frontiers | Frontiers in Genetics

TYPE Technology and Code
PUBLISHED 11 August 2022
DOI 10.3389/fgene.2022.956781

CCAS: One-stop and comprehensive annotation system for individual cancer genome at multi-omics level

Xinchang Zheng[1,2†], Wenting Zong[1,2,3†], Zhaohua Li[1,2,3†], Yingke Ma[1,2], Yanling Sun[1,2], Zhuang Xiong[1,2,3], Song Wu[1,2,3], Fei Yang[1,2,3], Wei Zhao[1,2,3], Congfan Bu[1,2], Zhenglin Du[1,2], Jingfa Xiao[1,2,3]* and Yiming Bao[1,2,3]*

[1]National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, Beijing, China, [2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, Beijing, China, [3]University of Chinese Academy of Sciences, Beijing, China

OPEN ACCESS

EDITED BY
Lixin Cheng,
Jinan University, China

REVIEWED BY
Dechao Tian,
School of Public Health, Sun Yat-sen University, Shenzhen Campus, China
Jing Qin,
School of Pharmaceuticals Sciences (Shenzhen), Sun Yat-sen University, China

*CORRESPONDENCE
Jingfa Xiao,
xiaojingfa@big.ac.cn
Yiming Bao,
baoym@big.ac.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 30 May 2022
ACCEPTED 18 July 2022
PUBLISHED 11 August 2022

CITATION
Zheng X, Zong W, Li Z, Ma Y, Sun Y, Xiong Z, Wu S, Yang F, Zhao W, Bu C, Du Z, Xiao J and Bao Y (2022), CCAS: One-stop and comprehensive annotation system for individual cancer genome at multi-omics level.
Front. Genet. 13:956781.
doi: 10.3389/fgene.2022.956781

Due to the explosion of cancer genome data and the urgent needs for cancer treatment, it is becoming increasingly important and necessary to easily and timely analyze and annotate cancer genomes. However, tumor heterogeneity is recognized as a serious barrier to annotate cancer genomes at the individual patient level. In addition, the interpretation and analysis of cancer multi-omics data rely heavily on existing database resources that are often located in different data centers or research institutions, which poses a huge challenge for data parsing. Here we present CCAS (Cancer genome Consensus Annotation System, https://ngdc.cncb.ac.cn/ccas/#/home), a one-stop and comprehensive annotation system for the individual patient at multi-omics level. CCAS integrates 20 widely recognized resources in the field to support data annotation of 10 categories of cancers covering 395 subtypes. Data from each resource are manually curated and standardized by using ontology frameworks. CCAS accepts data on single nucleotide variant/insertion or deletion, expression, copy number variation, and methylation level as input files to build a consensus annotation. Outputs are arranged in the forms of tables or figures and can be searched, sorted, and downloaded. Expanded panels with additional information are used for conciseness, and most figures are interactive to show additional information. Moreover, CCAS offers multidimensional annotation information, including mutation signature pattern, gene set enrichment analysis, pathways and clinical trial related information. These are helpful for intuitively understanding the molecular mechanisms of tumors and discovering key functional genes.

KEYWORDS

comprehensive annotation, multi-omics, individual cancer patient, databases integration, web server

1 Introduction

Cancer is one of the leading causes of human death all over the world (Jemal et al., 2007; Ferlay et al., 2013; Torre et al., 2016). The occurrence and development of each cancer is driven by a unique set of abnormalities in its genome (Stratton et al., 2009; Garraway and Lander, 2013; Birkbak and McGranahan, 2020). Therefore, dissecting changes in the cancer genome at the multi-omics level could significantly improve our understanding of the molecular mechanisms of tumorigenesis and help the development of new treatments (Tebani et al., 2016; Olivier et al., 2019). To date, a series of large cancer genome sequencing projects have been launched as the next generation sequencing (NGS) technology becomes more and more widely used in cancer researches (Cerami et al., 2012; Gao et al., 2013; ITP-CAoWG, Consortium, 2020). Genome annotation, as an effective approach, provides a comprehensive perspective of cancers' abnormalities by using multi-omics data. However, there are still a number of challenges that need to be addressed. Firstly, inter-tumor heterogeneity is increasingly recognized as a serious barrier in annotating cancer genome at the individual patient level. Secondly, comprehensive annotation relies heavily on existing data resources that are often located in different data centers or research institutions, which poses a huge challenge to integrate those resources. Finally, additional essential knowledge such as clinical trials, drug interactions, literature of the abnormalities are needed because they have far-reaching significance for understanding tumors.

In order to make cancer genome annotation convenient and efficient, several tools, online databases, and web servers have been developed over the past decades. ANNOVAR (Wang et al., 2010), Ensembl-VEP (McLaren et al., 2016), and SnpEff (Cingolani et al., 2012) were developed as annotation tools for variants function based on population frequencies in normal or disease cohorts, as well as damage predictions at genomic level. PCAWG-Scout (Goldman et al., 2020a), UCSC Xena (Goldman et al., 2020b), and OpenCRAVAT (Pagel et al., 2020) were designed for complex visualization and analysis services of large scale cancer datasets. PCGR (Nakken et al., 2018), GenomeChronicler (Guerra-Assuncao et al., 2020), and PORI (Reisle et al., 2022) were developed for cancer genome annotation at the individual patient level, providing many useful functions, such as mutation signature analysis, mutation burden analysis, drug interactions, as well as clinical trials analysis. However, these tools are more focused on parsing genomic level data, while lacking comprehensive annotations based on the integration of multiple cancer-related databases, or have limitations in data analysis at the individual patient level.

