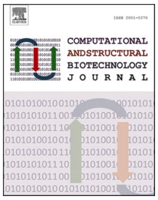
Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj



Software/web server article

CCLHunter: An efficient toolkit for cancer cell line authentication

Congfan Bu[a,b,1], Xinchang Zheng[a,b,1], Jialin Mai[a,b,c], Zhi Nie[a,b,c], Jingyao Zeng[a,b], Qiheng Qian[a,b,c], Tianyi Xu[a,b,c], Yanling Sun[a,b,c], Yiming Bao[a,b,c,*], Jingfa Xiao[a,b,c,*]

[a] National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China
[b] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Cancer cell line
Cell line authentication
Web server

ABSTRACT

Cancer cell lines are essential in cancer research, yet accurate authentication of these cell lines can be challenging, particularly for consanguineous cell lines with close genetic similarities. We introduce a new Cancer Cell Line Hunter (CCLHunter) method to tackle this challenge. This approach utilizes the information of single nucleotide polymorphisms, expression profiles, and kindred topology to authenticate 1389 human cancer cell lines accurately. CCLHunter can precisely and efficiently authenticate cell lines from consanguineous lineages and those derived from other tissues of the same individual. Our evaluation results indicate that CCLHunter has a complete accuracy rate of 93.27%, with an accuracy of 89.28% even for consanguineous cell lines, outperforming existing methods. Additionally, we provide convenient access to CCLHunter through standalone software and a web server at https://ngdc.cncb.ac.cn/cclhunter.

1. Introduction

As an in vitro model system, cancer cell lines are widely used in different fields of biological research, especially in cancer studies and drug discovery [1,2]. However, problematic issues arise when using cell lines during experiments, such as genomic variation and cell contamination, which could cause cell heterogeneity as cell lines proliferate, issues that have been overlooked for a long time [3–5]. It was reported that as of August 2017, 32,755 research papers may have used the wrong cell line [6]. Such a high proportion of misidentified cell lines has collected increasing attention from researchers [7]. To date, at least 21 journals, including *Nature*, have stipulated the requirements of cell line authentication in the submitted articles [8,9].

At present, the methods of cell line authentication are mainly divided into experiments-based and bioinformatics analysis-based approaches. Experiment-based cancer cell lines (CCLs) authentication methods generally include isoenzyme analysis [10], HLA type [11], and short tandem repeat (STR) detection [12]. Bioinformatics analysis-based CCL authentication methods primarily utilize information in next-generation sequencing (NGS) data, such as single nucleotide polymorphisms (SNP) and differential gene expression, which contains more genetic information. The SNP-based methods, such as CEL-ID and Uniquorn calculate the similarity of cell line SNPs between reference and sample using Pearson correlation [13,14]. However, somatic mutations in cancer cell lines often exhibit high individual specificity, leading to excessive noise in cell line authentication [15,16]. Zhang et al. proposed an expression-based method called CCLA, which reduces noise by selecting specifically expressed genes and mapping their values to pathways. Then, it uses refined data to build predictive models with random forests [17].

However, all existing methods apply the same treatment standard to all cell lines, ignoring the genetic differences within each cell line. This could decrease the authentication accuracy for consanguineous cell lines. We define a cell line directly isolated from tissue as a primary-derived cell line (primCCL). In contrast, a cell line isolated from a consanguineous cell line is referred to as a post-derived cell line (postCCL), constituting approximately 19% (18,542/97,506, data from Cellosaurus) of all existing cell lines [18]. Due to the purposeful design of cell lines, the incorrect authentication of postCCLs can result in conflicting experimental outcomes. Currently, the authentication of

[*] Corresponding authors at: National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China.

E-mail addresses: baoym@big.ac.cn (Y. Bao), xiaojingfa@big.ac.cn (J. Xiao).

[1] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors

https://doi.org/10.1016/j.csbj.2023.09.040
Received 3 August 2023; Received in revised form 28 September 2023; Accepted 28 September 2023
Available online 29 September 2023

postCCLs remains a significant challenge.

To authenticate cell lines, particularly post-derived ones, with sensitivity and robustness, we propose a new method called CCLHunter. This approach integrates variant, kindred topology, and expression profile schemes. We use cosine similarity to eliminate noise from hundreds of thousands of SNPs. The resulting refined SNP barcode categorizes the cell line into the lineal kindred group (LKG). We then authenticate CCLs within the LKG by constructing a stable fluctuation unit (SFU) based on the kindred topology and expression profile. Our evaluation shows that CCLHunter accurately and effectively authenticates cell lines, including those closely genetically related.

2. Methods

2.1. Data collection

Genotype data for cell lines were collected from the CCLE [19] and COSMIC [20] projects. For cell lines in the CCLE, we downloaded the ‘Birdseed Call’ files from the DepMap Portal (https://depmap.org/portal/). Each cell line’s genotype calls were downloaded in CSV format for COSMIC. The genotypes of both datasets were generated from the Affymetrix SNP6.0 array [21]. For each probe, the related rsID, location in the genome, strand, allele A, and allele B were annotated using the GPL6801–4019.txt file, which was downloaded from the GEO website. If more than one probe were annotated into one SNP, the one with the highest confidence would be selected in the CCLE dataset. For the COSMIC dataset, as COSMIC lacks a confidence score similar to CCLE, to eliminate duplicate probes, we retain only the results from the initial probe at that locus. Finally, 860,975 SNP sites were filtered for downstream analysis. The cell line name was standardized using the Cellosaurus and manually mapped cell line names from the CCLE and the COSMIC dataset to Cellosaurus [18]. The CCLs’ kindred information was also extracted from Cellosaurus.

2.2. Stable SNP vector set filtration

The screening criteria revolve around the ability to be inherited as stably as possible, and the accuracy problems caused by sequence complexity are minimized. 436 SNP sites and corresponding genes were filtered for building the SNP barcode.

1) Each allele should be located within the CDS regions of recognized coding genes.
2) Each allele should be recognized as a biallelic variant, meaning that only two variants (including its reference) could appear in this location, as determined by the dbSNP ALFA project’s statistics (build id 20201027095038) [22], and its variation type should be a transversion.
3) The allele frequency of each locus in dbSNP [23] should fall within the range of 0.4–0.6.
4) The allele frequency of each locus in our curated CCLs should also fall within the range of 0.4–0.6.
5) Each locus should not be within the tandem repeat regions identified by the tandem repeat finder with default parameters [24]
6) Each SNP should not be located within the linkage disequilibrium regions [25,26]
7) Only the one farthest from the CDS boundary will be retained if two or more alleles are located on the same coding gene.

