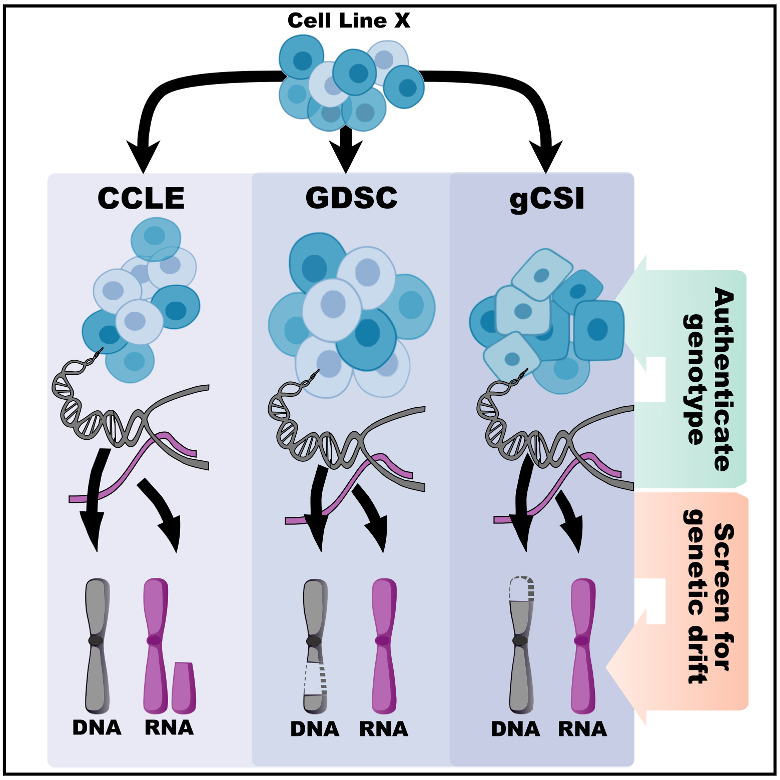
Cell Systems

Report

Assessment of Genetic Drift in Large Pharmacogenomic Studies

Graphical Abstract



Highlights

- Less than 1.6% of cell lines in pharmacogenomic studies have genetic identity issues
- A median of 4.5%–6.1% of the total genome is drifted between homonymous cell lines
- Genetic drift is present within each dataset between different data types
- Small association between total genetic drift and discordances in drug response

Authors

Rene Quevedo, Petr Smirnov,
Denis Tkachuk, ..., Zhaleh Safikhani,
Trevor J. Pugh, Benjamin Haibe-Kains

Correspondence

trevor.pugh@uhnresearch.ca (T.J.P.),
bhaibeka@uhnresearch.ca (B.H.-K.)

In Brief

Our study explores the genetic identity and stability of 1,497 cell lines used across the three largest pharmacogenomic studies published to date. We find that there are extensive shifts in the genetic profiles of cell lines between different assays of the same study and between the same assays of different studies. As genetic drift in cell lines is known to have a multitude of functional and phenotypic consequences, these findings act as a cautionary tale that extends beyond the scope of pharmacogenomic studies. We highlight the need for guidelines and resources that leverage existing data to screen and maintain the genetic stability of cell lines throughout a study.

Quevedo et al., 2020, Cell Systems *11*, 393–401
October 21, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cels.2020.08.012

CellPress

Cell Systems

CellPress

Report

Assessment of Genetic Drift in Large Pharmacogenomic Studies

Rene Quevedo,[1,2] Petr Smirnov,[1,2] Denis Tkachuk,[1] Chantal Ho,[1] Nehme El-Hachem,[3] Zhaleh Safikhani,[1,2] Trevor J. Pugh,[1,2,4,*] and Benjamin Haibe-Kains[1,2,4,5,6,7,*]
[1]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada
[2]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
[3]Integrative System Biology, Institut de Recherches Cliniques de Montréal (IRCM), Montréal, QC, Canada
[4]Ontario Institute for Cancer Research, Toronto, ON, Canada
[5]Department of Computer Science, University of Toronto, Toronto, ON, Canada
[6]Vector Institute, Toronto, ON, Canada
[7]Lead Contact
*Correspondence: trevor.pugh@uhnresearch.ca (T.J.P.), bhaibeka@uhnresearch.ca (B.H.-K.)
https://doi.org/10.1016/j.cels.2020.08.012

SUMMARY

Genomic instability affects the reproducibility of experiments that rely on cancer cell lines. However, measuring the genomic integrity of these cells throughout a study is a costly endeavor that is commonly forgone. Here, we validate the identity of cancer cell lines in three pharmacogenomic studies and screen for genetic drift within and between datasets. Using SNP data from these datasets encompassing 1,497 unique cell lines and 63 unique pharmacological compounds, we show that genetic drift is widely prevalent in almost all cell lines with a median of 4.5%–6.1% of the total genome size drifted between any two isogenic cell lines. This study highlights the need for molecular profiling of cell lines to minimize the effects of passaging or misidentification in biomedical studies. We developed the CCLid web application, available at www.cclid.ca, to allow users to screen the genomic profiles of their cell lines against these datasets. A record of this paper's transparent peer review process is included in the Supplemental Information.

INTRODUCTION

Cell lines are adequate model systems when exploring the link between genetic variation and pharmacological response; however, their usage in biomedical research is plagued by issues of misidentification and genetic instability (Freedman et al., 2015; Korch and Varella-Garcia, 2018; Vaughan et al., 2017). Reviews by Korch and Varella-Garcia (Korch and Varella-Garcia, 2018) illustrate the prevalence of these issues and present guidelines to ensure cell line authentication and integrity throughout a study. Common issues regarding ambiguous cell line annotations and their usage in biomedical research can be addressed by referencing the Cellosaurus knowledge resource (Bairoch, 2018), while cell line identity can be validated using the cell line authentication using short tandem repeat (CLASTR) web portal (Robin et al., 2020). Despite these resources, short tandem repeats are more susceptible to genetic drift and mismatch repair deficiencies than the use of single-nucleotide polymorphisms (SNPs) (Fan and Chu, 2007; Korch et al., 2012; Much et al., 2014), and there are limited resources available to monitor the stability of cell lines throughout and between studies (Ben-David et al., 2018).

Genetic instability is a hallmark of cancer that is particularly heightened in cancer cell lines, thus, affecting the stability of genomic profiles across and between studies (Ben-David et al., 2018; Kleensang et al., 2016). Ben-David et al. examined the extent of genetic drift between 106 cell lines shared between the Broad Institute (Cancer Cell Line Encyclopedia [CCLE]) and the Sanger Institute (Genomics of Drug Sensitivity in Cancer [GDSC]). In their study, they reported an estimated 19% discordance in non-silent mutations and 26% (range of 7%–99%) discordance in copy-number variants (CNVs). More focused analyses on genetic stability of a single cell line (MCF-7, A549, and HeLa) show that subclonal populations with different genomic backgrounds exist at the cell bank level (Kleensang et al., 2016), and that passaging or inter-laboratory effects can amplify genetic drift in these cell lines (Ben-David et al., 2018; Kleensang et al., 2016; Liu et al., 2019). More concerning is that these instabilities can lead to a plethora of functional consequences such as morphological differences, altered proliferative capacity, a diverse gene expression pattern, variable responses to alleviate proteotoxic stress, and a major impact on drug response (Ben-David et al., 2018; Kleensang et al., 2016; Liu et al., 2019). Ben-David et al. contributed to genetic stability issue by developing the web portal Cell STRAINER (strain instability profiler) (Ben-David et al., 2018), which allows users to compare the copy-number profile of their cell lines against CCLE (Barretina et al., 2012) and flag discordant regions.

