BBA - Reviews on Cancer 1872 (2019) 122–137

Contents lists available at ScienceDirect

BBA - Reviews on Cancer

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

Review

Comprehensive elaboration of database resources utilized in next-generation sequencing-based tumor somatic mutation detection

Peng Gao[a,b,c], Rui Zhang[a,c,*], Jinming Li[a,b,c,*]

[a] National Center for Clinical Laboratories, Beijing Hospital, National Center of Gerontology, Beijing, People's Republic of China
[b] Graduate School, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, People's Republic of China
[c] Beijing Engineering Research Center of Laboratory Medicine, Beijing Hospital, Beijing, People's Republic of China

ARTICLE INFO

Keywords:
Database
Somatic mutation
Tumor
Next-generation sequencing
Precision oncology

ABSTRACT

The rapid evolution of next-generation sequencing (NGS)-based tumor genomic profile detection and the emergence of molecularly targeted therapies have enabled precision oncology. In NGS-based analysis, various types of databases have been developed to perform different functions. However, many problems still exist when using these public databases. Therefore, it is important to better understand the characteristics and limitations of each database and have them complement each other to provide useful clinical evidence for NGS testing. In this review, we elaborate on the important role of databases and their concrete applications in NGS-based somatic mutation detection. We introduce the typically used databases for sequence alignment, variant filtration, and variant interpretation, and compare the differences between the databases with similar functions. Subsequently, we determine the limitations of each database and provide the corresponding solutions. Furthermore, we present an overview diagram to clearly illustrate the database used in the entire NGS-based somatic mutation detection pipeline.

1. Introduction

Precision oncology, which has developed rapidly and become mainstream in clinical practice, is defined as the identification of targetable alterations in the molecular profiling of tumors to hinder the aberrant phenotype characterized in cancer [1] The aim of precision oncology is to improve diagnosis and identify optimal therapies for cancers, thus enhancing the survival and quality of life of patients [2,3]. Large-scale cancer genomics projects [4,5] have systematically described the molecular lesions in human cancer genomes and laid the foundation for precision oncology. In recent years, a range of next-generation sequencing (NGS)-based approaches have been increasingly applied in precision oncology [6–8]. This is because they can deliver a full qualitative and quantitative analysis of sequences from tumor materials; determine somatic mutations in a large number of tumor-associated genes through a timely, cost-effective process; and, notably, provide properly targeted therapeutics based on the detected actionable mutations.

Despite these advanced applications, the analysis of these complex NGS data remains highly challenging for both researchers and oncologists, especially in NGS-based tumor somatic mutation analysis. First, the NGS data are known to be error-prone and the post-bioinformatics analytic pipelines are always different across laboratories. These complicated bioinformatics analytic processes, such as sequence alignment

Abbreviations: NGS, next-generation sequencing; NCBI, National Center for Biotechnology Information; SNP, single nucleotide polymorphism; HGP, human genome project; UCSC, University of Santa Cruz; COSMIC, the Catalog of Somatic Mutations In Cancer; TCGA, The Cancer Genome Atlas; ICGC, International Cancer Genomics Consortium; WGS, whole genome sequencing; WES, whole exome sequencing; ExAC, the Exome Aggregation Consortium; MCG, My Cancer Genome; PCT, Personalized Cancer Therapy; GRC, Genome Reference Consortium; dbSNP, the database of short genetic variations; dbVar, the database of genomic structural variations; HGVS, Human Genome Variation Society; SND, single nucleotide differences; OMIM, Online Mendelian Inheritance in Man; HGMD, Human Gene Mutation Database; LOVD, Leiden Open Variation Database; CH, clonal hematopoiesis; AMP, Association for Molecular Pathology; ACMG, American College of Medical Genetics and Genomics; CAP, College of American Pathologists; FDA, the Food and Drug Administration; CIViC, Clinical Interpretation of Variants in Cancer; the AACR Project GENIE, the American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange; SNV, single nucleotide variation; SV, structural variation; CNV, copy number variantion; FusionGDB, fusion gene annotation database; DGV, the database of genomic variants; CNVVdb, copy number variantions across vertebrate genomes; CNVD, copy number variantion in disease database; CTG, ClinicalTrials.gov; TTD, The therapeutic target database; PharmGKB, The Pharmacogenomics Knowledgebase; GDSC, The Genomics of Drug Sensitivity in Cancer; NCCN, The National Comprehensive Cancer Network