All of the refined SNPs were listed in the Supplementary Table 2.

2.3. Candidate primCCLs selection

The genotype and the depth of query CCL in SNP barcode were extracted using samtools [27]. In detail, all reads aligned to this location were extracted, and only quality scores with both read and alignments higher than Q30 were retained. For genotype determination, we require that each haplotype has a minimum depth of 3 reads and accounts for at least 15% of all reads; otherwise, we consider this location as a missing location labeled as ’NN.’ After one-hot coding, the cosine similarity was calculated between each CCL in the library. The most similar one with query was recorded as the primCCLs candidate.

2.4. Lineal kindred group building

If this CCL recursively had any direct or collateral relatives (including its different tissues), the topology was extracted as the lineal kindred group of this cell line. Each cell line shares this and only this same LKG in the kindred topology. CCLHunter will re-identify which cell line is the most likely in this LKG based on the stable fluctuation unit.

2.5. Stable fluctuation unit selection

We used a stable fluctuation unit (SFU) to select the gene set between two cell lines with the greatest difference among the expressed genes as stable as possible. First, using the depth values obtained from the previous step, we calculated the mean value for all gene pairs and selected the gene set falling within the 25th to 75th percentiles. Next, we calculated the fold change for all gene pairs and selected genes with fold changes greater than the 75th percentile or less than the 25th percentile. The objective of this step is to mitigate the impact of expression outlier genes on authentication while preserving as many differentially expressed genes as possible. Finally, the shared part of the two gene sets was considered as stable fluctuation units of the two cell lines.

2.6. Test data set preparation

Three data sets were used to test CCLHunter and other software. We collected RNA-seq data from 17 cell lines in SRA (Supplementary Table 3). These cell lines are all post-derived cell lines, which makes them difficult to be authenticated by the previous methods. The information on these 17 cell lines is listed in Supplementary Table 3. The second and third data sets are CHCC (E-MTAB-2706 dataset in EBI) [28] and CellMiner [29], released by other projects, respectively. CellMiner provided expression and genotype data for 60 cell lines, and CHCC provided expression data for 622 cell lines. Because the CHCC project has no available genotype data, when testing CCLHunter, we assume that CCLHunter can correctly locate the LKG of the postCCL, which should be the first step for CCLHunter using SNPs, so the actual accuracy will be influenced by the accuracy of the first-step SNP-based method.

2.7. RNA-seq data processing

All cell line RNA-seq data were downloaded from the NCBI SRA database. We used the fastq-dump module of sratoolkit version 2.8.2–1 [30] to convert sra data to fastq format with the parameter of –split 3 –gzip. STAR [31] was used for read alignment with parameters of –twopassMode Basic –outSAMtype BAM SortedByCoordinate. The data size and alignment ratio are shown in Supplementary Table 3. All sorted BAM were quantified at the gene level by FeatureCounts with default parameters [32].

2.8. Performance comparison with other tools

Three well-known authentication tools, CCLA[17], CEL-ID[13] and Uniquorn[14], were used to compare with CCLHunter. All software used default parameters to get the result and then mapped to Cellosaurus to uniform the cell line name. For CEL-ID and Uniquorn, the input BAM file was obtained from STAR version 2.7.6a[31] in the previous RNA-seq processing step. The gene expression matrix for testing CCLA was generated using FeatureCount [32] with default parameters. Due to the variant caller dependence, two different variant callers, Varscan2[33] and Freebayes[34] were used to build the input of the CEL-ID and Uniquorn. According to the guidance of GATK [35] best practices, GATK's MarkDuplicates, SplitNCigarReads, and BaseRecalibrator were used to filter and correct the data. Subsequently, the filtered BAM was processed in the subsequent variant calling steps. The input for CEL-ID was created using mpileup2snp of varscan2 with the parameters of –output-vcf 1 –p-value 0.01, while the input for Uniquorn was generated using freebayes with its default settings. Given that the CHCC and CellMiner projects only provide expression data and limited variant locus information, we could not assess the accuracy of SNP-based methods using these test sets. For the CHCC data sets that only furnish the expression profile, we hypothesize that CCLHunter can accurately identify the correct LKG through SNP analysis during testing.

3. Results

3.1. The schema of CCLHunter

The workflow of CCLHunter is shown in Fig. 1. We provided a novel CCL authentication method to promote the resolution of CCL identification. First, we collected the hierarchical relationship links of CCLs from Cellosaurus [18]. Through the kindred topology, we could authenticate the relatives of any cell line record, such as their parents, children, or siblings. All immediate relatives of a cell line and themselves are called lineal kindred groups (LKGs). The topology of the LKG serves as a guide to inform CCLHunter whether it should utilize expression data to refine the CCLs further.

Secondly, we collected SNP array instead of preparing variation data from CCLE[19] and COSMIC[20]. Furthermore, somatic SNVs provided in the CCLE or COSMIC projects have strong specificity related to individuals/organisms, thus decreasing the specificity as CCL 'population'. 1389 cell lines were collected, including 1042 cell lines in CCLE and 1006 cell lines in COSMIC (Supplementary Table 1). Among them, 168 postCCLs had LKGs containing more than one CCL. Through SNP filtering (see method), we could greatly reduce the scale of data to be scanned (from 860,975 to 436). Simultaneously, individual-specific noise was minimized as much as possible because the retained SNP could be considered as a candidate germline mutation (Fig. 2 D and E). The genotypes of these 436 SNPs were extracted from the reference array as the vector set of each CCL.