Here, we present Cancer genome Consensus Annotation System (CCAS), which is a comprehensive annotation server for individual cancer genome at multi-omics level. CCAS builds two ontology frameworks and integrates 20 data resources, which are commonly used in cancer researches. Information and knowledge in CCAS can be classified into 6 aspects: genomics, disease, normal/cancer cohorts, clinical trials, literature, and drug interactions (Supplementary Table S1), enabling comprehensive annotation at the individual patient level. The integration of these information allows CCAS to annotate not well studied abnormalities in patient-specific cancer subtypes by transferring knowledge across cancer subtypes and databases. Moreover, CCAS uses a two-step process to identify key functional genes that significantly change in the individual patient and play important roles in tumorigenesis. Furthermore, CCAS offers analysis including mutation signature pattern, gene set enrichment analysis. Overall, CCAS is aimed at annotating cancer genome precisely and effectively in the individual patient level.

2 Materials and methods

2.1 Data collection

To provide high-quality annotation results, CCAS integrated 20 resources (Supplementary Table S1) to build the annotation results at 6 aspects including genomics, disease, normal/cancer cohorts, clinical trials, literature, and drug interactions (Figure 1). Genomics aspect data were collected from Ensembl (Zerbino et al., 2018), dbNSFP (Liu et al., 2020), dbSNP (Sherry et al., 2001), HGNC (Tweedie et al., 2021), and UniProtKB (UniProt, 2021). Those resources provided knowledge of gene descriptions, IDs (gene IDs, protein IDs, and variant IDs) in different databases, protein function descriptions, and protein damage predictions. For the disease aspect, data were integrated from COSMIC (Tate et al., 2019), Disease Ontology (Schriml et al., 2022), MeSH (Baumann, 2016), single sample GSEA (ssGSEA) (Subramanian et al., 2005) and Reactome (Gillespie et al., 2022). Those data provided insights of patient's cancer subtype including disease description, disease synonymous names, disease ontology name, and related pathways. Besides, the mutation signature analysis and ssGSEA analysis were used to reveal underlying biological processes of the patient. For the normal/cancer cohorts aspect, ExAC (Karczewski et al., 2017), gnomAD (Karczewski et al., 2020), intOGen (Martinez-Jimenez et al., 2020), the 1000 Genomes Project (Genomes Project et al., 2010), Cancer Hotspots V2 (Chang et al., 2016; Chang et al., 2018), Cancer Genome Interpreter (Tamborero et al., 2018) were collected. Those data provided the frequencies of variants both in cancer and normal cohorts. The ClinicalTrials.gov (https://clinicaltrials.gov/ct2/home) database was integrated into Clinical trial aspect. Those data provided related clinical trials information of patient's cancer subtype including study design, eligibility criteria, and intervention. The Literature aspect was built mainly in the aid of the CancerMine (Lever et al., 2019) database, providing relationships between genes and cancer subtypes. DGIdb (Freshour et al., 2021) and Open Target

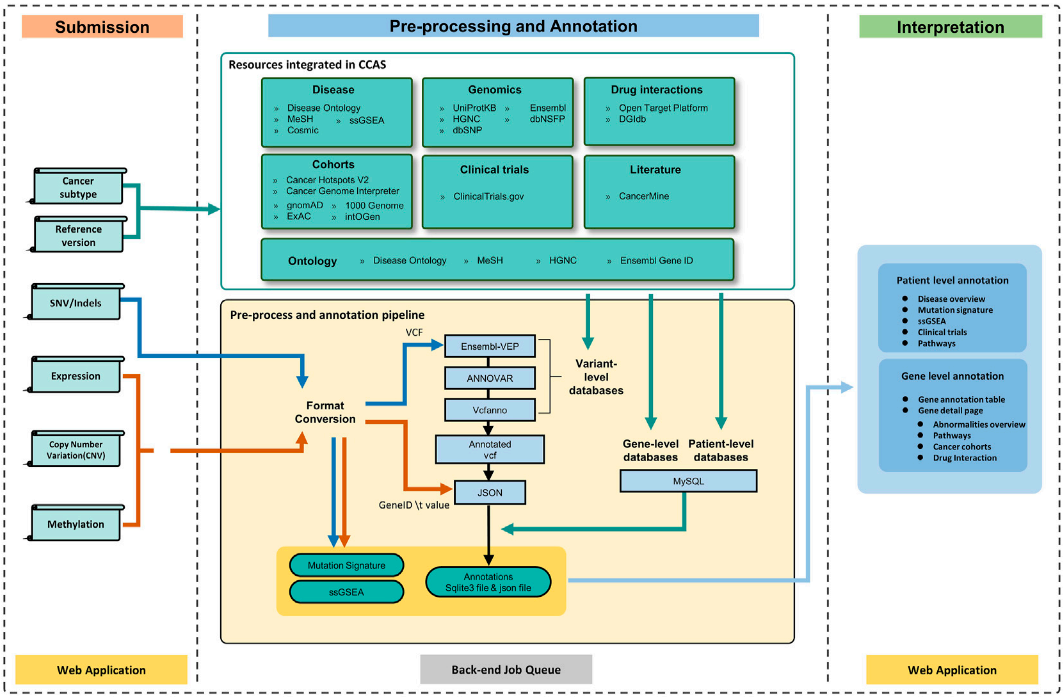


FIGURE 1
The workflow of CCAS can be divided into three modules: Submission, Pre-processing and Annotation, and Interpretation. After the user submits data to CCAS, CCAS first converts the format of the files. SNV/Indels data will be converted to VCF format and other data types will be converted to “Gene ID \t Value” format. CCAS will then annotate the patient data using the integrated data sources at multiple levels. Mutation Signature and ssGSEA calculations are also performed. The annotation results are stored in sqlite3 database (a single file database) and json file. CCAS has built user-friendly interface to help users navigate and interpret the annotation results, enabling efficient identification of key functional genes at the individual patient level.