* Corresponding authors at: National Center for Clinical Laboratories, Beijing Hospital, No.1 Dahua Road, Dongdan, Beijing, 100730, People's Republic of China.
E-mail addresses: ruizhang@nccl.org.cn (R. Zhang), jmli@nccl.org.cn (J. Li).

https://doi.org/10.1016/j.bbcan.2019.06.004
Received 25 April 2019; Received in revised form 16 June 2019; Accepted 26 June 2019
Available online 29 June 2019
0304-419X/

or variant filtration, may produce both false-positive and false-negative variant results [9–12]. Next, the collected variant clinical evidence from the literature is typically dispersive and disordered; furthermore, novel or rare variants with uncertain clinical significance are common, which slows the ability of variant interpretation. In addition, owing to the emergence of large amounts of novel targeted drugs, it is a challenge for clinicians to obtain the latest drug information in a timely manner, and this may lead to incomplete therapeutic treatment options for patients. Hence, different types of databases were established and utilized for NGS-based analytic processes to reduce these problems. These high-quality databases provide convenient data analysis service and data-sharing platforms for researchers. The databases also annotate tumor-specific or associated mutations, offer relevant clinical evidence to interpret their clinical significance and provide targeted drug information. Thus, they have become indispensable in precision oncology.

In this review, we first briefly describe some historical events that promoted the establishment of the databases. Subsequently, we elaborate the important role of the databases and their concrete applications in NGS-based somatic mutation detection. Finally, we present an overview diagram to illustrate the database used in the entire NGS-based somatic mutation detection pipeline. In our opinion, a detailed understanding of the applications and key characteristics of the typically used public databases provides useful information for the NGS testing, ensures reliable test results in clinical practice, and offers suitable therapeutic options for cancer patients.

2. Historical background of the database development

The evolution of sequencing technologies and many sequencing-related milestone events promoted the development of various types of databases (shown in Fig. 1). In 1977, Sanger sequencing was developed and the first DNA-based genome sequencing was completed (PhiX174) [13,14], leading to an increased speed of data sequencing. To realize data sharing for maximizing the value of every sequence, National Center for Biotechnology Information (NCBI) GenBank was established as a central data repository in 1982 [15]. In 1998, large-scale single-nucleotide polymorphisms were identified in the human genome and characterized human diversity [16], which promoted the formation of the Database of Short Genetic Variations (dbSNP) [17]. Four years later, the release of the International Hapmap Project further enriched the single nucleotide polymorphism (SNP) information in dbSNP [18]. The launch of the Human Genome Project (HGP) enabled the first complete version of the human genome sequence in 2003 [19]. The release of the HGP resulted in the emergence of many genomic databases, such as Ensembl [20] and the University of Santa Cruz (UCSC) Genome Browser [21]. Following the HGP, the Cancer Genome Project was launched, and the Catalog of Somatic Mutations In Cancer (COSMIC) database was established to gather somatic mutation information across different types of cancers [22].

Subsequently, multiple large-scale cancer genomic research studies such as The Cancer Genome Atlas (TCGA) [4,23] and the International Cancer Genomics Consortium (IGCG) [5] were conducted successively worldwide, and some database platforms based on TCGA or ICGC datasets were gradually generated, for instance, cBioPortal [24,25] and IntoGen [26]. Owing to the widespread availability of NGS technology, it became possible to realize the whole-genome sequencing (WGS) or whole-exome sequencing (WES) of hundreds of thousands of individuals; furthermore, the 1000 Genomes Projects [27,28], and Exome Aggregation Consortium (ExAC) [29,30] were launched to evaluate the population frequency across diverse ethnic types. Additionally, a catalog of somatic mutations was identified as potential predictive tumor biomarkers to targeted therapies, and many of them have been used in clinics to guide oncologists with regard to treatment options. Therefore, databases including My Cancer Genome (MCG) [31], Personalized Cancer Therapy (PCT) [32], Clinical Interpretation of Variants in Cancer (CIViC) [33], and OncoKB [34] were established to provide information on variant interpretation and targeted therapeutics. In 2017, the American Association for Cancer Research (AACR) initiated the AACR Project Genomics Evidence Neoplasia Information Exchange (GENIE) to promote precision oncology [35]. In addition, since 2000, some drug-related databases that provide information about drug structure, physical properties, potential targets, and detailed drug-associated clinical information have emerged, thereby contributing significantly to precision oncology.