Lastly, CCLHunter determined whether it was necessary to continue to refine the cell line through the expression matrix according to the

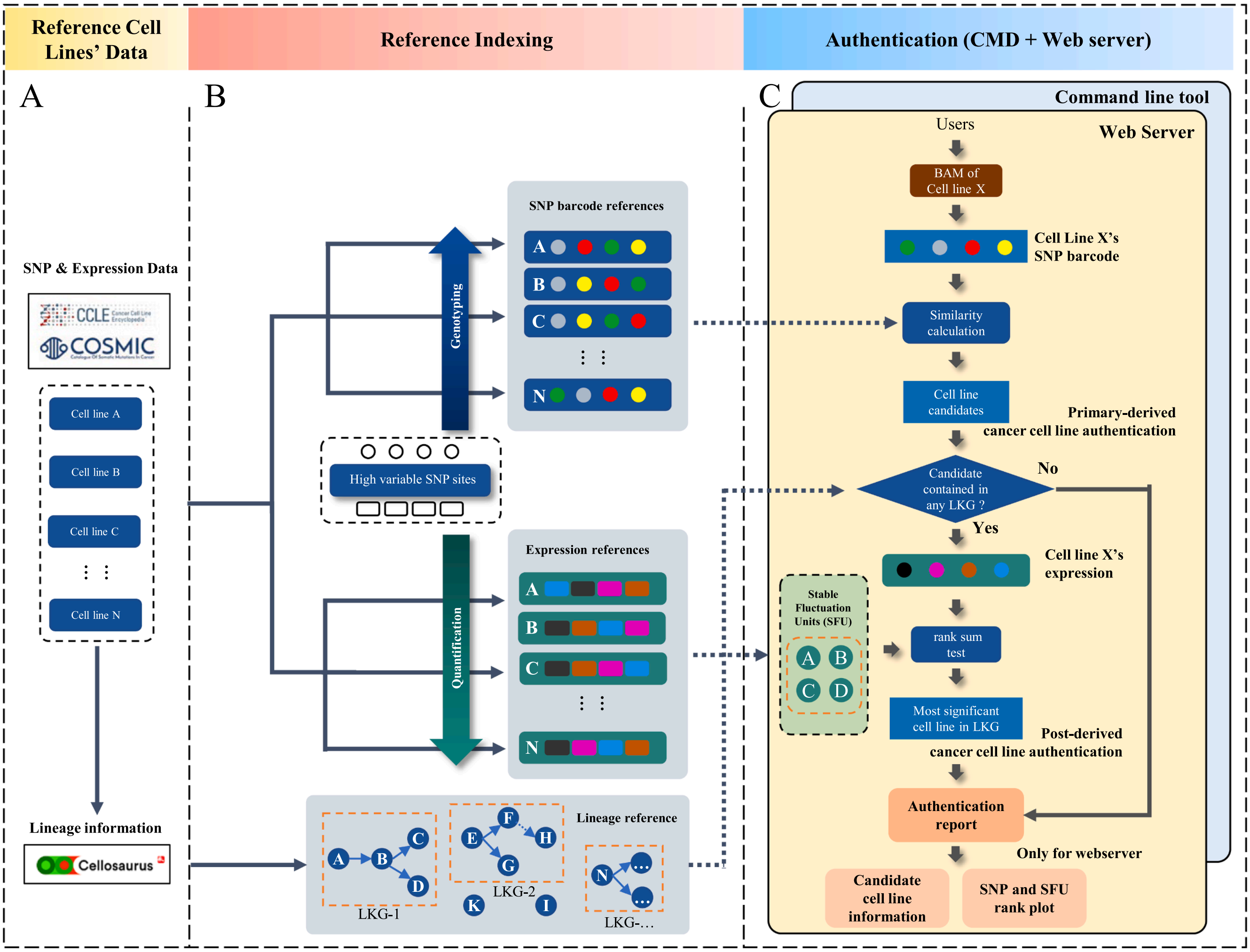


Fig. 1. Schema of CCLHunter. A) The reference data used in CCLHunter to build internal references. Among them, CCLE and COSMIC were used to extract the genotype and expression data, and Cellosaurus was used to extract lineage topology information. B) Internal reference built-in CCLHunter. Firstly, high-variable SNP sites and their related genes refined by rules were selected as a benchmark to build both the SNP barcode and expression matrix for each cell line. Secondly, the lineage reference with kindred topology and the standard name was extracted from Cellosaurus. C) Authentication of the user-provided cancer cell line. CCLHunter will extract the SNP barcode from BAM and calculate the similarity between the user-provided and reference cancer cell lines to identify the primary-derived cancer cell line as the candidate. Then, if the candidate cell line is contained in any LKG, SFU will be constructed dynamically and compared against each reference cell line in LKG. The final results will be treated as the most significant reference cell line. CCLHunter can be run in both command line and Web server mode.