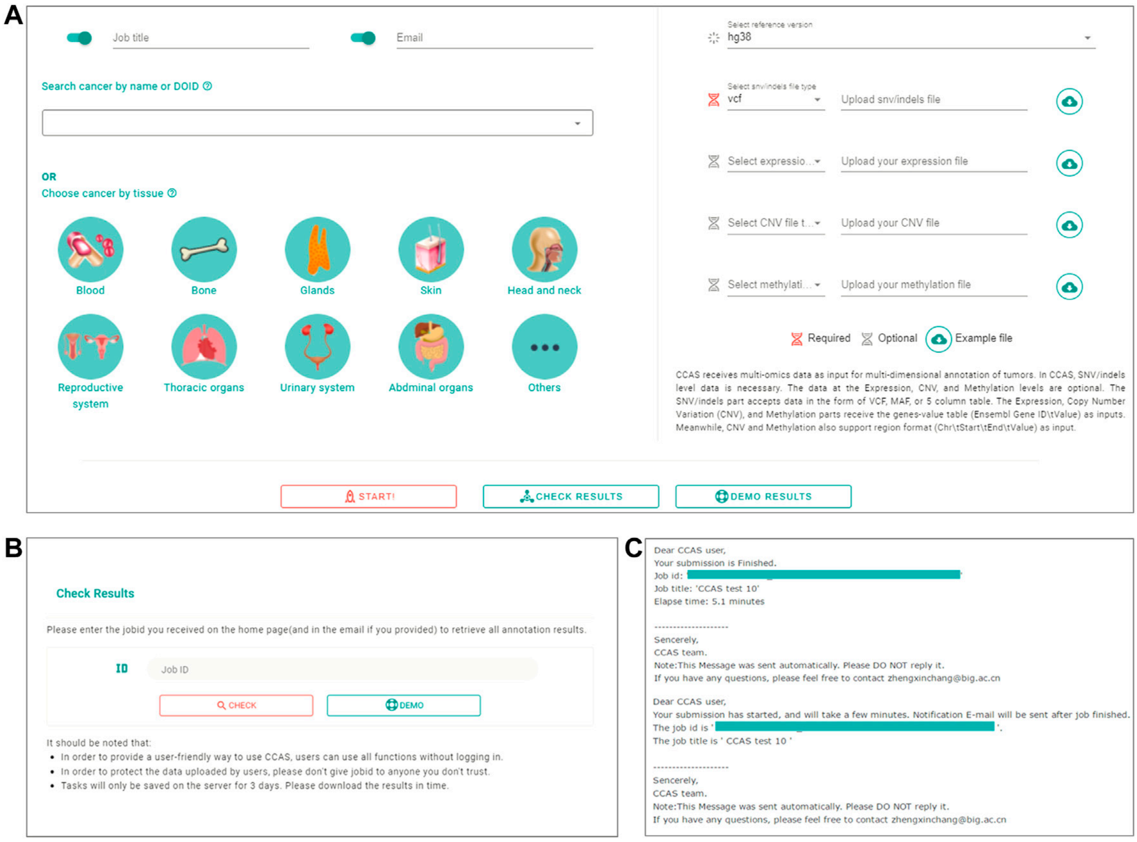
Platform (Koscielny et al., 2017; Carvalho-Silva et al., 2019) were used to build the Drug interactions aspect, providing potential drug interactions of abnormalities. Disease Ontology (Schriml et al., 2022), MeSH (Baumann, 2016), Ensembl (Zerbino et al., 2018), and HGNC (Tweedie et al., 2021) databases were used to build the ontology frameworks which were used to integrate data from multiple resources. The detailed description of the databases can be found at CCAS documentation (https://ngdc.cncb.ac.cn/ccas/docs/#/, 2.3 Data sources integrated into CCAS).

2.2 Ontology frameworks construction

To integrate multiple data sources, we built two ontology frameworks respectively: ontology of cancers and genes. For the ontology of cancers, we downloaded data from Disease Ontology and parsed them by the Pronto package (https://pypi.org/project/pronto/). Cancer subtypes with MeSH IDs were recursively extracted starting from the node “cancer” (DOID: 162). Ultimately, 395 cancer subtypes were integrated into the CCAS. To make it easier for users to specify cancer types, all cancer subtypes were manually classified into 10 groups according to the human tissue type. MeSH terms corresponding to each cancer were fetched by using NCBI’s E-utilities. For ontology of genes, we retrieved gene IDs from the Ensembl database (release version 104) and converted them to NCBI gene IDs and UCSC gene IDs by using the HGNC database.

2.3 Data standardization and integration

Human protein records were extracted from UniProtKB’s XML file using Python library of BeautifulSoup4. Data in Open Target Platform, DGIdb, CancerMine, and intOGen were downloaded in tabular format. Pathway information along with diagrams were extracted from the Reactome database. Data from Cancer Hotspots V2 were converted into the VCF format and indexed by Tabix (Li, 2011) after sorting by chromosomes. For the ClinicalTrials.gov database, NCT ID (Clinical trial ID) and other metadata were extracted by the Python XML module. After that, clinical trials with drugs were retained. The MeSH terms in clinical trial records were linked to Disease Ontology by MeSH IDs. Data in Ensembl, dbNSFP, 1000 Genomes Project, ExAC, gnomAD, and dbSNP were



CCAS receives multi-omics data as input for multi-dimensional annotation of tumors. In CCAS, SNV/indels level data is necessary. The data at the Expression, CNV, and Methylation levels are optional. The SNV/indels part accepts data in the form of VCF, MAF, or 5 column table. The Expression, Copy Number Variation (CNV), and Methylation parts receive the genes-value table (Ensembl Gene ID\tValue) as inputs. Meanwhile, CNV and Methylation also support region format (Chr\tStart\tEnd\tValue) as input.

Please enter the jobid you received on the home page(and in the email if you provided) to retrieve all annotation results.

It should be noted that:

- In order to provide a user-friendly way to use CCAS, users can use all functions without logging in.
- In order to protect the data uploaded by users, please don't give jobid to anyone you don't trust.
- Tasks will only be saved on the server for 3 days. Please download the results in time.