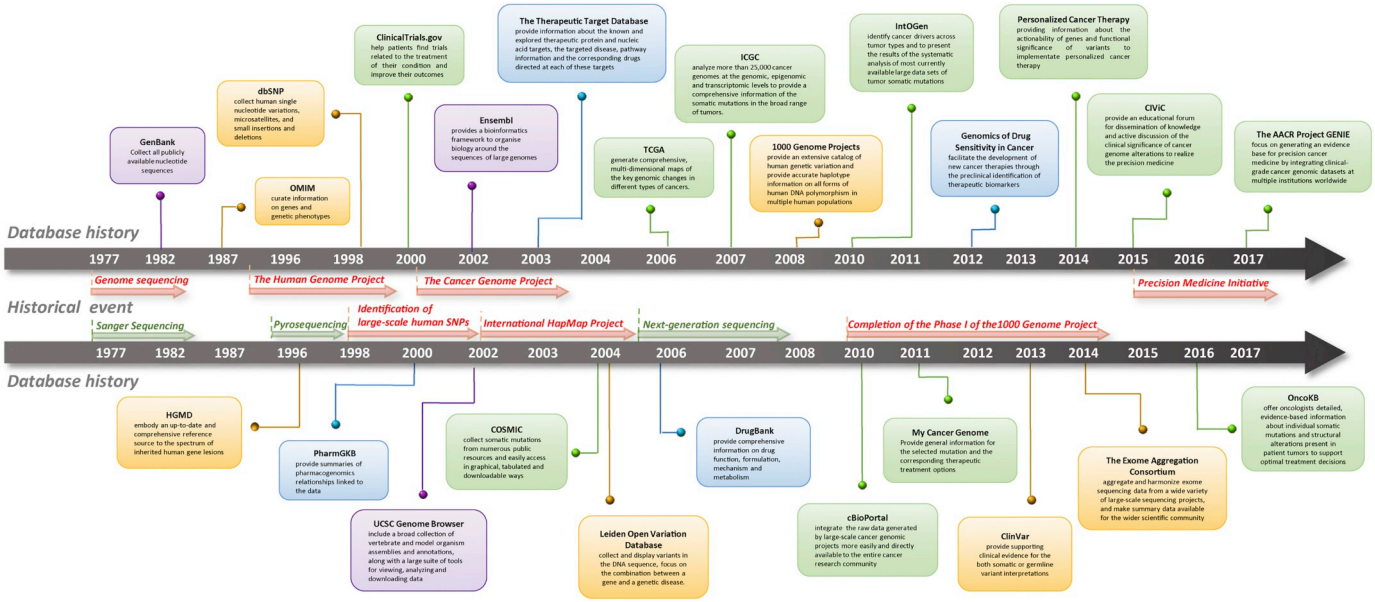

Fig. 1. Overview of the historical background for the various types of database development. Purple boxes represent the databases associated with sequence alignment. Yellow boxes represent the databases associated with variant filtration. Green boxes represent the databases associated with variant interpretation. Blue boxes represent the drug-associated databases.

Table 1
Database resources in variant filtration for tumor somatic mutation detection.