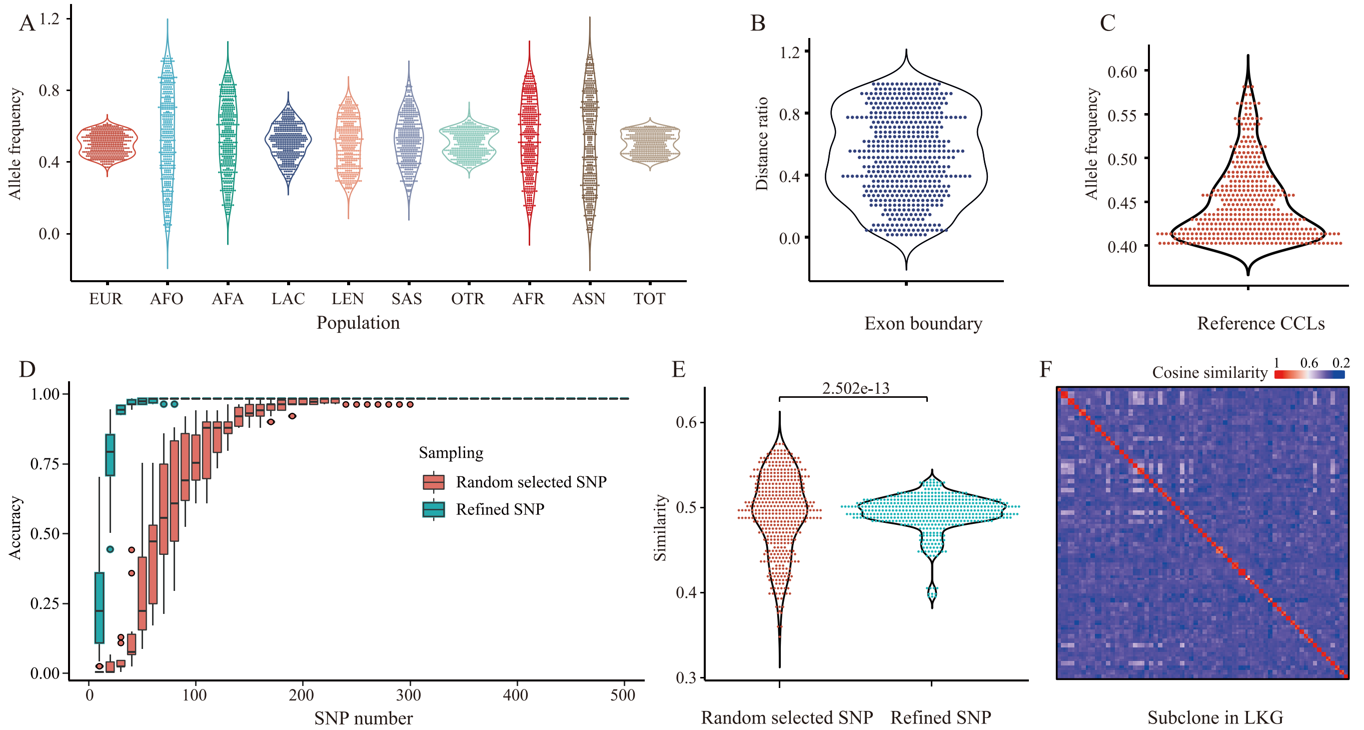


Fig. 2. The distribution and characteristics of refined 436 SNP alleles. A) The population genotype frequency of 436 refined SNPs recorded in ALFA. B) Distance between the refined SNP and its overlapping exon boundary. C) Genotype frequency of refined SNP in the collected cell lines. D) Use gradient randomly selected SNP and our refined SNP to test the SNP-based authentication accuracy. E) Compared with randomly selected, refined SNPs show more robust similarity scores. F) Cosine similarity heatmap with postCCLs. The results show that the SNP-based method cannot distinguish the postCCLs in the LKG.

topology of the best hit of the CCL candidate. If a candidate has any immediate relatives in LKG, it was considered untrustworthy regardless of the results from the SNP-based method. In such cases, it was necessary to determine which candidate was most likely the best match through the gene expression in LKG (Fig. 1B). The expression profile of all 168 postCCLs was mapped to a common name curated from Cellosaurus for expression-based CCL authentication. We then selected genes' expression value as stable fluctuation unit (SFU) (see detail in method) by comparing the input CCL and each CCL contained in LKG to obtain the authenticated report through the rank sum test.

3.2. *Authentication of prim-derived cell lines using refined SNP barcode*

After the screening, we obtained 436 SNP alleles for the rough primary authentication to the LKG (Fig. 2 and Supplementary Table 2). Fig. 2 A-C displays the attributes of the refined SNP barcode. According to the population frequency information of the Allele Frequency Aggregator (ALFA) [22], the allele frequency of refined SNP is approximately 0.5 in the total population. It was more divergent in Asian and African populations, which might be due to insufficient data in these regions. All refined SNPs tend to maintain stable genotype frequencies in the cell lines we collected and those contained in ALFA. This stability contributes to a reduction in false positive SNP identification. Furthermore, we found that all refined SNPs are located in the non-variable region [36] and are evenly distributed across all chromosomes without prior knowledge (Supplementary Figure 1), illustrating the robustness of SNP extraction and the unbiased nature of our screening conditions.

To access whether the refined SNPs contain sufficient information entropy, we used a dataset from CellMiner [37] containing genotype and expression profile data for 60 cell lines to authenticate CCL into LKG levels using our refined SNPs and randomly selected SNPs, respectively (Fig. 2D). The results showed that all cell lines could be exactly distinguished when using the refined SNPs or the number of random-selected SNPs in the coding region was greater than 300. We also compared the similarity distribution between the 436 SNP barcode and the randomly selected SNPs for CCLs authenticating (Fig. 2E). The results demonstrated that the similarity of the refined SNP was significantly more stable than that of the randomly selected ones (KS test, P-value=2.50e-13), indicating the SNP refinement enhances the method's robustness when applied to different real datasets.

After conducting our analysis, we discovered that both the SNP-based methods (whether refined or not) and expression-based methods could not accurately distinguish between cell lines with postCCLs, as shown in Table 1 and Supplementary Table 4. When comparing various cell lines with multiple LKGs in CCLE and COSMIC, we found that the similarity between cell lines within LKGs was significantly higher than that among general cell lines (as illustrated in Fig. 2F). This finding highlights the limitations of relying solely on a single information-based approach to differentiate postCCLs. To address this limitation, CCLHunter has been developed to enhance authentication resolution by extracting stable fluctuation units for cell lines derived from postCCLs.

3.3. *Authentication of post-derived cell lines in LKG using stable fluctuation unit*

The postCCLs in LKG were distinguished using dynamically extracted

Table 1
Evaluation of CCLHunter's accuracy in 3 separate datasets.

	SNP-based	SFU-based	Combined
RNA-seq collected from SRA	76.47% (13/17)	76.47% (13/17)	100% (17/17)
CHCC	/	73.12% (291/398)	96.48% (384/398)*
CellMiner	78.33% (47/60)	23.33% (14/60)	83.33% (50/60)

* : We default that the combined method uses the correct LKG of this CCL.

hot genes as stable fluctuation units (SFUs). For example, IGR-37 and IGR-39 are two cancer cell lines derived from the melanoma tissue of a European man, where IGR-39 is derived from the primary site and IGR-37 is derived from the site of metastasis to inguinal lymph nodes. When comparing the genotypes collected with RNA-seq data in SRA (Supplementary Table 3), we found that they had nearly identical genotypes (Fig. 3A), which may lead to unreliable authentication results (Fig. 3B).

SFU composed of hot genes was used for the rank sum test (Fig. 3C-F and Supplementary Figure 2). We only focused on the internal expression of hot genes in SFU to obtain a qualitative result, rather than any specific value. This approach was chosen because RNA-seq or sequencing data from other methods from different samples can introduce non-negligible technical and biological errors into sampling [38, 39]. Consequently, we observed that the expression-based SFU could significantly distinguish IGR-37 and IGR-39 reciprocally (Fig. 3C, E).