Dear CCAS user,
Your submission is Finished.
Job id: '[redacted]'
Job title: 'CCAS test 10'
Elapse time: 5.1 minutes

Sencerely,
CCAS team.
Note:This Message was sent automatically. Please DO NOT reply it.
If you have any questions, please feel free to contact zhengxinchang@big.ac.cn

Dear CCAS user,
Your submission has started, and will take a few minutes. Notification E-mail will be sent after job finished.
The job id is '[redacted]'.
The job title is ' CCAS test 10 '

Sencerely,
CCAS team.
Note:This Message was sent automatically. Please DO NOT reply it.
If you have any questions, please feel free to contact zhengxinchang@big.ac.cn

FIGURE 2
The workflow of submitting data and checking job progress in CCAS. **(A)** Submitting portal at home page. CCAS receives data at multi-omics level including SNV/Indels (required), Expression, Copy Number Variation (CNV), and Methylation along with, job title, notification email, the Disease Ontology ID, and reference version. **(B)** Check results page in CCAS. On this page the user can check the progress of the job. **(C)** Notification emails sent to users at the start of a job and at the end of a job.

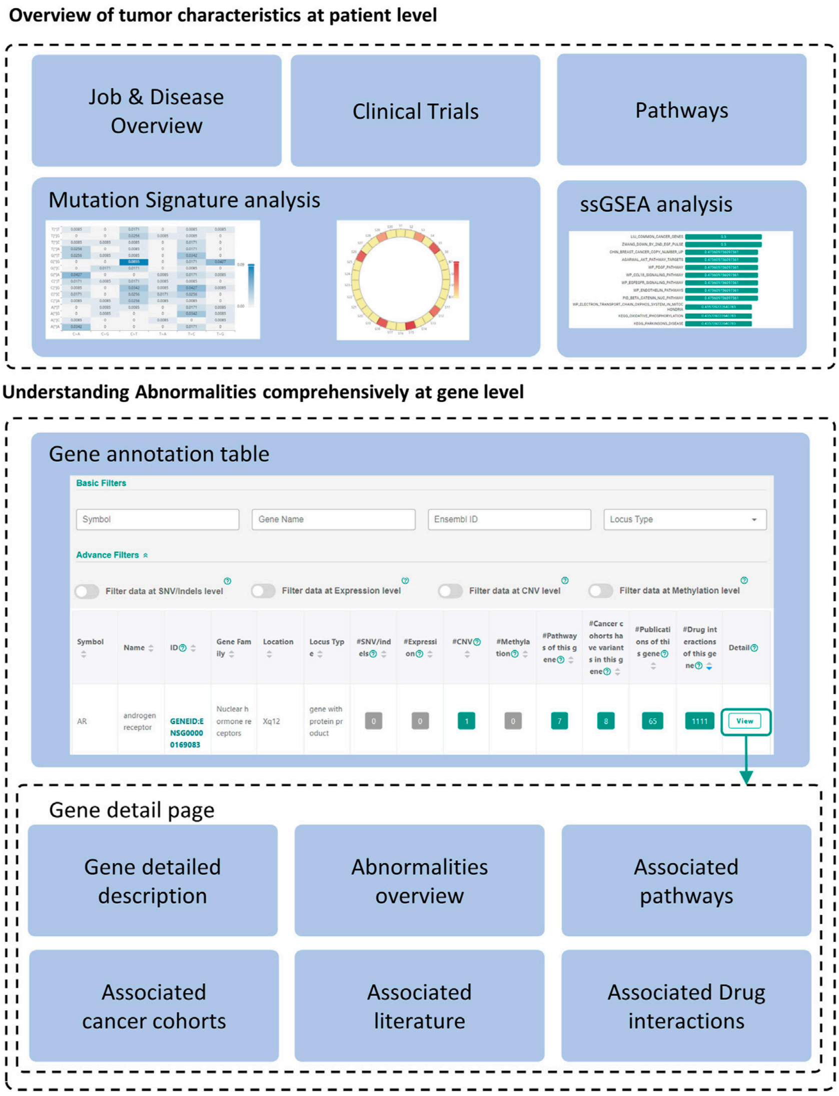
retrieved by using Ensembl-VEP (McLaren et al., 2016) and ANNOVAR (Wang et al., 2010). CrossMap tool was used to convert data with different genome coordinates (Zhao et al., 2014).

2.4 Overall workflow of CCAS

The overall workflow of CCAS can be divided into three modules: Submission, Pre-processing and Annotation, and Interpretation (Figure 1).

The Submission module was used to collect user's uploaded data as well as the reference genome version and cancer subtype of the patient (Figure 1). A submission portal was built to provide user-friendly interface at the home page of the web application. Cancer subtype can be selected at the left part of portal. The multi-omics data files along with reference genome version can be uploaded at the right part of the submission portal (Figure 2A). The mandatory inputs were reference genome version, cancer subtype, and the data file in SNV/Indels level.

The Pre-processing and Annotation module consists of four main parts: format conversion, variant level annotation, gene level annotation, and patient level annotation (Figure 1). At the format conversion part, SNV/Indels level files including mutation annotation format (MAF) or 5 columns tabular (5coltsv) format were converted into the VCF format. "chr" prefix was added if it did not exist. For data in expression, CNV and methylation level, files in region format were converted to 2 columns table ("Ensembl Gene ID \t Value") by using bedtools (https://bedtools.readthedocs.io/) (Figure 1). At the variant level annotation part, data in SNV/Indels level were annotated. the pipeline integrated ANNOVAR, Ensembl-VEP, Vcfanno (Pedersen et al., 2016), vt-normalize (Tan et al., 2015), DeconstructSigs (Rosenthal et al., 2016), GSVA package (Hanzelmann et al., 2013), and GSEAbase (https://bioconductor.org/packages/GSEABase/) package to conduct the entire annotation. Briefly, vt-normalize was used to normalize the variants in the VCF file, then split multi-allele variants into different records. Then, CCAS used Ensembl-VEP


Overview of tumor characteristics at patient level
Job & Disease Overview
Clinical Trials
Pathways
Mutation Signature analysis
ssGSEA analysis
Understanding Abnormalities comprehensively at gene level
Gene annotation table
Basic Filters
Symbol
Gene Name
Ensembl ID
Locus Type
Advance Filters
Filter data at SNV/Indels level
Filter data at Expression level
Filter data at CNV level
Filter data at Methylation level
Gene detail page
Gene detailed description
Abnormalities overview
Associated pathways
Associated cancer cohorts
Associated literature
Associated Drug interactions