Database	Website	Supported by	Database characteristics	Variant type	Other international collaborative research projects	Utilization	Limitations	Ref
1000 Genomes project	http://www.internationalgenome.org/	The Wellcome Trust	Variant-centric	Genetic polymorphisms	NA	Find most genetic variants with frequencies of at least 1% in the population studies, provide a comprehensive resource on human genetic variation	Inefficiently remove out pseudogene interferences	[27,28]
dbSNP	http://www.ncbi.nlm.nih.gov/snp	NCBI	Variant-centric	Short genetic variations less than 50 base pairs in length	1000 Genomes project/gnomAD-Genomes/TOPMED/Genetic variation in the Estonian population/The Avon Longitudinal Study of Parents and Children/UK 10 K study-Twins/ ClinVar/HGVS/	Collect human single nucleotide variations, microsatellites, and small-scale insertions and deletions and give a detailed reference SNP information for researchers	The listed SNPs are not completely reliable.	[17]
dbVar	http://www.ncbi.nlm.nih.gov/dbvar	NCBI	Variant-centric	Genomic structural variants greater than 50 base pairs in length	1000 Genomes project/ClinGen/CNV Global Population Survey/Short Tandem Repeat Population Survey	Collect human genomic structural variation, provide a detailed reference structure variation information for researchers	Limited structure variants information	[47]
ExAC	http://exac.broadinstitute.org	The Exome Aggregation Consortium	Variant-centric	Variants on the exome	1000 Genomes/Bulgarian Trios/ Finland-United States Investigation of NIDDM Genetics (FUSION)/GoT2D/ Inflammatory Bowel Disease/ METabolic Syndrome In Men/Jackson Heart Study/Myocardial Infarction Genetics Consortium/ESP/NIMH Controls/SIGMA-T2D//SISu/Swedish Schizophrenia & Bipolar Studies/T2D-GENES/ Schizophrenia Trios from Taiwan/TCGA/TSAICG	Aggregate and harmonize exome sequencing data from a variety of large-scale sequencing projects and serve as a useful reference set of allele frequencies for severe disease studies, give a more comprehensive representation of very rare variants and allow for more accurate minor allele frequency calculations	The use of non-HGVS standard variant nomenclature	[29,30,48]
HGMD	http://www.hgmd.org	Cardiff University	Gene-centric	Germline variants	1000 Genomes Project	Provide an up-to-date and comprehensive reference source to the spectrum of inherited human gene lesions, aid the development of the post-NGS variant interpretation and exome prioritization algorithms	The public version does not contain annotation information and cannot update timely	[60–62]
OMIM	http://www.omim.org	Johns Hopkins University	Gene-genotype	Germline variants	Entrez Gene/Nomenclature/RefSeq/ GenBank/Protein/UniGene/CFMDB/ CCR/HGMD/ClinVar/PubMed/	Curate information on genes and genetic phenotypes	The use of non-HGVS standard variant nomenclature and only covers a limited number of variants on each gene	[66,67]
ClinVar	www.ncbi.nlm.nih.gov/clinvar	NCBI	Variant-centric	Germline/somatic variants	NA	Provide supporting clinical evidence for the both somatic or germline variant interpretations, give clinical evidence for therapeutic treatment	Failure in resolving the conflicting variant interpretations	[68,69]
Leiden Open Variation Database	https://www.lovd.nl/3.0	Leiden University Medical Center	Gene-centric	Germline variants	NA	Collect and display variants in the DNA sequence, focus on the combination between a gene and a genetic (heritable) disease, help to diagnose and advise patients carrying a genetic disease.	Failure in giving a certain assessment towards the variants with inconsistency clinical significance	[70,71]

Abbreviation: NCBI, National Center for Biotechnology Information; HGVS, Human Genome Variation Society; SNP, Single Nucleotide Polymorphism; CNV, Copy Number Variation; ESP, NHLBI-GO Exome Sequencing Project; NIMH, National Institute of Mental Health; SISu, Sequencing in Suomi; TCGA, the Cancer Genome Atlas; TSAICG, Tourette Syndrome Association International Consortium for Genomics; CFMDB, Cystic Fibrosis locus-specific Mutation Database; CCR, Coriell Cell Repository; HGMD, Human Gene Mutation Database; NA, Not Available.

3. Applications of the databases in NGS-based somatic mutation analysis

3.1. Sequence alignment

The accurate alignment of raw reads to the human reference genome is the first and critical step in NGS-based somatic mutation analysis [36,37]. The human reference genome sequence is necessary in the alignment process and can be freely downloaded from reference databases, such as National Center for Biotechnology Information (NCBI) GenBank [38], the UCSC Genome Browser database [21], and Ensembl [20]. The NCBI GRCh37 reference sequence from the Genome Reference Consortium (GRC) and the hg19 genome assembly from UCSC are identical but differ slightly in nomenclature and have been extensively used as reference assembly in most researches for many years. In December 2013, the GRCh38/hg38 assembly was released; this was a major update to the reference genome with improvements including fewer gaps, few sequencing errors, and centromere sequences modeled in comparison with GRCh37/hg19 [39]. Notably, the genomic coordinates, mRNA transcripts, and exon boundary definitions between these two versions are different.