Our study expands the work of Ben-David et al. by exploring the genetic identity and stability of 1,497 cell lines used across the three largest pharmacogenomic studies published to date: CCLE/CTRPv2, GDSC, and the Genentech cell line screening initiative (gCSI) (CCLE Consortium and GDSC Consortium, 2015; Haibe-Kains et al., 2013; Hatzis et al., 2014; Haverty et al., 2016; Safikhani et al., 2016). In contrast to the method utilized by Cell STRAINER (Ben-David et al., 2018), we use an SNP allelic-fraction method that allows us to leverage RNA sequencing (RNA-seq) data in addition to SNP array within each dataset to screen for genetic stability both between and within each institution. Furthermore, by comparing measures of chromosomal instability (CIN) and genetic drift to distances between drug dose-response curves, we show that both measures have a significant, albeit weak association with pharmacological response for certain drugs. Thus, our results support the notion that the inherent karyotypic instability of cancer cell lines may have an adverse effect on the reproducibility of pharmacological screens.

RESULTS

Datasets

To quantify genetic instability of human cancer cell lines used in large-scale pharmacogenomic studies, we reprocessed 1,006 publicly available Affymetrix SNP arrays from the GDSC (Yang et al., 2013), 1,213 from the CCLE project (Barretina et al., 2012) as well as 668 Omni 2.5Million SNP arrays from the gCSI (Haverty et al., 2016). Using the paired RNA-seq from CCLE, GDSC, and gCSI we searched all datasets for both inter- and intra-institutional genotype concordance, genetic drift, and pharmacological agreement (Figure S1).

Authentication of Cancer Cell Lines

To ensure the identity of all cancer cell lines used across the three datasets, we used 5,701 SNPs uniformly scattered across the genome to check their genotypic identity. We did pairwise comparisons between the CCLE, GDSC, and gCSI pharmacogenomic datasets. While our predictions found numerous cases of isogenicity in differently annotated cell lines (Figure 1; $n_{CCLE\text{-}GDSC}$ = 68, $n_{CCLE\text{-}Gcsi}$ = 103, and $n_{GDSC\text{-}gCSI}$ = 86), these were largely rectified by cross-referencing the Cellosaurus database (Bairoch, 2018). After this correction, we found that all datasets contained ‘‘problematic cell lines,’’ indicating known contamination or misidentification. We further corrected for cell line pairs that were not directly linked in Cellosaurus (e.g., HCT-15 and DLD-1 share the same parental cell line but are not explicitly annotated as linked) and resulted in only 6 cell lines in CCLE, 3 in GDSC, and 11 in gCSI that were either misidentified (e.g., Hs-571.T in CCLE mapping to TOV-112D in GDSC and gCSI), known problematic with an unreported pair (e.g., cell line KE-97 pairing with KMS-18), have an unknown relation another cell line pair (e.g., HLC-1 and HMC-1-8 having different STR profiles with unknown relation), or simply mislabeled (e.g., DOR13 in gCSI is actually DOV13) (Table S1). Overall, given the scale of these initiatives, there was a commendable amount of concordance between cell lines with very few cases of misidentification or some unknown relation between cell line pairs.

We next ensured the identity of cell lines from RNA-seq data generated from GDSC and CCLE. Of the 462 samples from GDSC, 19 samples had no matching counterpart in any dataset, and 4 matched cell lines of a different annotation or lineage (HAL-01, KPD, CHL-1, and LB771-HNC) (Table S2). Of the 711 samples analyzed from CCLE, 44 did not match any other cell lines, and 6 matched cell lines of different annotation or lineage (NCI-H1819, LN-464, U343, HSSRB, HCC-1588, and PLB-985). Finally, the CCRF-SB was the only cell line out of the 675 tested that matched a cell line of a different annotation or lineage, CTV-1 found in the GDSC dataset. However, the CTV-1 line is labeled as a likely misidentified in Cellosaurus (https://web.expasy.org/cellosaurus/CVCL_1150). The imperfect concordance between genotype and gene expression datasets suggests the absence of DNA/RNA co-isolation and the possibility of genetic drift within individual datasets.

Detection of Genetic Drift

To investigate the potential of intra- and inter-institutional genetic drift in cell lines, we first quantified the amount of karyotypic discordances using log-R ratio (LRR) or B-allele frequency (BAF) alone between SNP-array-processed samples. We estimated the fraction of the genome that drifted between 586 isogenic cell lines between the GDSC and CCLE datasets using LRR (Table S3). We found that most cell lines show evidence of karyotypic drift as measured by the fraction of the genome drifted (k) using LRR (k_{LRR} = 0.09 ± 0.11), with one cell line exhibiting 72.6% of the genome being karyotypically different between datasets (Figure 2A). Using BAF, we found a significantly smaller portion of the genome exhibiting genetic drift (k_{BAF}= 0.02 ± 0.05) with one cell line exhibiting 60% of the genome drifted (Figure 2A).

As a simple observation, we found that the estimates made using BAF were less sensitive than those made using LRR, which were capable of detecting abnormalities derived from subclonality or heterogeneity (Figure S2). While the majority of cases agreed upon regions drifted using LRR and BAF approaches (Figure 2C), there were few cases with a large discrepancy between LRR and BAF estimates (Figure 2A). We identified that drift estimates using LRR were overly sensitive to the presence of subclonality or heterogeneity within a given cell line population (Figure S2) and that BAF estimates of drift were confounded by noise, especially in loss-of-heterozygosity (LOH) regions (Figure 2C). Furthermore, the step-like increase in ploidy of segments does not directly translate to detectable differences in BAF (e.g., AAB → AAAB; Figure 2D) in highly aneuploidy cell lines with high ploidy, such as JHOS-2.

Given these limitations, we sought to quantify how accurate BAF estimates of drift were when compared with LRR estimates. We assessed accuracy by detecting the minimum percent of the genome that is concordant in drifted/non-drifted regions between BAF- and LRR-calculated drift profiles. We found that 90% of all cell line pairs have at least 78.4% of the drift profiles for the genome in concordance, 80% of all cell line pairs have at least 85.3% concordance in drift profiles, and 70% of all cell line pairs have at least 89.4% concordance (Figure 2B). These results highlight that despite the large discordances between approaches seen in cell lines, such as JHOS-2 and HCC1937 (Figures 2D and 2E), these cases represent only a minority of cases, while the majority of drifted/non-drifted regions can be accurately called using BAF alone.