3.4. Performance of CCLHunter in real datasets

We selected three representative datasets from different sources to assess the accuracy of CCLHunter, particularly for postCCLs. The first dataset comprises RNA sequences from 17 cell lines, including 10 LKGs downloaded from SRA (Supplementary Table 3). These CCLs each have at least two postCCLs (e.g., IGR-37 and IGR-39) in the same LKG. The other two datasets are derived from CHCC[28] and CellMiner [29], with genotype and expression profile data provided on their respective project websites.

Our results demonstrate the superior performance of the combined method compared to the single evidence-based approach. As shown in Table 1, We found that the misidentified cell lines were also mostly grouped in the same LKG, suggesting that the SNP-based approach's resolution can reach the LKG level. The accuracy of CCLHunter using

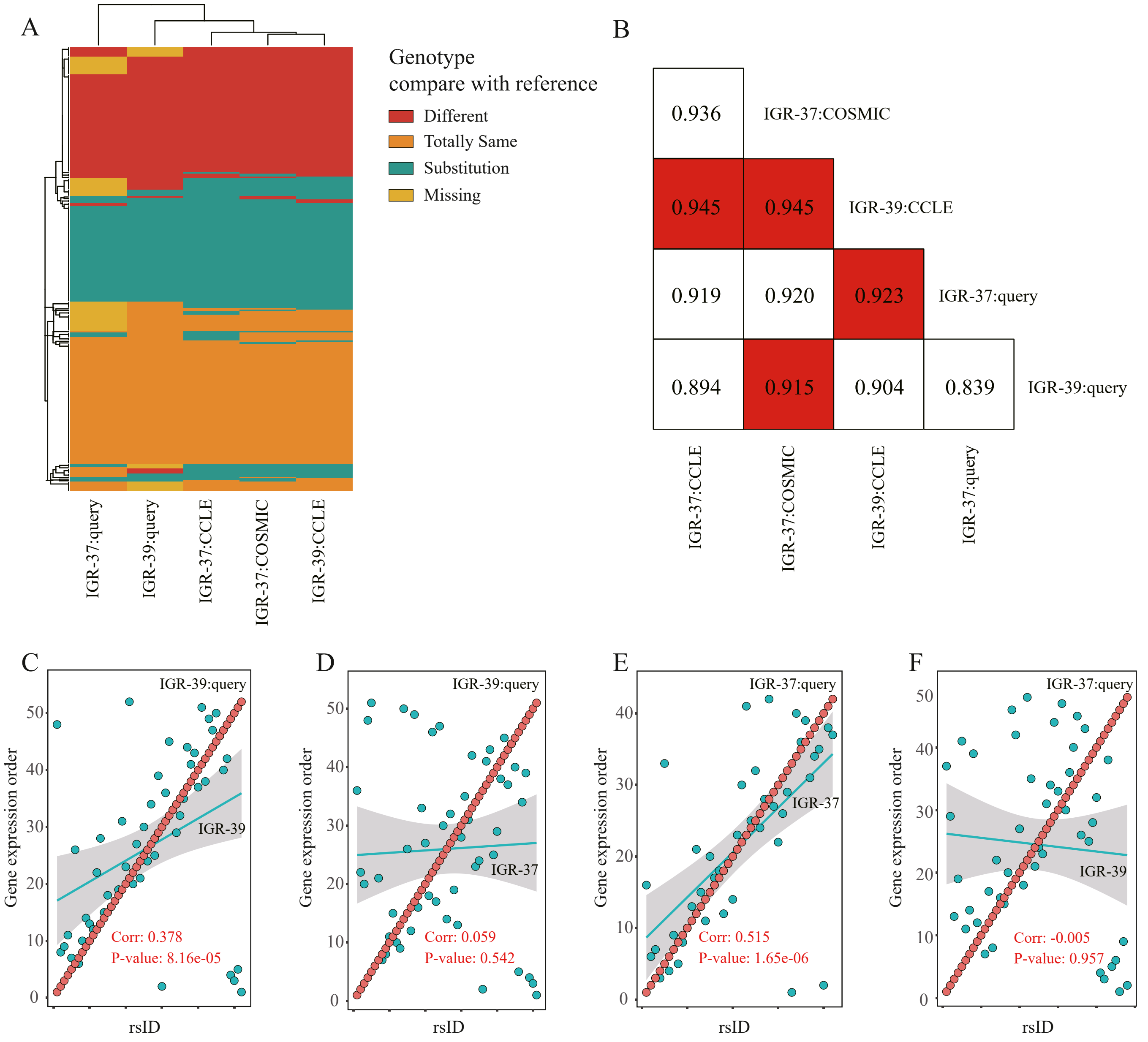


Fig. 3. Authentication of homologous CCLs IGR-37 and IGR-39 from the same individual using SFU sampling. A) shows the high similarity between related cell lines, such as IGR-37 and IGR-39, which proves the limitation of using an SNP-based method to authenticate cell lines with close relationships. B) The cosine similarity between these CCLs. The best match of the four CCLs is marked with orange. The similarity between almost each cell line pair is around 0.9 and disordered, indicating SNV information's need to provide more details to locate homologous cell lines accurately. C) to F) shows that SFU sampling in LKG enables CCLHunter to accurately locate consanguineous cell lines from the same individual.

refined 436 SNPs does not differ significantly from the method using all SNPs, suggesting that using refined SNPs is sufficient to achieve LKG-level authentication. At the same time, the expression-based method tended to misclassify cell lines outside the LKG. All software incorrectly identified cell lines as primCCL within the same LKG when they should have been postCCL. The above analysis highlights the necessity of combining SNP and expression-based methods.

3.5. Comparison with other approaches

To comprehensively evaluate the performance of CCLHunter, we compared it with three mainstream software, CEL-ID, Uniquorn and CCLA, for CCLs authentication using RNA-seq data. CCLHunter only requires alignment information, whereas the other methods have various prerequisites, such as variant calling or specific data filtering. CCLA and CCLHunter offer web servers for quick authentication and exploration of cell lineage relationships. In addition to those cell lines with obvious recognition, we also tested the ability of this software to authenticate easily confused CCLs from postCCLs (Supplementary Table 4). The evaluation results indicate that all approaches could effectively authenticate cell lines with distinct genetics (primCCLs). However, their performance in postCCL authentication could have been more consistent (Table 2). CCLHunter emerged as the featured software, achieving the highest accuracy of 89.28% (50/56) in complete test sets and 100% (17/17) in RNA-seq test sets, outperforming the other software. The SNP-based approaches, Uniquorn and CEL-ID, tend to have difficulties correctly identifying postCCLs within the same lineage, ranging from 64.71% (11/17) to 76.47% (12/17) accuracy at the exact post-derived level. Although the expression-based approach of CCLA does not have this same error tendency, its accuracy in postCCL authentication was still the same as the other two SNP-based approaches (76.47%, 12/17). The failed authenticated cell lines were nearly identical (PLB-985, U-138MG, IGR-39, and OPM-1), consistent with previous tests (Supplementary Table 4).