Symbol	Name	ID	Gene Family	Location	Locus Type	#SNV/indels	#Expression	#CNV	#Methylation	#Pathways of this gene	#Cancer cohorts have variants in this gene	#Publications of this gene	#Drug interactions of this gene	Detail
AR	androgen receptor	GENEID:ENSG00000169083	Nuclear hormone receptors	Xq12	gene with protein product	0	0	1	0	7	8	65	1111	View

FIGURE 3
Overview of the annotation results. The annotation results of CCAS consists of patient level annotation and gene level annotation. Gene detailed pages can be viewed by clicking "View" button at the end of each record in the gene annotation table.

to annotate variants with data from the Ensembl database, and used ANNOVAR to annotate the VCF file with dbNSFP, 1000 Genomes Project, ExAC, gnomAD, and dbSNP database. Vcfanno was used to annotate variants with the Cancer Hotspots and Cancer Genome Interpreter database. Output of this part was an annotated VCF file. At the gene level annotation part, annotated VCF and data in other level were converted in to json format. All abnormalities in different level were converted to gene level and were annotated by multiple databases including DGIdb, CancerMine, Reactome, intOGen, Open Target Platform, UniProtKB, and ClinicalTrials.gov. User specified reference genome version was used both in variant level annotation and gene level annotation. At the patient level annotation part, Disease Ontology, ClinicalTrials.gov, and Reactome were used. Briefly, Disease overview information were extracted from Disease Ontology database. Related clinical trials were annotated according to the cancer subtype by using ClinicalTrials.gov database. Pathways information aggregated related pathway of each abnormal gene. Besides, the mutation signature analysis and ssGSEA analysis were performed using DeconstructSigs package, GSVA package, and GSEAbase package (Figure 1). Output of this module had a single sqlite3 database file (https://www.sqlite.org/index.html), an annotated json file, Mutation signature analysis results, and ssGSEA analysis results (Figure 1).

The Interpretation module was used to help users understand the annotation results (Figure 1). The annotation results can be divided into two parts: patient level annotation and gene level annotation (Figures 1, 3). The patient level annotation results included Disease overview, Mutation signature analysis, ssGSEA, Clinical trials and Pathways. The gene level annotation results included a gene annotation table and gene detail pages for each gene. Annotation results from variant level databases were integrated in the gene detail page for each gene. Several filters were built in CCAS to help users to filter abnormalities. Basic filters were used to filter abnormalities by gene symbols, gene names, Ensembl gene IDs and locus types. Advance filters were built to data at different level. For SNV/Indels level data, CCAS provided filters based on the IMPACT value from the Ensembl-VEP tool. For other level data, range filters were developed for filtrations (Figure 1; Supplementary Figure S2A).

2.5 Web server implementation

The web application was compatible with major web browsers, including Firefox and Chrome. CCAS used the front-end and back-end separation mode. The back-end APIs was built using FastAPI (https://fastapi.tiangolo.com/). MySQL relational database was used for data storage. The front-end pages were constructed using Vue.js (https://vuejs.org/index.html) along with Vue-router (https://router.vuejs.org/) and Vuex (https://vuex.vuejs.org/index.html). Axios (http://www.axios-js.com/) was used to send AJAX requests to convey data from the back-end. The whole system was deployed in the Nginx server (http://nginx.org/). CentOS (https://www.centos.org/) was used to host pipelines and web applications. Back-end job queue and annotation pipeline were built by using Python, R and Shell scripts, and running for each user submission.

2.6 Run annotation

Users can submit a job on the home page, fill in the job title and notification email, select the cancer type, upload the file and specify the file type. Submission is started by clicking the “Start” button (Figure 2A). Users can check the progress of the annotation on the Check Results page (Figure 2B). An email notification will be sent to the user at the beginning and the end of the job (Figure 2C). The whole annotation process typically takes around 5–10 min, but depends on the size of the uploaded data.

3 Results

3.1 Glance of the annotation results

The annotation results of CCAS can be divided into two parts: patient level annotation and gene level annotation (Figure 3). The patient level annotation provides a whole picture on patient’s tumor characteristics. The gene level annotation offers a summarizing table combined with filters and gene detail pages for each abnormal gene (Figure 3). Tables can be searched, sorted, and downloaded. Most figures are interactive to show additional information. In a word, users can easily understand the tumor characteristics and screen for key functional genes in the individual patient level by using CCAS.

3.1.1 Overview of tumor characteristics at patient level

Patient level annotation presents the overview of the patient’s tumor. CCAS shows the type of data submitted by the user, the synonymy of the disease and the associated IDs in the “Job & Disease Overview” section. In order to decipher biological processes involved in tumorigenesis, CCAS calculates the mutation signature based on the patient’s SNV profile and compares it with COSMIC mutation signatures in the “Mutation Signatures” section. In addition, CCAS provides the results of the ssGSEA analysis, which helps user to gain insight into the patient’s tumor characteristics from the enriched gene sets in the “ssGSEA” section. Furthermore, CCAS provides disease-related clinical trials to help users understanding the progress of cancer treatment in the “Clinical trials” section. Finally, CCAS provides the pathways consisting of all the abnormal genes in the “Pathways” section (Figure 3; Supplementary Figure S1).

3.1.2 Understanding abnormalities comprehensively at gene level

The gene annotation table provides a detailed view of abnormal gene functions. The left side of the table shows basic information, including gene symbol and gene name. The right side shows the number of annotation hits in patient’s multi-omics data and the resources integrated in CCAS (Figure 3; Supplementary Figure S2B).