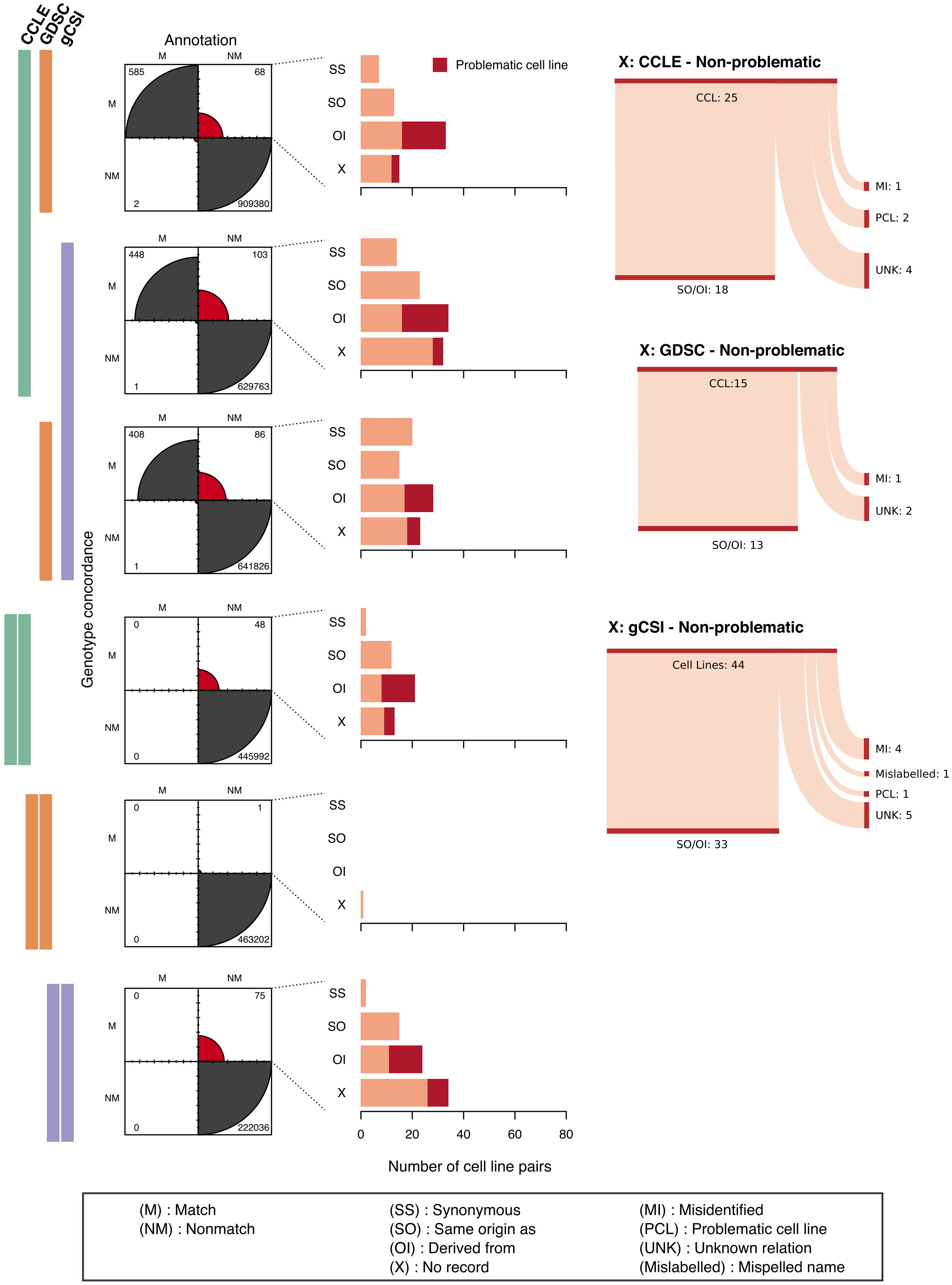


Figure 1. Concordance between Genotypes Using B-Allele Frequency

Confusion matrices for predicted isogenicity (matching; M) and assumed isogenicity based on annotations were computed for each pairwise combination of datasets: the predicted isogenic lines with non-matching (NM) annotations were closely investigated by annotations from the Cellosaurus dataset to look for synonymous cell lines (SS) or cell lines deriving from the same origins (OI/SO). All cell lines with no obvious link (X) were further manually reviewed to look for associations and reported as misidentified (MI) for clear cases of misidentification/contamination or unknown (UNK) where there is not enough metadata to make an informed decision.

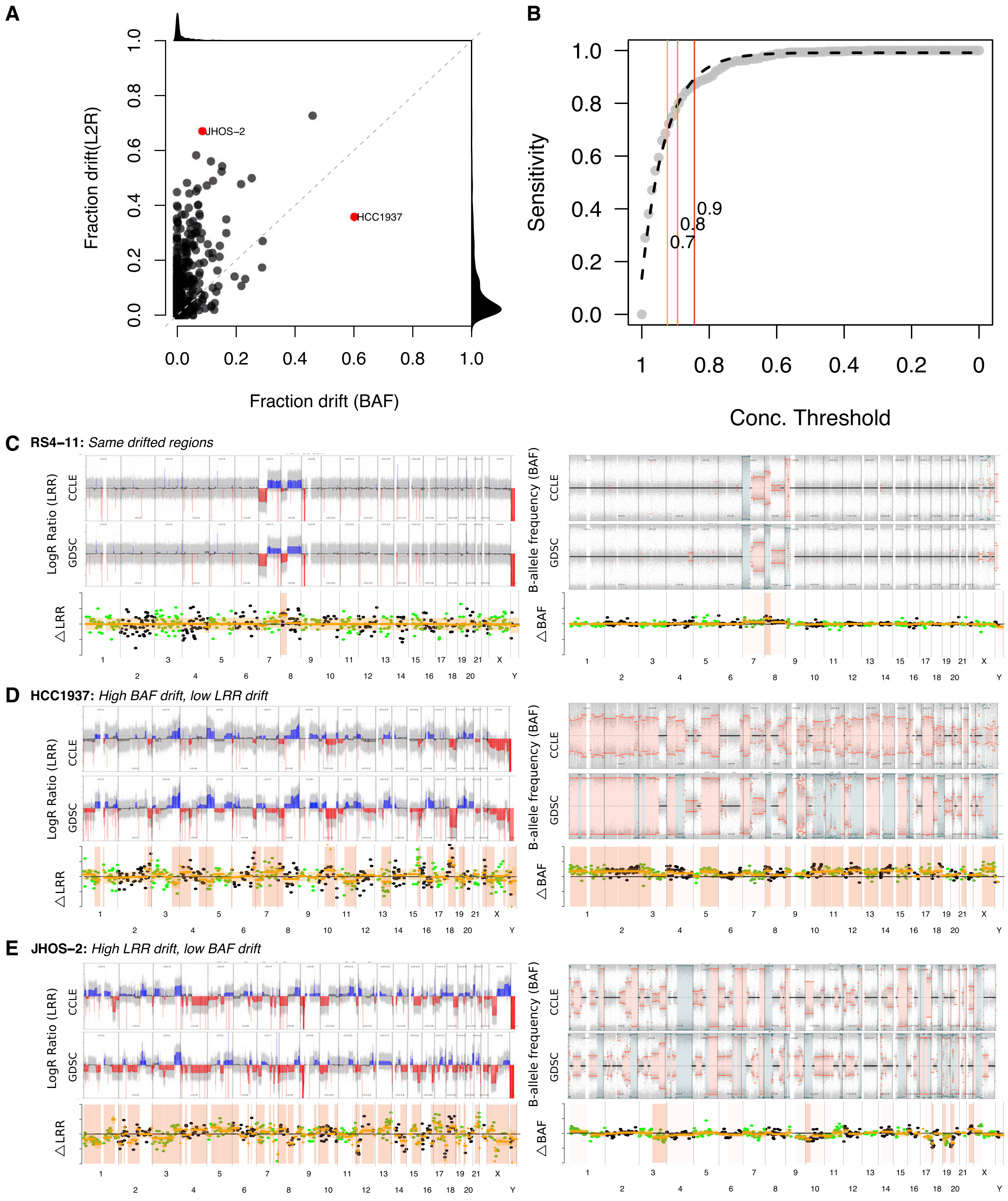


Figure 2. Measures of Genetic Drift between GDSC and CCLE

(A) Estimates of total genomic drift are plotted against each, one derived from the LRR and one from the BAF. Distributions of both estimates are represented as kernel density plots in the margins.