However, although there are some differences between each version of the human reference sequence, consistent alignment results should be guaranteed when the same version of the human reference sequence is used. This is of significance for subsequent tumor mutation analysis. During the sequence alignment process, some related artifacts may be generated. These artifacts typically remain at a low variant frequency and can be distinguished easily from true mutations by the high levels of recurrence within different samples [40]. Many alignment tools have been developed such that millions of short reads are mapped efficiently to the human reference genome. These tools include Bowtie [41], Bowtie2 [42], BWA [43], and SOAP2 [44].

3.2. Variant filtration

3.2.1. Filtration of single-nucleotide polymorphism

When analysis-ready reads are aligned to the human reference genome, sequent filter steps are utilized to remove false-positive variants. For somatic mutation analysis, these false-positive variants are considered as single-nucleotide polymorphisms or germline mutations. It is typical to compare normal and tumor samples against the reference genome and analyzed the differences between them to determine the tumor-specific mutations. However, this method has not been routinely used in most clinical laboratories, owing to the doubled cost and turnaround time [45]. In the absence of matched normal samples, the typical practice is to compare the detected mutations with large-population databases to filter the previously described SNPs and the typical variants that exist naturally in human populations [37,46]. A variant with a population frequency within a database of more than 1% is considered benign and should be excluded from further analysis. The typically used population databases include the 1000 Genomes Project [27], dbSNP [17], the Database of Genomic Structural Variations (dbVar) [47], and the Exome Aggregation Consortium (ExAC) [29].

The 1000 Genomes Project was launched to produce an extensive catalog of human genetic variations and provide DNA polymorphism information in diverse human populations [27,28]. The project characterizes more than 95% of variants within human genome regions and discerns the variants with a 1% or higher allele frequency in five major population groups, representing a significant progress in the detailed description of human DNA polymorphisms. The dbSNP database has served as a public repository for a broad collection of genetic polymorphisms since its first release and contains the 1000 Genomes Project data. This database primarily collects fewer than 50 base pairs of genetic variations [17]. Corresponding with the dbSNP is the dbVar, which summarizes more than 50 base pairs of structural variants [47]. In the dbSNP and dbVar databases, each variant contains an accession number and the variant information include the genomic coordinates, allele frequency, Human Genome Variation Society (HGVS) nomenclature, and variant clinical significance. When the variations are identified and cataloged in the databases, laboratories can directly use the relevant polymorphism information in further applications. The ExAC database summarizes more than 60,000 individuals exome sequencing data from diverse human populations and variant allele population frequency information can also be used to discern the SNPs [29,48]. However, ExAC does not use the standard HGVS nomenclature and this limitation may increase the false-negative possibilities [48]. Detailed differences between these databases are listed in Table 1.

In SNP filtration, some problems exist in the databases that may result in missing or wrongly filtered somatic mutations. First, the SNPs listed in the population databases above are not completely reliable. It was reported that only 63% of the SNPs in dbSNP were validated for human populations. Previous research has indicated that the false-positive rate of dbSNP was 15–17% owing to sequencing errors caused by variant calling algorithms [49]. Another study indicated that up to 8.32% of the SNPs in the dbSNP are artifacts because of the highly similar genes presented in the human genome. These SNPs may not be real interindividual DNA sequence variations, but single nucleotide differences (SNDs) [50]. These SNDs are sequence artifacts between two highly similar duplicated sequence genes, which are typically generated by uncritical bioinformatics alignments or PCR amplification using improper primers that cannot better distinguish between highly similar DNA segments [50]. Next, the population databases such as dbSNP may include some cancer-related somatic mutations. These somatic mutations that emerged in dbSNP may also occur as the cancer susceptibility SNPs and may have interactions with other somatic alterations and contribute to tumor formation [51,52]. A study investigated the overlapping somatic mutations in tumor samples by comparing the dbSNP and COSMIC databases. The results indicated that among the 653 cancer-associated somatic mutations in COSMIC, 39% of them were found in dbSNP [53]. Therefore, indiscriminate filtering based on public population databases to remove SNPs may result in an increased likelihood of false negatives. The method to reduce this type of problem is to use dbSNP to filter out only the typical SNPs when the population minor allele frequency is higher than 1%. Another method is to use the 1000 Genomes Project database instead of dbSNP, because the possibility of removing cancer-associated mutations is much smaller [53]. Subsequently, some special gene regions such as pseudogene regions should be considered. Pseudogenes are sequences that have high similarity with specific protein-coding genes, which are non-functional copies of gene fragments incorporated into the genome either by retrotransposition or the duplication of genomic DNA [54,55]. It is known that more than 8000 pseudogenes have been identified in the human genome [56]. Databases such as the 1000 Genomes Project and ExAC cannot efficiently remove pseudogene interferences. Therefore, it is necessary to use the Pseudogene database [57,58] or some pseudogene identification tools [59] to filter the interfaced variants in the pseudogenic region.