3.6. CCL authentication and visualization via webserver

We provide two versions of CCLHunter to adapt to different scenarios: web server and standalone program. The standalone program can simultaneously authenticate large cell line data in batch mode. In addition to providing authentication results, the web server can display various cell line information and authentication details by offering sorted BAM or JSON files outputted by the standalone version.

Our web server offers four functional modules: Browse, Task submission, Download, and Documentation (Fig. 4). The browse page provides a detailed description of each CCL recorded in CCLHunter, including the accession name, the original individual information, the reference refined SNP barcodes, and the publications. Task submission is designed to help users annotate and visualize the output of the CCLHunter standalone program. If users are unfamiliar with bioinformatics applications, they can run CCLHunter online by following the guidance provided in the task submission module. The detailed authorization results, including the similarity of the refined SNP barcode and the SFU rank plots, are displayed on the candidate report page. Subsequently, users can download the standalone program and related data on the download module. Finally, CCLHunter provides a user-friendly document containing detailed manuals for standalone programs and web servers.

4. Discussion

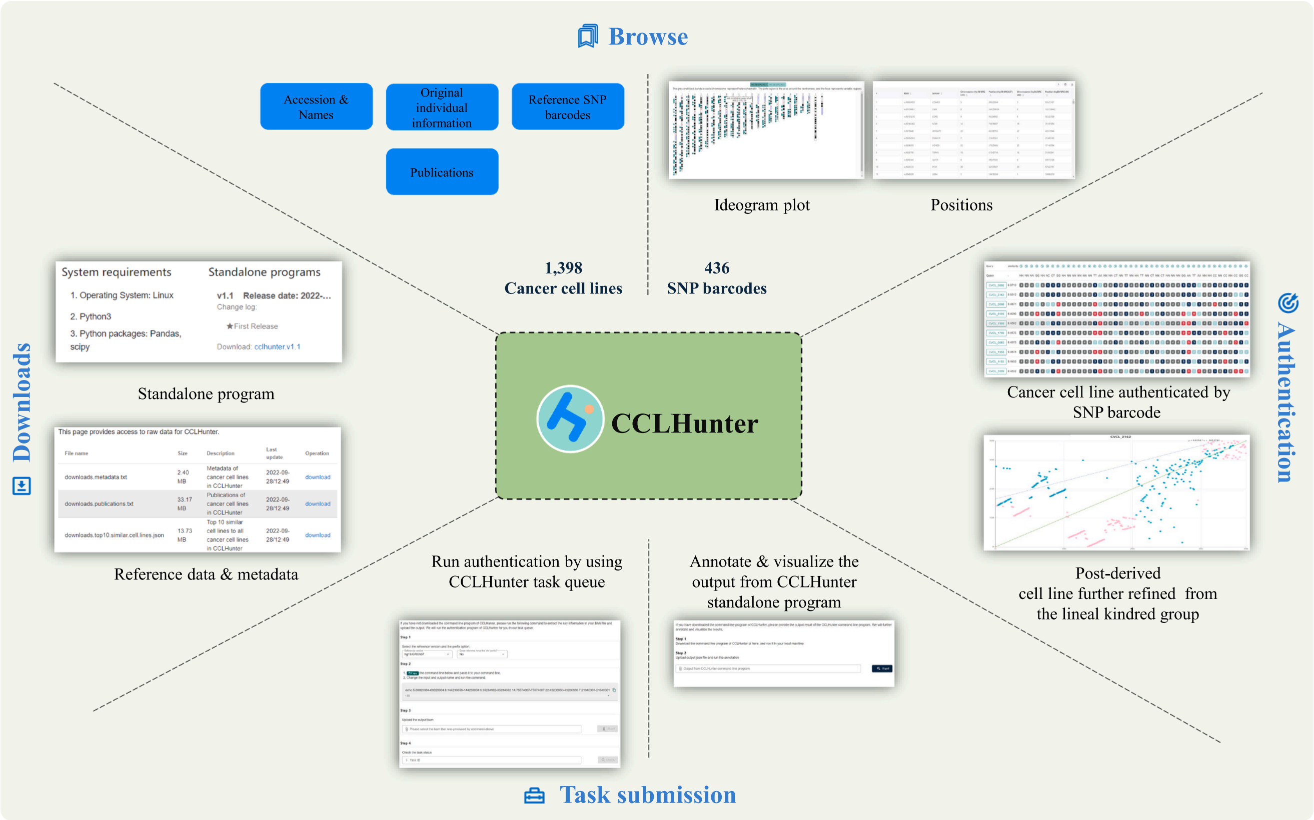
Cancer cell lines offer an almost unlimited source of material for experimental subjects and retain most of the genetic information from the original tissue, making them invaluable tools for cancer research. However, due to issues related to culture, misidentification, and classification, researchers are increasingly aware of the need to ensure the accuracy of cell lines in their experiments [6,8,9,40,41]. We proposed a new method called CCLHunter that can accurately authenticate CCLs from close relatives or even the same individual named postCCLs. The STR-based method serves as the gold standard for CCL authentication. With the maturation of high-throughput sequencing technology, RNA-seq has become an indispensable method for cancer cell line research. However, it can be challenging to authenticate CCLs using RNA-seq data, mainly because not all STRs are involved in transcription. Our method provides another option for cancer cell line authentication, enabling users to assess public data at scale without additional effort on cell line identification. Nonetheless, well-established STR-based authentication methods remain indispensable for large cell line repositories like ATCC and DSMZ [42,43], despite some research pointing out certain limitations in STR-based authentication methods [44–47].