(B) A BAF-drifted saturation curve is fitted using LRR-drifted regions as the ground truth. Three different cutoffs of 90%, 80%, and 70% saturation are represented on the plot.

(C–E) Genetic drift plots generated by CCLid using the Euclidean distance between LRR or BAF of each SNP are represented for a positive control (C) and two problematic cases; one with high LRR and low BAF (D) and one with high BAF and low LRR (E).

We then compared cell lines from gCSI and CCLE (Table S3). Unlike our previous GDSC-CCLE comparison, gCSI was analyzed using a different SNP-array platform; Omni 2.5Million as compared with Affymetrix SNP 6.0. Despite these differences, we saw the same distribution of LRR drift and BAF as observed between GDSC and CCLE (Figures S3A and S3B), suggesting a platform-independent analysis and that similar levels of genetic drift were observed across research institutions. Similarly, we observed poor concordance between BAF- and LRR-estimated drift, with extreme outliers representing different components, such as noise of the assay in Hs-294T (Figure S3C), or a highly aneuploid genome that does not translate well to BAF discordances, such as in HCC1599 (Figure S3D).

To look for evidence of intra-institutional genetic drift, we first compared the BAF-generated drift profiles made from RNA-seq data with the profiles generated from SNP array across all 3 datasets (Table S4). By comparing all datasets to the CCLE dataset, we found that we could accurately call the same drifted/non-drifted regions in 88% and 85.2% of the genome for 90% of the overlapping samples with GDSC or gCSI, respectively (Figures S4A–S4C). In an exemplary case of inter- and intra-institutional drift between GDSC and CCLE, the drift profiles inferred by RNA-seq in the HuP-T4 cell line had regions of the genome that overlapped with SNP-array estimates from GDSC (Chr4q, 6p, 7, 10p, 12q, and 17), regions that overlapped with CCLE (Chr3q, 4p, 8, 12p, 16p, and 21), and regions that agreed with neither (Chr 6q, 16q, and 20; Figure S4D). Meanwhile, there are cases such as VM-CUB-1 where large variance in RNA-seq estimates of BAFs confounds a proper analysis, resulting in regions predicted to be drifted appearing as the inverse of the SNP-array BAF analysis (Figure S4E). Such errors are expected as the noise from RNA-seq data diminishes confidence of BAF-estimated regions, making RNA-seq estimates less sensitive to any high ploidy or drifted regions. Overall, these results suggest that genetic drift can be inferred from RNA-seq data and that this issue of heterogeneity is prevalent both between and within studies.

Association between Genetic Drift and Drug Response

As aneuploidy is a consequence of chromosomal instability, we analyzed whether there was an association between high chromosomal instability and high levels of genetic drift between isogenic cell lines. Using RNA-seq from CCLE, GDSC, and gCSI datasets, we quantified chromosomal instability between isogenic cell lines by taking the mean expression from a set of 70 genes that are known to reflect chromosomal instability (CIN70) (Carter et al., 2006). To complement this analysis, we calculated the total genomic region of significantly drifted regions across all cell lines (STAR Methods). We did not observe any correlation between the level of genetic drift between cell lines and the CIN70 score (GDSC-CCLE: $r_{CN\_CIN} = -0.01$, $r_{BAF\_CIN} = 0.01$; GNE-CCLE: $r_{CN\_CIN} = -0.03$, $r_{BAF\_CIN} = 0.01$; Figures 3A and 3D). However, we observed a large correlation between CIN70 scores and the variance of total genetic drift for those cell lines (GDSC-CCLE: $r_{CN\_CIN} = 0.66$, $r_{BAF\_CIN} = 0.49$; GNE-CCLE: $r_{CN\_CIN} = 0.67$, $r_{BAF\_CIN} = 0.53$), suggesting that CIN70 reflects the capability of the cell lines to drift more than the actual observed drift.

We next investigated whether the discrepancy between pharmacological responses in isogenic cell lines could be attributed to genetic drift or chromosomal instability. For this analysis, we considered 63 drugs found in at least 2 datasets, 20 of which are found in all three. Between two datasets and per drug, we investigated (1) the correlation between the area between the drug dose-response curves (ABC) of isogenic cell lines and the CIN70 score and (2) the correlation between the ABC of isogenic cell lines and estimates of total genomic drift (BAF or CN). We found that between CCLE and GDSC, the FGFR tyrosine kinase inhibitor AZD4547 had the strongest correlation between CIN70 and ABC ($p = 0.16$, q value = 0.06) while other drugs such as dinacicilib, topotecan, teniposide, and trametinib all showed a trend to significance between CIN70 and ABC ($q < 0.25$; Figures 3B and 3C). The dual IGF-IR/insulin receptor inhibitor, BMS−754807, showed the strongest correlation between genetic drift and ABC ($p = 0.46$, q value = 0.34) with a strong negative correlation between CIN70 and ABC ($p = -0.31$, q value = 0.43); however, these estimates were made using only 37 samples and did not reach statistical significance. Unfortunately, we could not test the robustness of these estimates as none of these drugs were found in the gCSI dataset. Furthermore, there were no significant correlations between genetic drift or CIN70 and drug response discordance when comparing gCSI and CCLE (Figures 3E and 3F). Hence, our findings suggest that only a small part of the discrepancy between drug response in cancer cell lines can be linked to either chromosomal instability or genetic drift.

CCLid

In order to replicate our work in existing and future datasets, we developed the cancer cell line identification (CCLid) toolkit as both an R package (https://github.com/bhklab/CCLid) and web application (cclid.ca) to address the lack of a publicly available resource for genotype-based cell line authentication (Korch and Varella-Garcia, 2018). This tool was designed to use generic VCFs generated from variant callers, such as MuTect(2) (Cibulskis et al., 2013; Korch and Varella-Garcia, 2018). The variant calls are then overlapped with up to 507,452 SNPs found across 1,497 unique cancer cell lines from the GDSC, CCLE, and gCSI datasets. We minimize the memory footprint by binning the genome into 500-kb bins and select 5,701 representative SNPs within each region with the smallest variance to ensure a uniform and equal representation of the entire genome. We have pipelines implemented that use the resulting sample by BAF matrix for three major analyses: (1) predicting cellular identity, (2) deconvolution of cross-contaminated samples, and (3) approximation of genetic drift (Figure 4).