3.2.2. Filtration of germline mutations

In cancer research, it is important to distinguish somatic from germline variants because these two types of variants typically yield different effects in tumor formation. Somatic mutations can be tissue-specific and are typically present in somatic cells. Germline variants are inherited and always present in germ cells. A major application of NGS somatic mutation analysis is to distinguish somatic mutations in tumor tissues from germline variants in normal tissues. However, sample contamination, sequencing errors, insufficient variant coverage, and infiltration of germline mutations often yield significant challenges. False positives emerge when germline variants fail in calling because of a low variant frequency in the normal sample. Hence, the disease databases in variant filtration are critical for effectively reducing the false-positive rate to ensure the accuracy of the somatic mutation detection.

It is noteworthy that a large portion of the SNP information overlaps with the germline mutation information. We discuss this separately because we consider the SNPs as nonpathogenic, while the germline mutation is regarded as pathogenic. The most typically used germline mutation databases are the Online Mendelian Inheritance in Man (OMIM), Human Gene Mutation Database (HGMD), ClinVar, and Leiden Open Variation Database (LOVD).

HGMD is a comprehensive repository for the germline spectrum of inherited human disease and can be accessible in either a public or professional version. The public version maintains basic functions but does not contain any of the annotation information, and is freely accessible to academic institutions or nonprofit organizations. The professional version provides variant annotation information including genomic coordinates, HGVS nomenclature, and relevant literature [60–62]. In clinical practice, HGMD can facilitate the filtering of germline-associated mutations. Additionally, the genetic information resource in the HGMD has been added to some post-NGS variant interpretations and exome prioritization algorithms [63,64]. OMIM is another germline variant repository database. Unlike HGMD, OMIM does not merely include the germline associated variants; it also contains a few polymorphisms and somatic variants [65]. OMIM summarizes crucial and novel variant information based on the literature, thus aiding in clinical and molecular genetic research via a user-friendly searchable website [66,67]. Each variant has its corresponding OMIM number, providing information on the title, variant, and a detailed description. OMIM does not provide statistics on allelic variants because it only includes a limited number of variants on each gene. Additionally, the variant information is not precisely obeyed by the HGVS nomenclature; therefore, searching a specific variant in OMIM may be challenging. ClinVar is an archival database that collects and aggregates submitted reports and interpretations of clinical and functional significance of the germline-original variants [68,69]. Each submitter can provide supporting evidence for the clinical significance of the variants, and users can easily obtain and compare different variant interpretations submitted by different submitters towards the same variant. Until 2016, more than 158,000 submitted variant interpretations representing more than 125,000 variants affecting more than 26,000 genes were contained in ClinVar [69]. In addition to its primary function, the ClinVar database collected variant information to allow for discussion of many variant interpretations. The shared data can facilitate the identification of differences in the variant interpretation. ClinVar provides a valuable platform to demonstrate these conflicting interpretations, but its limitation lies in a failure to resolve these differences for submitters. In addition, germline mutation information can be found in the LOVD, which provides information on phenotypes, family data, functional data, variant frequency, and associated literature [70,71]. Notably, the database is gene-centered based on HGVS recommendations and focuses on its usability and flexibility. Submitters and curators can determine the pathogenicity of a variant, but users still need to assess the variant in an inconsistent situation by themselves.