Gene detail pages are used to display comprehensive information about genes. Gene basic information, including gene IDs in various databases, and gene functional description are shown on the top of the page (Figure 3; Supplementary Figure S3A). The lower part of the page shows the patient’s abnormalities in the gene (Figure 3; Supplementary Figures S3B–G). Especially, CCAS describes the abnormalities at the SNV/Indels level, including the frequency of variants in normal and cancer populations, and damage predictions. This helps users to gain deep insight into the variants. In addition, CCAS provides pathway information to help users to understand the gene function (Figure 3; Supplementary Figure S3C). Gene frequencies are provided if the gene has been detected in cancer cohorts (Figure 3; Supplementary Figure S3D). The Literature section provides current research status on this gene (Figure 3, Supplementary Figure S3E). Finally, CCAS offers interactions of genes and drugs, which helps users to evaluate whether a gene is targetable (Figure 3, Supplementary Figures S3F,G).

3.2 Identifying key functional genes at individual patient level

Key functional genes not only have significant functional changes in patients' tumors, but also play a key role in tumorigenesis. Because of the heterogeneity between tumors, these key functional genes may be different at the individual patient level and have not been well studied in the current tumor type. CCAS provides complete annotation on those genes by transferring knowledge across cancer subtypes and databases. Based on the CCAS annotation results, users can find key functional genes through a two-step process. Firstly, filters can be used to screen significant functionally changed genes. The basic filters can filter genes by gene symbols, gene names, Ensembl gene IDs and locus types. Advanced filters can be applied to specific data types (SNV/Indels, expression, CNV and methylation) (Figure 3; Supplementary Figure S2A). Secondly, essential genes in tumorigenesis are screened by examining information in associated literature, drug interactions, pathways, and cancer cohorts (Figure 3).

3.3 Case study

To evaluate the performance of CCAS, we carry out a case study for a patient with prostate cancer (DOID:10283). The patient's multi-omics data are downloaded from GDC data portal (https://portal.gdc.cancer.gov/) at SNV/Indels level (MAF format), expression level (tabular format), CNV level (tabular format) and methylation level (tabular format). We perform basic filtering on data at expression level, CNV level and methylation level to simulate input data by users (Details can be found at https://ngdc.cncb.ac.cn/ccas/docs/#/, 5. Case study). The results of the case study can be viewed by clicking the demo button on the home page or the check results page. At the patient level annotation, the mutation signature analysis reveals that the tumor cells may have a deficiency of DNA mismatch repair function. ssGSEA analysis indicates that multiple cancer related pathways harbor abnormalities including the *AKT* pathway and the *PDGF* pathway, which are consistent with previous studies (van der Poel, 2004; Shorning et al., 2020; Shen et al., 2021). At the gene level annotation, by selecting high impact variant at the SNV/Indels level filter above the gene annotation table, users obtain four genes with significant functional alterations: *ARID1A*, *ZFHX3*, *GADL1*, and *ARID2*. Based on the results, *ARID1A* has 2 related pathways, 70 related cancer cohorts, 55 related publications, and 7 related drug interactions. The gene detail page of *ARID1A* shows that *ARID1A* is a subunit of the SWI/SNF chromatin remodeling complex, and plays an important role in changing chromatin structure by altering DNA-histone contacts within a nucleosome in an ATP-dependent manner. Abnormalities occur at SNV/Indels levels (Abnormalities in user's upload data section). Moreover, *ARID1A* is involved in 2 pathways including the *RUNX1* pathway, which plays an important role in the development of leukemia (Pathways section) (Kaisrlikova et al., 2022). The literature section indicates that *ARID1A* is observed in a variety of cancers including bladder cancer (Saito et al., 2018; Cao et al., 2020), ovarian cancer (Kim et al., 2016), liver cancer (Sun et al., 2017) and colon cancer (Mathur et al., 2017; Iftekhar et al., 2021). The Cancer cohorts section also reveals *ARID1A* mutations in multiple cancer subtypes, which is consistent with the Literature section. The Drug interactions section suggests that Atezolizumab is likely to interact with this gene. In summary, we suggest that although *ARID1A* is not frequently mutated and well-studied in prostate cancer, it may be one of the important factors in tumorigenesis of prostate tumors and may act as a potential biomarker for this cancer.

Taken together, we conclude that CCAS provides complete annotation on the individual cancer genome both at patient level and gene level by integrating 20 data resources. Especially, genes which are not frequently mutated and well-studied in the patient's cancer subtype can be well annotated in CCAS.

4 Discussion

Cancer is known as a complex disease and is often driven by abnormalities in key cancer genes that occur in cells at multiple omics levels (Chakraborty et al., 2018; ITP-CAoWG, Consortium, 2020). With the explosion of cancer genome data, cancer genome annotation has become an effective way to uncover the underlying mechanisms of tumorigenesis and help the development of treatment strategies (Tebani et al., 2016; Olivier et al., 2019). However, there are still some challenges to be addressed. Firstly, inter-tumor heterogeneity, as a fundamental characteristic of cancer genome, causes incomplete annotation in individual patients. Abnormalities that play crucial roles in individual patients may have low population frequencies and may not be well studied in the cancer type. Secondly, knowledge which is important for cancer genome annotation is usually deposited in different databases with various data structure. Finally, vital knowledge such as clinical trials, drug interactions, literature is lacking in cancer genome annotation.

Existing tools have been developed to facilitate annotation on cancer genome but have limitations on providing more comprehensive annotation for individual patients at multi-omics level. CCAS is designed to annotate multi-omics data from the individual patient and has the following features: Firstly, CCAS has built two ontology frameworks to integrate resources. To date, CCAS has enrolled 20 widely recognized databases in the field. Secondly, within CCAS, knowledge about normal/cancer cohorts, clinical trials, literature, and drug interactions are integrated, providing deep insights into patient's tumor characteristics. Thirdly, genes which are not frequently mutated and well-studied in one cancer subtype can be well annotated in CCAS by transferring knowledge from other cancer subtypes. This can help users to understand deeply of heterogenous cancer genomes with the aid of existing knowledge across cancer subtypes. Moreover, CCAS provides a two-step process to identify key functional genes that are significantly changed in the patient and play important roles in tumorigenesis, which may provide aid to biomarker identification. Finally, CCAS has a user-friendly web interface, one-click input data submission, smooth and efficient data analysis. No installation or command lines skills are necessary for using CCAS, making it very efficient for users. The current version of CCAS still has some shortcomings, which only integrates knowledge in the resources but with the lack of consensus score to evaluate abnormalities in patients. In a future version, we plan to design an algorithm to support consensus ranking score for each abnormality.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

YB and JX conceived the study and revised the manuscript. XZ, WZo, ZL, WZh, and CB performed the server front-end and back-end. XZ carried out the case studies. XZ wrote the manuscript. WZo, ZL, YM, YS, ZX, SW, FY, WZh, CB, ZD, YB, and JX polished the language and gave many constructive suggestions. All authors read and revised the final manuscript.