While our analysis was done using 5,701 SNPs, we showed that a minimum of 60 SNPs were required for accurate genotyping (Figure S5); opening up the possibility for usage in sequencing datasets, including targeted panels greater than 75 kb in size (Zhao et al., 2003). Similarly, at this threshold, we were able to detect and deconvolute the overall fraction of a cross-contaminated sample in an *in silico* dilution series as low as a 4:1 ratio (Figure 4B). While we demonstrated that detection of cell mixtures is limited to a 4:1 ratio, we were able to accurately deconvolute when given prior knowledge of which cell lines may constitute the admixed sample (Figure 4B). Using these tools will allow researchers to not only screen the genetic identity of their cell lines, but test for cases

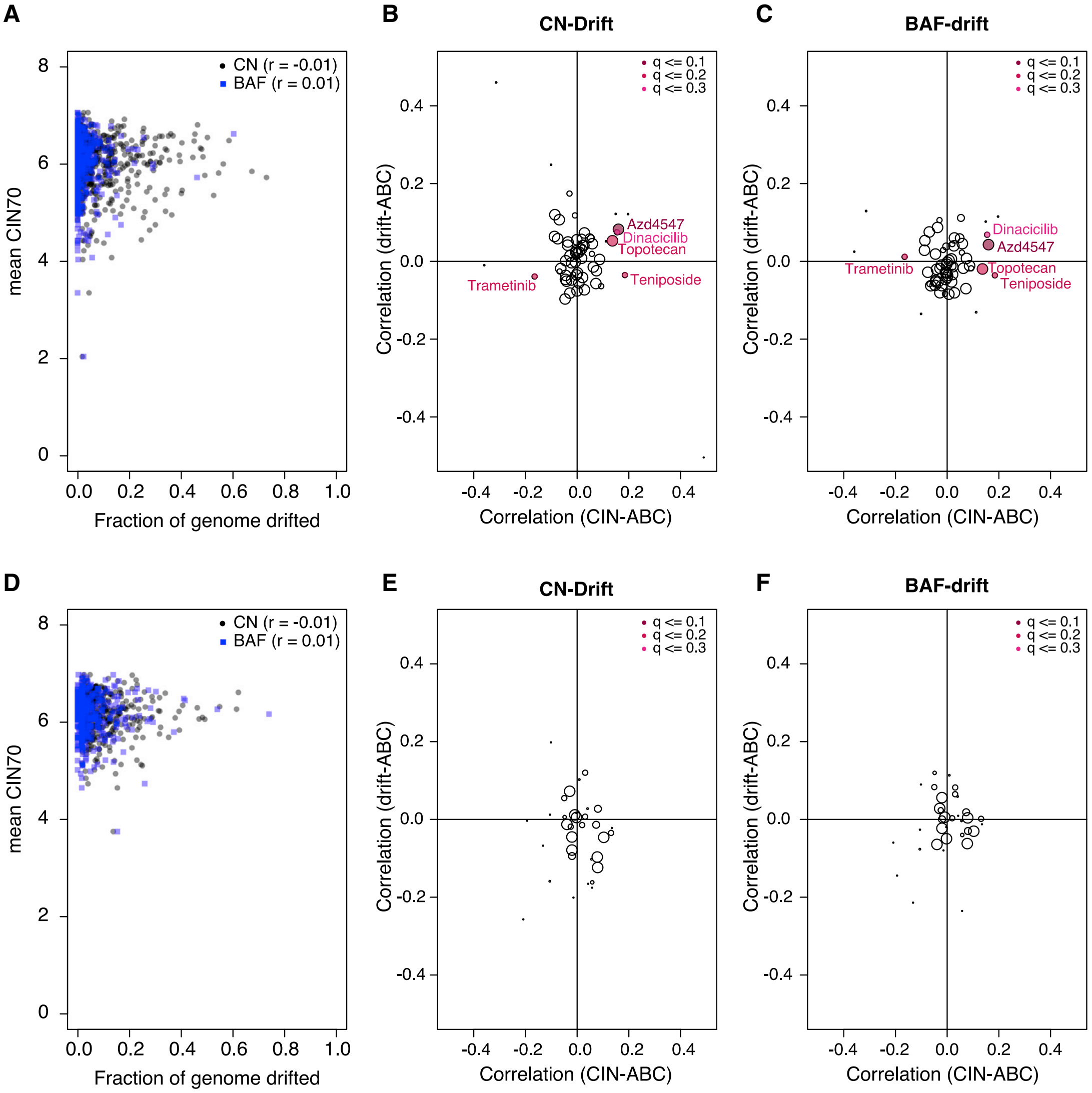


Figure 3. Associations of Phenotypic Data with Measures of Genetic Drift

(A and D**)** Plots of total genomic fraction with evidence of drift against the mean CIN70 score for GDSC compared with CCLE (A) and for gCSI compared with CCLE (D).

(B, C, E, and F) Correlation of CIN70 scores to the ABC for each drug between, plotted against the correlation of total genomic fraction drifted to ABCs. The size of the circle represents the number of cell lines used in the correlation, while the color of the circle represents the false discovery rate (FDR) adjusted q value. Drugs with an adjusted q value less than 0.3 are annotated on the plot. GDSC compared with CCLE are represented in plots (B and C) and gCSI compared with CCLE are in plots (E and F). Plots are separated using either CN estimates of drift (LRR) (B and E) or BAF estimates of drift (C and F).

of cross-contamination with or without prior knowledge of cell line mixtures.

With the surge of transcriptomic profiling in the current literature, researchers can still ensure the identity or drift of their cell lines in the absence of genomic data. Not only can simple genotype checking be done using RNA-seq data, but in the pharmacogenomic datasets used in this study, we found that an average of 1,519 SNPs distributed across 500-kb bins were sufficient to estimate regions of drift in isogenic cell lines (Figure 4C). We hope other researchers can benefit from these resources to not only validate the genetic identity of their cell lines, but also to use them in longitudinal research designs to ensure genetic integrity across multiple passages and treatments.

DISCUSSION

Our study expands on the 106 cell line GDSC-CCLE molecular profile comparison work by Ben-David et al. (2018) by comparing 1,497 cell lines found across the three largest pharmacogenomic studies published to date (Barretina et al., 2012; Haverty et al., 2016; Yang et al., 2013). We detected 2 out of 688 unique cell lines that existed in 2 or more datasets where the genotype was discordant despite synonymous cell line annotation (0.3%), significantly fewer than the expected ∼18% misidentification rate (MacLeod et al., 1999). A further analysis revealed that there were 9 out of a possible 1032 cell lines were isogenic despite all annotations depicting these as definitively different lines (0.9%), and 11 cell lines