It is noteworthy that the design of the genetic mutation testing is key to selecting the sequent method of filtering the germline mutations [72]. Obviously, the used germline mutation filtering methods for these smaller gene assays, including “hotspot” or “targeted” NGS panels, are different from those used for WES or WGS. When laboratories use a smaller gene panel that contains only a limited number of genes to detect actionable mutations, a tumor-only sequencing approach can be conducted owing to the small likelihood of identifying important germline variants [73]. Hence, performing a second matched-normal sample test may add to the patient's cost without improving the detection accuracy. It is wise to use public disease databases that include a large amount of germline mutation information as a filter to distinguish the putative somatic mutations. For larger gene panels, simply using databases to filter is insufficient. Previously published research compared tumor-only and tumor-normal analyses in either an exome-sequencing or a targeted 111 gene panel. The results revealed that the tumor-only analysis failed to completely identify germline variants, and up to one-third of actionable mutations were classified incorrectly as somatic when they were in fact germline [74]. These results demonstrate that tumor-normal analysis is more accurate for precisely determining somatic and germline mutations in larger gene panels or in WES and WGS. Additionally, matched-normal samples can be used to distinguish frequent clonal hematopoiesis (CH) - derived alterations in patients with solid tumors. CH is clonal expansion owing to somatic genomic alterations in leukemia-associated genes and usually occurs in aging human hematopoietic stem cells [75,76]. The presence of CH may generate an incorrect classification of blood-derived somatic mutations as tumor-derived somatic mutations, thus increasing the rate of false positives [77]. A sequencing result in recent research indicated that 5% of the patients had at least one CH-derived mutations misclassified as tumor-derived when the matched-normal samples were absent [78]. Failure to identify these blood-derived mutations may result in improper therapies; therefore, a paired sample analysis must be utilized in this situation.

In variant filtration analysis, it is crucial to select suitable filter methods. This can reduce the likelihood of misidentifying the detected variants, lower the rates of false positives or false negatives, and benefit tumor diagnosis and patient treatment. When analyzing “hot-spot” mutations in a small gene panel, databases can be used instead of matched-normal samples for filtering analysis. Instead of dbSNP the 1000 Genome Project can better filter the variants with a population frequency higher than 1% because the 1000 Genome Project may reduce the likelihood of filtering the cancer-related mutations. In addition, ExAC can be used to supplement the 1000 Genome Project data, and ClinVar can be used to filter the existing germline mutations. When considering genome-scale mutation testing, it is wise and more accurate to conduct a tumor-matched filter analysis because it not only reduces many false positives including SNPs or germline mutations but can also recognize and filter out CH-derived mutations. However, when a matched-normal sample is absent, it is necessary to use numerous existing public databases to remove the variant interferences. First, population-based databases including dbSNP, the 1000 Genome Project, and ExAC are combined to filter the typical SNPs. In addition, it is necessary to remove existing germline variants collected in locus-specific databases including HGMD, OMIM, ClinVar, and LOVD. Finally, some special types of interferences, such as pseudogene interferences, should be dealt with its specific pseudogene database or tools. However, for the interference of CH-derived mutations, there is currently no corresponding database, and the only solution to filter them is the paired sample. An accurate variant filter process provides valid somatic mutation detection results, which is a vital prerequisite for sequent variant interpretation.

3.3. Variant interpretation

3.3.1. Identification of putative somatic mutations

After completing the filtration of SNPs and germline mutations, the putative somatic mutations should be further compared with the variant catalog database to determine whether they were previously reported in the literature or are novel mutations. COSMIC is the broadest somatic mutation information database [22]. It gathers and performs a detailed analysis of somatic mutation data from a variety of public resources and can be easily accessible in a variety of graphical and tabulated methods. To date, this database has included a wide range of somatic mutations throughout the human genome and has focused on more than 400 key cancer genes [79]. The database is updated four times a year to rapidly respond to the changing trends in cancer genetics. Except for the storage of variant information to help users retrieve and identify candidate somatic mutations, COSMIC also provides several functional sections such that other mutation-related information can be obtained conveniently. In the Cancer Browser section, the mutation profiles across all the related genes for a specific tumor

histology can be selected in a high to low-frequency ranking. The Cell Lines Project section includes the molecular profiles of more than 1000 cancer cell lines that are typically used in laboratory research, and can assist users in the evaluation of pharmaceutical activity and efficacy. In the Drug Resistance section, 25 drugs are detailed, and 360 somatic mutations that have caused drug resistance in 2134 tumor samples are described [79–82].