Funding

Strategic Priority Research Program of the Chinese Academy of Sciences (XDB38030200 and XDB38030400); National Key Research and Development Program of China (2021YFF0703704, 2017YFC0908403, and 2017YFC0907502); Key Technology Talent Program of the Chinese Academy of Sciences; Genomics Data Center Operation and Maintenance of Chinese Academy of Sciences (CAS-WX2022SDC-XK05); The Professional Association of the Alliance of International Science Organizations (ANSO-PA-2020-07); The Open Biodiversity and Health Big Data Programme of IUBS.

Acknowledgments

We acknowledge the GDC data portal developed by National Cancer Institute. The multi-omics data stored in the project has provided great help to case study in CCAS.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.956781/full#supplementary-material

References

Baumann, N. (2016). How to use the medical subject headings (MeSH). *Int. J. Clin. Pract.* 70 (2), 171–174. doi:10.1111/ijcp.12767

Birkbak, N. J., and McGranahan, N. (2020). Cancer genome evolutionary trajectories in metastasis. *Cancer Cell* 37 (1), 8–19. doi:10.1016/j.ccell.2019.12.004

Cao, Q., Wang, C., Ding, Y., Xu, D., Qian, S., Shen, H., et al. (2020). ARID1A upregulation predicts better survival in patients with urothelial bladder carcinoma. *J. Int. Med. Res.* 48 (4), 300060519895687. doi:10.1177/0300060519895687

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., et al. (2019). Open targets platform: new developments and updates two years on. *Nucleic Acids Res.* 47 (D1), D1056–D65. doi:10.1093/nar/gky1133

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2 (5), 401–404. doi:10.1158/2159-8290.CD-12-0095

Chakraborty, S., Hosen, M. I., Ahmed, M., and Shekhar, H. U. (2018). Onco-multi-OMICS approach: A new frontier in cancer research. *Biomed. Res. Int.* 2018, 9836256. doi:10.1155/2018/9836256

Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., et al. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* 34 (2), 155–163. doi:10.1038/nbt.3391

Chang, M. T., Bhattarai, T. S., Schram, A. M., Bielski, C. M., Donoghue, M. T. A., Jonsson, P., et al. (2018). Accelerating discovery of functional mutant alleles in cancer. *Cancer Discov.* 8 (2), 174–183. doi:10.1158/2159-8290.CD-17-0321

Cingolani, P., Platts, A., Wang, le L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6 (2), 80–92. doi:10.4161/fly.19695

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J. W., Comber, H., et al. (2013). Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *Eur. J. Cancer* 49 (6), 1374–1403. doi:10.1016/j.ejca.2012.12.027

Freshour, S. L., Kiwala, S., Cotto, K. C., Coffman, A. C., McMichael, J. F., Song, J. J., et al. (2021). Integration of the drug-gene interaction database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res.* 49 (D1), D1144–D1151. doi:10.1093/nar/gkaa1084

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6 (269), pl1. doi:10.1126/scisignal.2004088

Garraway, L. A., and Lander, E. S. (2013). Lessons from the cancer genome. *Cell* 153 (1), 17–37. doi:10.1016/j.cell.2013.03.002

Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467 (7319), 1061–1073. doi:10.1038/nature09534

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., et al. (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* 50 (D1), D687–D692. doi:10.1093/nar/gkab1028

Goldman, M. J., Craft, B., Hastie, M., Repecka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* 38 (6), 675–678. doi:10.1038/s41587-020-0546-8

Goldman, M. J., Zhang, J., Fonseca, N. A., Cortes-Ciriano, I., Xiang, Q., Craft, B., et al. (2020). A user guide for the online exploration and visualization of PCAWG data. *Nat. Commun.* 11 (1), 3400. doi:10.1038/s41467-020-16785-6

Guerra-Assuncao, J. A., Conde, L., Moghul, I., Webster, A. P., Ecker, S., Chervova, O., et al. (2020). GenomeChronicler: The personal genome project UK genomic report generator pipeline. *Front. Genet.* 11, 518644. doi:10.3389/fgene.2020.518644

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

Iftekhar, A., Berger, H., Bouznad, N., Heuberger, J., Boccellato, F., Dobrindt, U., et al. (2021). Genomic aberrations after short-term exposure to colibactin-producing *E. coli* transform primary colon epithelial cells. *Nat. Commun.* 12 (1), 1003. doi:10.1038/s41467-021-21162-y

ITP-CAoWG, Consortium (2020). Pan-cancer analysis of whole genomes. *Nature* 578 (7793), 82–93. doi:10.1038/s41586-020-1969-6

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., and Thun, M. J. (2007). Cancer statistics, 2007. *Ca. Cancer J. Clin.* 57 (1), 43–66. doi:10.3322/canjclin.57.1.43

Kaisrlikova, M., Vesela, J., Kundrat, D., Votavova, H., Dostalova Merkerova, M., Krejcik, Z., et al. (2022). *RUNX1 mutations contribute to the progression of MDS due to disruption of antitumor cellular defense: a study on patients with lower-risk MDS*. Leukemia.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141, 456 humans. *Nature* 581 (7809), 434–443. doi:10.1038/s41586-020-2308-7

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* 45 (D1), D840–D5. doi:10.1093/nar/gkw971