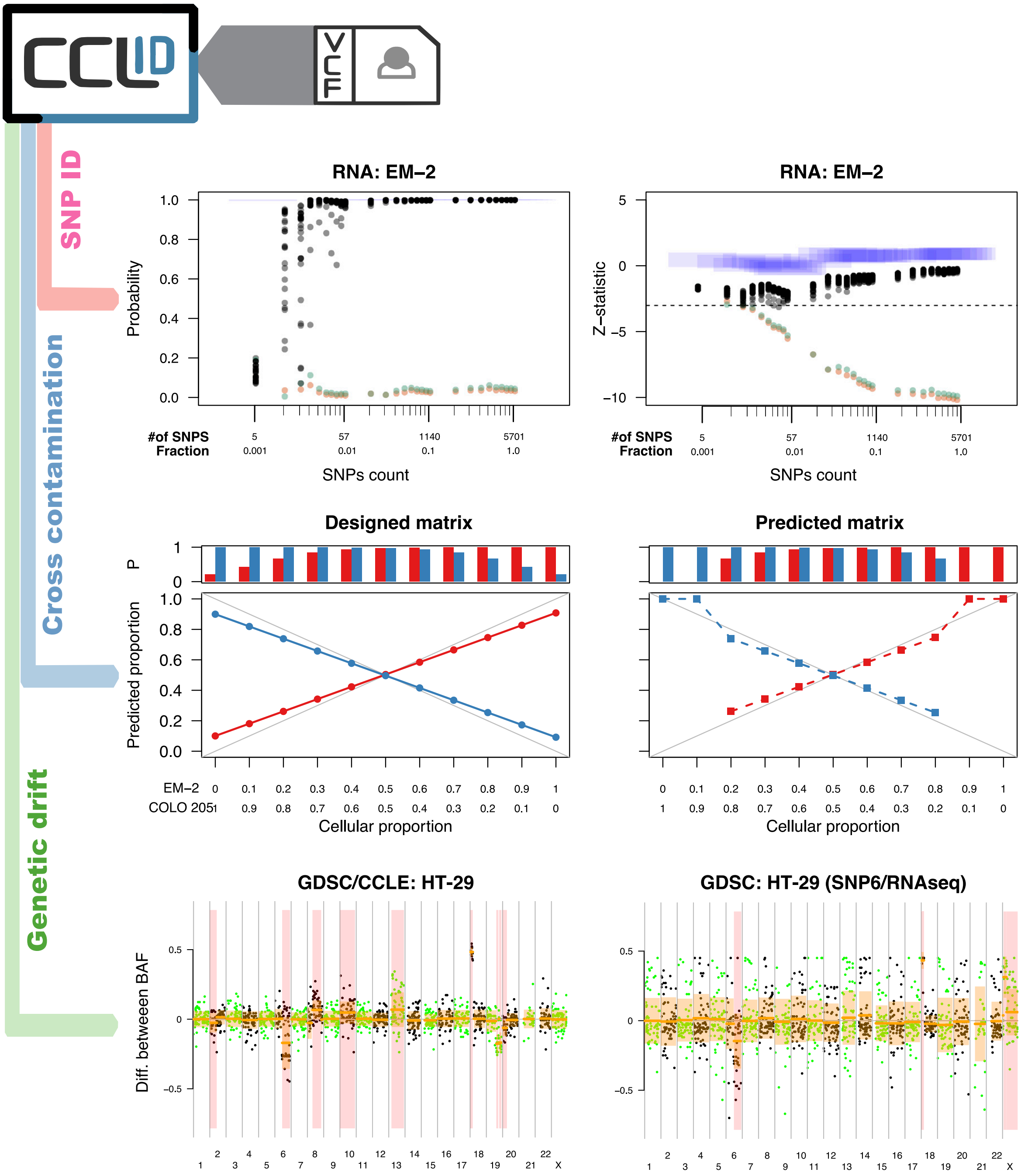


Figure 4. Demonstration of the CCLid Toolkit to Analyze BAF Data

(A) Two measures of genotype concordance using cell line data from RNA-seq data compared against the SNP-array counterpart, the left plot uses a logistic regression, and right plot estimates z statistic from the "mismatch" distribution.

(B) NMF is used to deconvolute an *in silico* mixture of EM-2 and COLO-205 cell lines where the left plot uses prior knowledge of EM-2 and COLO-205 and the right plot uses CCLid to predict matching cell lines prior to deconvolution.

(C) Genetic drift estimates are made using isogenic cell lines between SNP-array data from two datasets, and between RNA-seq and SNP-array data from the same dataset where red areas are flagged as drifted regions.

that were isogenic and differently annotated, but with metadata that did not definitively rule out the possibility of originating from the same line (1.1%). There was a vast range of genetic drift in the 1,497 unique cancer cell lines analyzed: up to 71.1%–72.6% of the genome in the GDSC-CCLE comparisons (median_{CN} = 4.5%, median_{BAF} = 2.5%), and up to 61.8%–73.6% in the CCLE-gCSI comparisons ($\text{median}_{CN=}$ 6.1%, median_{BAF} = 1.9%). However, some of these numbers are confounded by subclonal detection using the L2R approach and from noise in non-genomic data or high copy-number regions using the BAF approach.

Genetic drift between isogenic cell lines can create variable responses to therapeutic compounds as illustrated by the MCF-7 cell line work done by both Ben-David et al. and Kleensang et al. (Ben-David et al., 2018; Kleensang et al., 2016). Our CCLid tool expands on the genetic stability tool, Cell STRAINER, (Ben-David et al., 2018) by allowing for a karyotypic comparison using discrepancies in BAF between RNA-seq and SNP array rather than Cell STRAINERs L2R approach, which is limited to genomic data only. Our BAF-centric approach shows evidence of both intra- and inter-institutional genetic drift between cell lines, suggesting that different passages or separate stocks with distinct molecular profiles may have been used for pharmacological profiling. These confounders, especially the presence of intra-institutional drift, affects downstream pharmacogenomic analyses as chromosomal instability can lead to rapid shifts in the genomic landscape across each passage (Thompson and Compton, 2008), and the landscape assayed may not accurately represent the cells at the time of treatment. Our results emphasize the need for measures of genetic stability throughout an experiment and for molecular profiling to be done as close to pharmacological profiling as possible.

While excellent guidelines exist for handling (Geraghty et al., 2014) and authenticating cell lines (Korch and Varella-Garcia, 2018), there is an absence of protocols that detect and monitor the stability of genetic profiles throughout a study. To aid in stability screening, we developed CCLid to allow users to compare the genomic and transcriptomic profiles of cell lines across 1,497 unique cell lines spread across 3 datasets. This extra versatility has the added benefit of being amenable to multi-omic datasets where L2R cannot be accurately estimated on certain data types (i.e., transcriptomic). To maintain the identity and stability of the cell lines throughout a study, we propose the use of stability screeners such as Cell STRAINER or CCLid in addition to:

(1) Co-isolating DNA and RNA from the same sample and same day (or as close to) to limit the effects of genetic drift
(2) Validating the identity of cell lines using their SNP genotypes as they are more stable than karyotype or STR and less susceptible to microsatellite instabilities (Much et al., 2014)
(3) Establish a baseline genetic profile and compare to the profile prior to, and following experimental manipulation to ensure maintenance of genetic identity and assess the level of genetic drift
(4) Limit the number of passages as whole chromosome missegregation occurs once every 2–5 cell divisions (Laughney et al., 2015; Lupski, 2007; Sebat et al., 2007; van Ommen, 2005)

While these practices can be implemented using any available sequencing data, they will be most reliable and sensitive with genomic data. We believe that our tool will benefit researchers to ensure the integrity of their cancer cell lines in retrospective and prospective research.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.08.012.

ACKNOWLEDGMENTS

We thank the investigators of the Genomics of Drug Sensitivity in Cancer (GDSC), the Cancer Cell Line Encyclopedia (CCLE), and Genentech (gCSI) who have made their invaluable data available to the scientific community. B.H.-K. was supported by the Gattuso Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre, the Canadian Institute of Health Research, Natural Sciences and Engineering Research Council, and the Stand Up To Cancer Canada.