We refined the number of SNPs used from nearly one million to 436. Since the refined SNPs are remarkably stable, we can confirm the genotypes by the nucleotide distribution of the site rather than relying on a statistical model generated by the variant caller. However, whether using refined SNP or other existing methods, it shows a high error rate for post-derived cell lines. It is potentially because the postCCLs in the same LKG often have direct lineage relationships (e.g., U-138MG and U-118MG), or are derived from the same individual (e.g., IGR-37 and IGR-39), or even the same cell line (e.g., HL-60 and PLB-985), which makes them have similar nucleotide polymorphisms and gene expression

Table 2
The comparison of CCLHunter with other approaches.

	Uniquorn	CEL-ID	CCLA	CCLHunter
Implementation	R package	R package	webserver	Python & webserver
Sequencing type	DNA or RNA-seq	RNA-seq	RNA-seq or microarray	RNA-seq
Input data format	vcf	vcf	expression matrix	bam
Evidence	SNP	SNP	expression specificity	SNP & expression specificity & kindred topology
Precondition	variant calling+ alignment	alignment + variant calling	alignment /quantification	alignment
Dependency	GATK/freebayes	varscan2	-	-
Preprocessing	need to download additional train sets	need specific format to filter from vcf	-	-
# reference CCLs	1516	934	1219	1389
Declared accuracy	96%, 95%	-	96.58%, 92.15%	93.27%
Accuracy in primCCLs	100%	100%	100%	100%
Accuracy in postCCLs	64.71% (11/17)	76.47% (12/17)	76.47% (12/17)	100% (17/17) and 89.27% (50/56) in total sets

* : Due to the data format constraint, Uniquorn, CEL-ID, and CCLA only use the RNA-seq test sets for testing, while CCLHunter uses the CHCC, Cellminer, and RNA-seq test sets at the same time


Browse
Accession & Names
Original individual information
Reference SNP barcodes
Publications
Ideogram plot
Positions
1,398
Cancer cell lines
436
SNP barcodes
CCLHunter
Downloads
Standalone program
Reference data & metadata
Authentication
Cancer cell line authenticated by SNP barcode
Post-derived cell line further refined from the lineal kindred group
Run authentication by using CCLHunter task queue
Annotate & visualize the output from CCLHunter standalone program
Task submission

System requirements

1. Operating System: Linux
2. Python3
3. Python packages: Pandas, scipy

Standalone programs

v1.1 Release date: 2022-...
Change log:
★First Release
Download: cclhunter.v1.1

This page provides access to raw data for CCLHunter.

File name	Size	Description	Last update	Operation
downloads.metadata.txt	2.40 MB	Metadata of cancer cell lines in CCLHunter	2022-09-28/12:49	download
downloads.publications.txt	33.17 MB	Publications of cancer cell lines in CCLHunter	2022-09-28/12:49	download
downloads.top10.similar.cell.lines.json	13.73 MB	Top 10 similar cell lines to all cancer cell lines in CCLHunter	2022-09-28/12:49	download

If you have not downloaded the command line program of CCLHunter, please run the following command to extract the key information in your BAM file and upload the output. We will run the authentication program of CCLHunter for you in our task queue.

Step 1
Select the reference version and the prefix option.

Step 2
1. Copy the command line below and paste it to your command line.
2. Change the input and output name and run the command.

Step 3
Upload the output bam

Step 4
Check the task status

If you have downloaded the command line program of CCLHunter, please provide the output result of the CCLHunter command line program. We will further annotate and visualize the results.

Step 1
Download the command line program of CCLHunter at here, and run it in your local machine.

Step 2
Upload output json file and run the annotation

Fig. 4. Function model of CCLHunter in webserver.

patterns. As a result, using single evidence previously can only authenticate cell lines that may contain the same LKG but cannot be accurate to specific CCL. Through mutual testing of datasets in the reference, we consider 0.65 to be a reasonable value for distinguishing the credibility of authenticated cell lines. For cell lines with a SNP similarity score of less than 0.65, you should verify whether your expected cell line exists in our database. To overcome the resolution problem of the postCCL, we only used the SNP-based method to reduce the candidate set to an LKG and then finally authenticate the cell line through the expression profile. In LKG, hot genes are dynamically selected to form SFU, finally used for cell line authentication. Then, the candidate is considered the gene set with the most stable expression and the largest difference as far as possible in the paired sample set of comparison [48].

However, there should be certain limitations to our approach. Firstly, we must extract nucleotide and depth information of 436 positions simultaneously in bam, so CCLHunter currently only supports RNA-seq data. Additionally, expression changes under very different conditions can lead to different results, which is a limitation of all expression-based approaches. Secondly, although about 19% of CCLs are postCCL, only some cell lines are documented by CCLE and COSMIC, and CCLHunter needs access to them. Therefore, CCLHunter will only conduct cell line authentication based on SNP evidence. In future updates, we plan to collect more cell line data from different datasets to make our reference set more representative. Simultaneously, we will also enrich various meta-information about cell lines through data retrieval to help reproducibility and standardization of research.

5. Conclusion

To address the urgent need for accurate authentication of cancer cell lines, we have developed a new method called CCLHunter. It greatly utilizes and effectively leverages multidimensional data encompassing single nucleotide polymorphisms, expression profiles, and kindred topology to enhance the accuracy of cell line authentication, thereby circumventing the limitations of being based on a single piece of information. The evaluation results unequivocally demonstrate that our method is considerably superior to existing and widely used methods for identifying cancer cell lines, particularly consanguineous ones. This approach can potentially standardize biomedical research and enhance the reproducibility of results in cancer research.

CRediT authorship contribution statement

C.F.B and X.C.Z: data collection, bioinformatics analysis, and manuscript writing; X.C.Z: conceptualization, web server work; C.F.B: methodology; J.L.M, J.Y.Z, Q.H.Q, T.Y.X, Y.L.S, and Z.N: revising; J.F.X and Y.M.B: conceptualization, revising, funding acquisition and supervision. All authors have read and approved the final manuscript.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors declare that they do not use any AI and AI-assisted technologies in the writing process.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors express their gratitude to all participants of this study. The authors would like to thank CCLE, COSMIC, SRA, and Cellosaurus project group colleagues. This study was supported by Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38030400 to J.X. and XDB38030200 to Y.B.]; National Natural Science Foundation of China [32170669 and 31970634 to J.X.]; National Key Research Program of China [2020YFA0907001 to J.X.]; the Youth Innovation Promotion Association of the Chinese Academy of Sciences [2022098 to J.Z.].