Kim, M., Lu, F., and Zhang, Y. (2016). Loss of HDAC-mediated repression and gain of NF-κB activation underlie cytokine induction in ARID1A- and PIK3CA-mutation-driven ovarian cancer. *Cell Rep.* 17 (1), 275–288. doi:10.1016/j.celrep.2016.09.003

Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., et al. (2017). Open targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* 45 (D1), D985–D94. doi:10.1093/nar/gkw1055

Lever, J., Zhao, E. Y., Grewal, J., Jones, M. R., and Jones, S. J. M. (2019). CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods* 16 (6), 505–507. doi:10.1038/s41592-019-0422-y

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27 (5), 718–719. doi:10.1093/bioinformatics/btq671

Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* 12 (1), 103. doi:10.1186/s13073-020-00803-9

Martinez-Jimenez, F., Muinos, F., Sentis, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., et al. (2020). A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* 20 (10), 555–572. doi:10.1038/s41568-020-0290-x

Mathur, R., Alver, B. H., San Roman, A. K., Wilson, B. G., Wang, X., Agoston, A. T., et al. (2017). ARID1A loss impairs enhancer-mediated gene regulation and drives colon cancer in mice. *Nat. Genet.* 49 (2), 296–302. doi:10.1038/ng.3744

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The Ensembl variant effect predictor. *Genome Biol.* 17 (1), 122. doi:10.1186/s13059-016-0974-4

Nakken, S., Fournous, G., Vodak, D., Aasheim, L. B., Myklebost, O., and Hovig, E. (2018). Personal cancer genome reporter: variant interpretation report for precision oncology. *Bioinformatics* 34 (10), 1778–1780. doi:10.1093/bioinformatics/btx817

Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D., and Cox, L. A. (2019). The need for multi-omics biomarker signatures in precision medicine. *Int. J. Mol. Sci.* 20 (19), E4781. doi:10.3390/ijms20194781

Pagel, K. A., Kim, R., Moad, K., Busby, B., Zheng, L., Tokheim, C., et al. (2020). Integrated informatics analysis of cancer-related variants. *JCO Clin. Cancer Inf.* 4, 310–317. doi:10.1200/CCI.19.00132

Pedersen, B. S., Layer, R. M., and Quinlan, A. R. (2016). Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol.* 17 (1), 118. doi:10.1186/s13059-016-0973-5

Reisle, C., Williamson, L. M., Pleasance, E., Davies, A., Pellegrini, B., Bleile, D. W., et al. (2022). A platform for oncogenomic reporting and interpretation. *Nat. Commun.* 13 (1), 756. doi:10.1038/s41467-022-28348-y

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* 17, 31. doi:10.1186/s13059-016-0893-4

Saito, R., Smith, C. C., Utsumi, T., Bixby, L. M., Kardos, J., Wobker, S. E., et al. (2018). Molecular subtype-specific immunocompetent models of high-grade urothelial carcinoma reveal differential neoantigen expression and response to immunotherapy. *Cancer Res.* 78 (14), 3954–3968. doi:10.1158/0008-5472.CAN-18-0173

Schriml, L. M., Munro, J. B., Schor, M., Olley, D., McCracken, C., Felix, V., et al. (2022). The human disease ontology 2022 update. *Nucleic Acids Res.* 50 (D1), D1255–D1261. doi:10.1093/nar/gkab1063

Shen, T., Wang, W., Zhou, W., Coleman, I., Cai, Q., Dong, B., et al. (2021). MAPK4 promotes prostate cancer by concerted activation of androgen receptor and AKT. *J. Clin. Invest.* 131 (4), 135465. doi:10.1172/JCI135465

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29 (1), 308–311. doi:10.1093/nar/29.1.308

Shorning, B. Y., Dass, M. S., Smalley, M. J., and Pearson, H. B. (2020). The PI3K-AKT-mTOR pathway and prostate cancer: At the crossroads of AR, MAPK, and WNT signaling. *Int. J. Mol. Sci.* 21 (12), E4507. doi:10.3390/ijms21124507

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature* 458 (7239), 719–724. doi:10.1038/nature07943

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Sun, X., Wang, S. C., Wei, Y., Luo, X., Jia, Y., Li, L., et al. (2017). Arid1a has context-dependent oncogenic and tumor suppressor functions in liver cancer. *Cancer Cell* 32 (5), 574–589. doi:10.1016/j.ccell.2017.10.007

Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M. P., Vivancos, A., Rovira, A., et al. (2018). Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* 10 (1), 25. doi:10.1186/s13073-018-0531-8

Tan, A., Abecasis, G. R., and Kang, H. M. (2015). Unified representation of genetic variants. *Bioinformatics* 31 (13), 2202–2204. doi:10.1093/bioinformatics/btv112

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47 (D1), D941–D7. doi:10.1093/nar/gky1015

Tebani, A., Afonso, C., Marret, S., and Bekri, S. (2016). Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations. *Int. J. Mol. Sci.* 17 (9), E1555. doi:10.3390/ijms17091555

Torre, L. A., Siegel, R. L., Ward, E. M., and Jemal, A. (2016). Global cancer incidence and mortality rates and trends--an update. Cancer epidemiology, biomarkers & prevention : a publication of the American association for cancer research, cosponsored by the. *Cancer Epidemiol. Biomarkers Prev.* 25 (1), 16–27. doi:10.1158/1055-9965.EPI-15-0578

Tweedie, S., Braschi, B., Gray, K., Jones, T. E. M., Seal, R. L., Yates, B., et al. (2021). Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.* 49 (D1), D939–D946. doi:10.1093/nar/gkaa980

UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49 (D1), D480–D489. doi:10.1093/nar/gkaa1100

van der Poel, H. G. (2004). Smart drugs in prostate cancer. *Eur. Urol.* 45 (1), 1–17. doi:10.1016/j.eururo.2003.08.011

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38 (16), e164. doi:10.1093/nar/gkq603

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., et al. (2018). Ensembl 2018. *Nucleic Acids Res.* 46 (D1), D754–D61. doi:10.1093/nar/gkx1098

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J. P., and Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 30 (7), 1006–1007. doi:10.1093/bioinformatics/btt730