "increase in Mean Squared Error" (incMSE). Node purity in random forest models refers to the homogeneity of the samples within each node of the decision trees comprising the forest. It measures how well a node separates samples of the same class from those of different classes. Higher purity indicates that the majority of samples within a node belong to the same class, resulting in clearer decision boundaries. "Increase in mean squared error" is a criterion used by random forest models to evaluate the effectiveness of splitting a node. It quantifies the reduction in overall variance that occurs when a node is split based on a particular feature. A larger increase in mean squared error suggests that splitting the node based on that feature results in greater improvement in prediction accuracy.

Acknowledgments

We gratefully acknowledge the valuable discussions with other members of Chen's group. Additionally, we appreciate the helpful suggestions provided by the reviewers.

Supplementary data

Supplementary data is available at *HMG Journal* online.

Conflict of interest statement: The authors declare that they have no competing interests.

Funding

This work was supported by the National Institutes of Health [grant numbers R01GM137428, R01NS125276].

Data availability

Related programming codes and data can be found at https://github.com/Steven51516/LCV.

References

1. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. *Nature* 2015;**526**:336–342.
2. Ashley EA. Towards precision medicine. *Nat Rev Genet* 2016;**17**: 507–522.
3. Roses AD. Pharmacogenetics and the practice of medicine. *Nature* 2000;**405**:857–865.
4. Meyer UA. Pharmacogenetics and adverse drug reactions. *Lancet* 2000;**356**:1667–1671.
5. Oates JT, Lopez D. Pharmacogenetics: an important part of drug development with a focus on its application. *Int J Biomed Investig* 2018;**1**:111.
6. Orrico KB. Basic concepts in genetics and pharmacogenomics for pharmacists. *Drug Target Insights* 2019;**13**:1177392819886875.
7. Taylor C, Crosby I, Yip V. *et al.* A review of the important role of CYP2D6 in pharmacogenomics. *Genes (Basel)* 2020;**11**:1295.

8. Bertilsson L, Dahl ML, Dalen P. *et al.* Molecular genetics of CYP2D6: clinical relevance with focus on psychotropic drugs. *Br J Clin Pharmacol* 2002;**53**:111–122.
9. Seelig A. P-glycoprotein: one mechanism, many tasks and the consequences for pharmacotherapy of cancers. *Front Oncol* 2020;**10**:576559.
10. Hodges LM, Markova SM, Chinn LW. *et al.* Very important pharmacogene summary: ABCB1 (MDR1, P-glycoprotein). *Pharmacogenet Genomics* 2011;**21**:152–161.
11. Daly AK. Pharmacogenetics and human genetic polymorphisms. *Biochem J* 2010;**429**:435–449.
12. Smith RP, Lam ET, Markova S. *et al.* Pharmacogene regulatory elements: from discovery to applications. *Genome Med* 2012; **4**:45.
13. Qiu W, Rogers AJ, Damask A. *et al.* Pharmacogenomics: novel loci identification via integrating gene differential analysis and eQTL analysis. *Hum Mol Genet* 2014;**23**:5017–5024.
14. Yang HC, Lin CW, Chen CW. *et al.* Applying genome-wide gene-based expression quantitative trait locus mapping to study population ancestry and pharmacogenetics. *BMC Genomics* 2014;**15**:319.
15. Simonovsky E, Schuster R, Yeger-Lotem E. Large-scale analysis of human gene expression variability associates highly variable drug targets with lower drug effectiveness and safety. *Bioinformatics* 2019;**35**:3028–3037.
16. Saliba AE, Westermann AJ, Gorski SA. *et al.* Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* 2014;**42**: 8845–8860.
17. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* 2018;**13**: 599–604.
18. Chen L, Zheng S. BCseq: accurate single cell RNA-seq quantification with bias correction. *Nucleic Acids Res* 2018; **46**:e82.
19. Yin Q, Chen L. CellTICS: an explainable neural network for cell-type identification and interpretation based on single-cell RNA-seq data. *Brief Bioinform* 2023;**25**:1–12.
20. Eraslan G, Drokhlyansky E, Anand S. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* 2022;**376**:eabl4290.
21. Consortium GT. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015;**348**:648–660.
22. Freshour SL, Kiwala S, Cotto KC. *et al.* Integration of the drug-gene interaction database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res* 2021;**49**:D1144–D1151.
23. Schwenk M. Drug transport in intestine, liver and kidney. *Arch Toxicol* 1987;**60**:37–42.
24. Glassman PM, Myerson JW, Ferguson LT. *et al.* Targeting drug delivery in the vascular system: focus on endothelium. *Adv Drug Deliv Rev* 2020;**157**:96–117.
25. Huang Y, Furuno M, Arakawa T. *et al.* A framework for identification of on- and off-target transcriptional responses to drug treatment. *Sci Rep* 2019;**9**:17603.
26. Chhibber A, French CE, Yee SW. *et al.* Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines. *Pharmacogenomics J* 2017;**17**:137–145.
27. Whirl-Carrillo M, McDonagh EM, Hebert JM. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012;**92**:414–417.
28. Tang X, Li R, Wu D. *et al.* Development and validation of an ADME-related gene signature for survival, treatment outcome and immune cell infiltration in head and neck squamous cell carcinoma. *Front Immunol* 2022;**13**:905635.
29. Kim JA, Ceccarelli R, Lu CY. Pharmacogenomic biomarkers in US FDA-approved drug labels (2000-2020). *J Pers Med* 2021;**11**:179.
30. Guney E, Menche J, Vidal M. *et al.* Network-based in silico drug efficacy screening. *Nat Commun* 2016;**7**:10331.