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.09.040.

References

[1] Sharma SV, Haber DA, Settleman J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. Nat Rev Cancer 2010;10:241–53.
[2] Wilding JL, Bodmer WF. Cancer cell lines for drug discovery and development. Cancer Res 2014;74:2377–84.
[3] Gillet JP, Varma S, Gottesman MM. The clinical relevance of cancer cell lines. J Natl Cancer Inst 2013;105:452–8.
[4] Capes-Davis A, Theodosopoulos G, Atkin I, et al. Check your cultures! A list of cross-contaminated or misidentified cell lines. Int J Cancer 2010;127:1–8.
[5] Korch C, Varella-Garcia M. Tackling the human cell line and tissue misidentification problem is needed for reproducible biomedical research. Adv Mol Pathol 2018;1:209–28. e236.
[6] Horbach S, Halffman W. The ghosts of HeLa: how cell line misidentification contaminates the scientific literature. PLoS ONE 2017;12:e0186281.
[7] Strong MJ, Baddoo M, Nanbo A, et al. Comprehensive high-throughput RNA sequencing analysis reveals contamination of multiple nasopharyngeal carcinoma cell lines with HeLa cell genomes. J Virol 2014;88:10696–704.
[8] ATCC. Cell line authentication publication requirements. https://www.atcc.org/the-science/authentication/cell-line-authentication-publication-requirements.
[9] Identity crisis, Nature 2009;457:935–936.
[10] Araujo SB, Patricio GF, Simoni IC, et al. Isoenzyme and molecular approach for authenticating and monitoring of animal cell lines. Acad Bras Cienc 2019;91:e20180487.
[11] Boegel S, Lower M, Bukur T, et al. A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. Oncoimmunology 2014;3:e954893.
[12] Dirks WG, Faehnrich S, Estella IA, et al. Short tandem repeat DNA typing provides an international reference standard for authentication of human cell lines. ALTEX 2005;22:103–9.
[13] Mohammad TA, Tsai YS, Ameer S, et al. CeL-ID: cell line identification using RNA-seq data. BMC Genom 2019;20:81.
[14] Fasterius E, Raso C, Kennedy S, et al. A novel RNA sequencing data analysis method for cell line authentication. PLoS ONE 2017;12:e0171435.
[15] Zhang Z, Hernandez K, Savage J, et al. Uniform genomic data analysis in the NCI Genomic Data Commons. Nat Commun 2021;12:1226.
[16] Fanfani V, Citi L, Harris AL, et al. The Landscape of the Heritable Cancer Genome. Cancer Res 2021;81:2588–99.
[17] Zhang Q, Luo M, Liu CJ, et al. CCLA: an accurate method and web server for cancer cell line authentication using gene expression profiles. Brief Bioinform 2021;22:bbaa093.
[18] Bairoch A. The cellosaurus, a cell-line knowledge resource. J Biomol Tech 2018;29:25–38.
[19] Ghandi, Huang M, Jane FW, Valbuena J, et al. Next-generation characterization of the cancer cell line encyclopedia. Nature 2019;569:503–8.
[20] Tate JG, Bamford S, Jubb HC, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 2019;47:D941–7.
[21] Affymetrix®. Genome-Wide Human SNP Array 6.0. https://www.affymetrix.com/support/downloads/package_inserts/genomewide_snp6_insert.pdf.
[22] stuff N. NCBI ALFAOpen-Access to dbGaP Aggregated Allele Frequency for Variant Interpretation.https://ncbiinsights.ncbi.nlm.nih.gov/2020/03/26/alfa/.
[23] Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2011;39:D38–51.
[24] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27:573–80.
[25] Barnes MR, Breen G. Genetic Variation: Methods and Protocols. Totowa, NJ: Humana; 2010.
[26] Price AL, Weale ME, Patterson N, et al. Long-range LD can confound genome scans in admixed populations. Am J Hum Genet 2008;83:132–5. author reply 135-139.
[27] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
[28] Klijn C, Durinck S, Stawiski EW, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol 2015;33:306–12.
[29] Shankavaram UT, Varma S, Kane D, et al. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genom 2009;10:277.
[30] NIH. SRA-Toolkit. https://hpc.nih.gov/apps/sratoolkit.html.
[31] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.
[32] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923–30.
[33] Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568–76.
[34] Erik Garrison G.M. Haplotype-based variant detection from short-read sequencing, arXiv 2012:1207.3907.
[35] Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 2017:201178.
[36] Shaffer L.G., McGowan-Jordan J., Schmid M. ISCN 2013: an international system for human cytogenetic nomenclature (2013): recommendations of the International Standing Committee on Human Cytogenetic Nomenclature. 2005.
[37] Reinhold WC, Sunshine M, Liu H, et al. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res 2012;72:3499–511.
[38] Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 2014;30:301–4.
[39] Hansen KD, Wu Z, Irizarry RA, et al. Sequencing technology does not eliminate biological variability. Nat Biotechnol 2011;29:572–3.
[40] Drexler HG, Dirks WG, Matsuo Y, et al. False leukemia-lymphoma cell lines: an update on over 500 cell lines. Leukemia 2003;17:416–26.
[41] American Type Culture Collection Standards Development Organization Workgroup ASN. Cell line misidentification: the beginning of the end, Nat Rev Cancer 2010;10:441–448.
[42] ATCC. ATCC STR search profile. https://www.atcc.org/search-str-database.
[43] Koblitz J, Dirks WG, Eberth S, et al. DSMZCellDive: Diving into high-throughput cell line data. F1000Res 2022;11:420.
[44] Bourré L. Cancer Cell Line Authentication. https://blog.crownbio.com/cancer-cell-line-authentication.
[45] Evrard C, Tachon G, Randrian V, et al. Microsatellite Instability: Diagnosis, Heterogeneity, Discordance, and Clinical Impact in Colorectal Cancer. Cancers (Basel) 2019;11.
[46] Zhang P, Zhu Y, Li Y, et al. Forensic evaluation of STR typing reliability in lung cancer. Leg Med (Tokyo) 2018;30:38–41.
[47] Chen A, Xiong L, Qu Y, et al. Opportunity of next-generation sequencing-based short tandem repeat system for tumor source identification. Front Oncol 2022;12:800028.
[48] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40.