AUTHOR CONTRIBUTIONS

R.Q. analyzed and interpreted the cancer cell line data, developed the CCLid tool, and wrote the manuscript. P.S. and Z.S. developed the PharmacoGx package, which all expression and drug-response data were leveraged from. D.T. and C.H. developed the CCLid web application. N.E.-H. did preliminary work on cell line annotation and data assembly. B.H.-K. and T.J.P. provided supervision and managed this project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 18, 2019
Revised: July 4, 2020
Accepted: August 18, 2020
Published: September 15, 2020

REFERENCES

Affymetrix. (2008). Quality control assessment in genotyping console. Lab. Investig. J. Tech. Methods Pathol. *93*, 970–982.

Bairoch, A. (2018). The Cellosaurus, a cell-line knowledge resource. J Biomol Tech *29*, 25–38.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The cancer cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., Strathdee, C.A., Dempster, J., Lyons, N.J., Burns, R., et al. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. Nature *560*, 325–330.

Birkbak, N.J., Eklund, A.C., Li, Q., McClelland, S.E., Endesfelder, D., Tan, P., Tan, I.B., Richardson, A.L., Szallasi, Z., and Swanton, C. (2011). Paradoxical relationship between chromosomal instability and survival outcome in cancer. Cancer Res *71*, 3447–3452.

Cancer Cell Line Encyclopedia Consortium, and Genomics of Drug Sensitivity in Cancer Consortium. (2015). Pharmacogenomic agreement between two cancer cell line data sets. Nature *528*, 84–87.

Carter, S.L., Eklund, A.C., Kohane, I.S., Harris, L.N., and Szallasi, Z. (2006). A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat. Genet. *38*, 1043–1048.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Fan, H., and Chu, J.Y. (2007). A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics *5*, 7–14.

Freedman, L.P., Gibson, M.C., Wisman, R., Ethier, S.P., Soule, H.R., Reid, Y.A., and Neve, R.M. (2015). The culture of cell culture practices and authentication—results from a 2015 Survey. BioTechniques *59*, 189–192.

Geraghty, R.J., Capes-Davis, A., Davis, J.M., Downward, J., Freshney, R.I., Knezevic, I., Lovell-Badge, R., Masters, J.R.W., Meredith, J., Stacey, G.N., et al. (2014). Guidelines for the use of cell lines in biomedical research. Br. J. Cancer *111*, 1021–1046.

Haibe-Kains, B., El-Hachem, N., Birkbak, N.J., Jin, A.C., Beck, A.H., Aerts, H.J.W.L., and Quackenbush, J. (2013). Inconsistency in large pharmacogenomic studies. Nature *504*, 389–393.

Hatzis, C., Bedard, P.L., Birkbak, N.J., Beck, A.H., Aerts, H.J., Stem, D.F., Shi, L., Clarke, R., Quackenbush, J., and Haibe-Kains, B. (2014). Enhancing reproducibility in cancer drug screening: how do we move forward? Cancer Res *74*, 4016–4023.

Haverty, P.M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., Neve, R.M., Martin, S., Settleman, J., Yauch, R.L., and Bourgon, R. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. Nature *533*, 333–337.

Kleensang, A., Vantangoli, M.M., Odwin-DaCosta, S., Andersen, M.E., Boekelheide, K., Bouhifd, M., Fornace, A.J., Jr., Li, H.-H., Livi, C.B., Madnick, S., et al. (2016). Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function. Sci. Rep. *6*, 28994.

Korch, C., Spillman, M.A., Jackson, T.A., Jacobsen, B.M., Murphy, S.K., Lessey, B.A., Jordan, V.C., and Bradford, A.P. (2012). DNA profiling analysis of endometrial and ovarian cell lines reveals misidentification, redundancy and contamination. Gynecol. Oncol. *127*, 241–248.

Korch, C., and Varella-Garcia, M. (2018). Tackling the human cell line and tissue misidentification problem is needed for reproducible biomedical research. Adv. Mol. Pathol. *1*, 209–228.e36.

Laughney, A.M., Elizalde, S., Genovese, G., and Bakhoum, S.F. (2015). Dynamics of tumor heterogeneity derived from clonal karyotypic evolution. Cell Rep *12*, 809–820.

Lin, X., and Boutros, P.C. (2020). Optimization and expansion of non-negative matrix factorization. BMC Bioinformatics *21*, https://doi.org/10.1186/s12859-019-3312-5.

Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E.G., Van Drogen, A., Borel, C., Frank, M., Germain, P.L., Bludau, I., et al. (2019). Multi-omic measurements of heterogeneity in HeLa cells across laboratories. Nat. Biotechnol. *37*, 314–322.

Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. Nat. Genet. *39*, S43–S47.

MacLeod, R.A.F., Dirks, W.G., Matsuo, Y., Kaufmann, M., Milch, H., and Drexler, H.G. (1999). Widespread intraspecies cross-contamination of human tumor cell lines arising at source. Int. J. Cancer *83*, 555–563.

Mayrhofer, M., Viklund, B., and Isaksson, A. (2016). Rawcopy: improved copy number analysis with Affymetrix arrays. Sci. Rep. *6*, 36158.

Much, M., Buza, N., and Hui, P. (2014). Tissue identity testing of cancer by short tandem repeat polymorphism: pitfalls of interpretation in the presence of microsatellite instability. Hum. Pathol. *45*, 549–555.

Nilsen, G., Liestøl, K., Van Loo, P., Vollan, H.K.M., Eide, M.B., Rueda, O.M., Chin, S.F., Russell, R., Baumbusch, L.O., Caldas, C., et al. (2012). Copynumber: efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics *13*, 591.

Robin, T., Capes-Davis, A., and Bairoch, A. (2020). CLASTR: the Cellosaurus STR similarity search tool - a precious help for cell line authentication. Int. J. Cancer *146*, 1299–1306.

Safikhani, Z., Smirnov, P., Freeman, M., El-Hachem, N., She, A., Rene, Q., Goldenberg, A., Birkbak, N.J., Hatzis, C., Shi, L., et al. (2016). Revisiting inconsistency in large pharmacogenomic studies. F1000Res *5*, 2333.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–449.

Smirnov, P., Kofia, V., Maru, A., Freeman, M., Ho, C., El-Hachem, N., Adam, G.A., Ba-alawi, W., Safikhani, Z., and Haibe-Kains, B. (2018). PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. Nucleic Acids Res *46*, D994–D1002.

Smirnov, P., Safikhani, Z., El-Hachem, N., Wang, D., She, A., Olsen, C., Freeman, M., Selby, H., Gendoo, D.M., Grossmann, P., et al. (2016). PharmacoGx: an R package for analysis of large pharmacogenomic datasets. Bioinformatics *32*, 1244–1246.

Thompson, S.L., and Compton, D.A. (2008). Examining the link between chromosomal instability and aneuploidy in human cells. J. Cell Biol. *180*, 665–672.

Van Loo, P., Nordgard, S.H., Lingjærde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., et al. (2010). Allele-specific copy number analysis of tumors. Proc. Natl. Acad. Sci. USA *107*, 16910–16915.

van Ommen, G.J. (2005). Frequency of new copy number variation in humans. Nat. Genet. *37*, 333–334.

Vaughan, L., Glänzel, W., Korch, C., and Capes-Davis, A. (2017). Widespread use of misidentified cell line KB (HeLa): incorrect attribution and its impact revealed through mining the scientific literature. Cancer Res *77*, 2784–2788.

Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res *41*, D955–D961.

Zhao, Z., Fu, Y.-X., Hewett-Emmett, D., and Boerwinkle, E. (2003). Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene *312*, 207–213.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Genomics of Drug Sensitivity in Cancer (SNP 6.0, n=1013)	(Yang et al., 2013)	EGAD00010000644
Cancer Cell Line Encyclopedia (SNP 6.0, n=1190)	(Barretina et al., 2012)	GSE36139
Genentech Cell Line Screening Initiative (Illumina 2.5M, n=668)	(Haverty et al., 2016)	EGAD00010000951

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Benjamin Haibe-Kains (bhaibeka@uhnresearch.ca)

Materials Availability

This study did not generate new materials.

Data and Code Availability

Our code and documentation are open-source and publicly available through the CellLineConcordance GitHub repository (https://github.com/bhklab/CCLid).

METHOD DETAILS

Pharmacogenomic Datasets

We analyzed 2,219 Affymetrix Genome-Wide Human SNP 6.0 (Affy6) arrays and 668 HumanOmni2.5-Quad (Omni) arrays from 1,497 unique cell lines profiled by 3 pharmacogenomic studies (STAR Methods), resulting in a comparison of 688 unique cell lines that are found in 2 or more datasets. CEL files and IDAT were downloaded from the accession codes outlined in the STAR Methods. Cell lines were linked to their Cellosaurus IDs (Bairoch, 2018) using functions included in CCLid that parses the cellosaurus distributed XML file (ftp://ftp.expasy.org/databases/cellosaurus; downloaded on February, 2020).

Preprocessing of Affymetrix SNP 6.0 Arrays

Raw Affy6 files were pre-processed using the rawcopy pipeline (Mayrhofer et al., 2016) implemented in the Easy Copy Number (EaCoN) package (https://github.com/gustaveroussy/EaCoN). Analysis was conducted using the GenomeWideSNP_6 Annotations Release 35 manifest. As per Affy6 guidelines, samples and datasets were removed based on poorly resolved genotype clusters and a bias toward restriction enzyme digestion during library preparation when assayed using affymetrix power tools (APT version 1.16.1) (Affymetrix, 2008). The B-allele frequency (BAF) and Log$_2$ Ratios (L2R) were exported from the EaCoN package prior to allele-specific segmentation using ASCAT version 2.5.1 (Tumor-Only mode) (Van Loo et al., 2010). By modifying the EaCoN package, we implemented measures to export total and allele-specific L2Rs, as well as modal copy numbers to a PharmacoGX PSet (Smirnov et al., 2016, 2018) for all datasets analyzed.

Preprocessing of Illumina Human Omni Arrays

Raw Omni files were preprocessed using GenomeStudio version 2 using the HumanOmni2.5-Quad version 1.0 manifest and HumanOmni2.5-Quad cluster file. Genotype was called using the Genotype Module. Signal intensities were converted to log-R ratios (LRR) and B-allele frequencies (BAF) using the cnvPartition CNV Analysis Plugin version 3.2.1 The LRR and BAF were used as input into a modified version of EaCoN, which was used fed into ASCAT version 2.5.1 (Tumor-Only mode) to generate copy-number profiles.

Preprocessing of RNAseq data

Raw FASTQ files were aligned to human genome version hg19 and transcript annotation GENCODE v19 using STAR (v 2.4.2a) two-pass method. SNPs that overlapped those found in the Affymetrix SNP6 arrays were force-called using MuTect2 and all SNPs and BAFs were strand-corrected.

QUANTIFICATION AND STATISTICAL ANALYSIS

Comparing Genotypes between Samples

To compute the concordance between genotypes of two samples, we focused on the BAF values of 507,452 SNPs that were found in the SNP-array data for all three datasets used in this study. To make the SNP values from the Omni array comparable to those using Affy6 arrays, we intersect based on genomic location and ensure that Ref and Alt alleles are in the same order as those in Affy6. We reduced the number of SNPs that were used for genotype concordance estimates by binning the genome into 500kb bins and selecting a single representative SNP from each bin based on the lowest variance. Using the resulting 5,701 SNPs uniformly scattered throughout the genome, we did a pairwise cell line comparison to calculate the summation of Euclidean distances between the BAFs for each SNP. We then trained a logistic regression model using the given cell line annotations and under-sampling the non-matching annotation group to rectify class imbalances. Finally, to measure the limit of detection using this regression approach, we selected a subset of SNPs from the 5,701 uniformly scattered SNPs and retrained a logistic regression model for each test.

Any two cell lines were considered isogenic if their predicted similarity was lower than a q-value of 0.05 (multiple testing adjusted). This value was compared to the cell line annotations to establish a confusion matrix for precision-recall estimations.

Cell line deconvolution

A given cell line was deconvoluted into mixtures of cell lines using the R package NNLM (Nonnegative Linear Models) (Lin and Boutros, 2020). A design matrix composed of SNPs by cell lines was used that incorporated prior knowledge of which cell lines composed the input sample. Non-negative matrix factorization was done on the input sample using this design matrix to estimate the relative proportion of each cell line from the design matrix in the input sample.

Comparing Karyotypes between Samples

To calculate concordance between copy-number profiles, we used two approaches: a BAF method and a L2R method. In the BAF approach, we first reduced BAF to the 0-0.5 range and calculated the Euclidean distance between the 5,701 uniformly scattered SNPs. However, if the input data was RNAseq, we removed all SNPs that had fewer than 5 reads supporting the SNP as this was too few to get an accurate BAF estimate. We then segmented the data using the copynumber R package (Nilsen et al., 2012) implementation of piecewise constant fit (pcf) with a minimum of 5 SNPs. Using all available SNPs, we calculated the standard deviation of BAF for each segmented region and used this value to draw a z-score threshold of ±4 to classify a region as significantly drifted.

Using a similar metric, we did the same analysis using the L2R values instead of the BAF. The ASCAT defined segments were used in lieu of SNP probeset L2R values. We calculated the pairwise Euclidean distance between all L2R-segments and median-normalized only if the medians of the two profiles were greater than 0.5 distance apart. This was done to mitigate the issue of two CN-profiles having slightly different L2R values due to non-concordant median-centering in two isogenic lines. For each segment, we calculated the standard deviation by taking the Euclidean distance between individual probeset L2Rs that populate that region. Finally, we used a z-score threshold of ±2 to classify a region as significantly drifted.

Comparing Phenotypes

Reprocessed drug curves and gene expression data were accessed from the R PharmacoGx (version 1.10.3) package (Smirnov et al., 2016). The recomputed areas above the curve (AAC) were used as measurements of drug sensitivity, with higher AACs corresponding to higher sensitivity (Safikhani et al., 2016). To quantify the distance between drug curves, we implemented the area between the curve (ABC) metric described by Safikhani and colleagues (Safikhani et al., 2016). The area was estimated by taking the unsigned area between two curves, intersected over overlapping concentration ranges and normalized by the length of the intersection interval.

Based on the cell lines' RNAseq profiles, we calculated the CIN70 chromosomal instability score (Carter et al., 2006) following a the approach implemented by Birkbak and colleagues (Birkbak et al., 2011) (CIN70 score = mean FPKM of the 70 genes that populate this geneset). Estimates of total genomic fraction drifted and CIN70 scores were used to calculate the Pearson correlation with the ABC metrics for each drug.