3.3.2. Assessment of clinical actionability for candidate somatic mutations

It is crucial to determine whether the candidate somatic mutation is "actionable." Actionable mutations are broadly defined as somatic alterations that have sufficient clinical evidence to support their role in predicting the response or resistance to cancer therapy [83]. According to guidelines from the Association for Molecular Pathology (AMP), American College of Medical Genetics and Genomics (ACMG), and College of American Pathologists (CAP) [84], clinical evidence is classified into four levels, and these levels of evidence are assigned to the mutations to evaluate their clinical significance and determine their clinical actionability. If a specific mutation can predict a response or resistance to the Food and Drug Administration (FDA)-approved or professional-guidelines-included therapy for a given tumor type (e.g., *EGFR* L858R mutation against gefitinib in non-small cell lung cancer), this type of mutation is level-A clinical evidence and is assigned to tier I. In this case, personalized treatment can be relatively straightforward and is considered as "standard of care." However, the number of available targeted therapeutic agents that have been clinically approved is low at this time. Thus, only a small subset of somatic mutations may be deemed actionable by this criterion. The clinical evidence for most mutations must be integrated from knowledge bases and the literature to classify their clinical significance, and the more sufficient clinical evidence always corresponds to higher tier variants and more treatment options. Hence, various types of databases have been utilized to supply valuable clinical evidence and linking experimentally determined tumor somatic mutations and clinical actionability.

My Cancer Genome (MCG) is a knowledge resource that develops a disease-organized approach to summarize the clinical evidence for specific somatic alterations in specific tumors and highlight the effect of mutations on tumor therapy [31]. Users can rapidly obtain updated information on an expanding list of genetic mutations that are involved in diverse signaling pathways affected by different types of cancers. This enables mutations to be matched to therapies such that it can be easily accessed by properly targeted therapeutics and clinical trials [85]. However, MCG presents some limitations because it is in the evolvement stage. For example, MCG contains only a small amount of mutations with strong clinical significance; more genes and mutations should be added with its development. Additionally, a variety of novel emerging targeted therapies and clinical trials should be updated in a timely and quick manner in MCG. The Personalized Cancer Therapy database (PCT) is used for clinical responses relevant to tumor mutations and was designed to encourage patients to participate in specific mutation-associated clinical trials. PCT allows users to easily obtain the variant clinical significance and actionability and their corresponding therapeutic implications, resulting in better guidance for clinicians and the best outcomes for patients [32,86]. Unlike MCG, which takes a disease-organized approach, PCT is designed in a gene-centric manner. The general gene information includes gene function, gene-associated signaling pathways, and relationships with other mediators in the signaling cascade. This information is shown on a website. Comparing this with MCG may be optimal to provide driver mutations in the selected disease. PCT provides treatment options for genomic alterations in diverse tumors, including FDA-approved drugs and clinical-trial investigational drugs. OncoKB is another comprehensive knowledge base for precision oncology, offering clinicians evidence-based information on somatic and structural aberrances in patient tumors to support proper treatment options [34]. OncoKB is organized in a gene or alteration-focus manner and creates a classification criterion according to the clinical significance of the variant. OncoKB conveys information on FDA-approved therapies and clinical-trial investigational agents. Notably, OncoKB emphasizes negative clinical outcomes to prevent the use of expensive off-label targeted therapies that are ineffective in specific mutational contexts. However, a limitation of OncoKB is that the clinical trial data are not included in the website, and users must navigate to another webpage to search for relevant clinical trials. CIViC is an expert-crowd sourced knowledge base that provides a free platform for disseminating knowledge and actively discussing the clinical implications of cancer genome alterations [33]. Its transparency represents a sustainable model that enables the standardization and comprehensive interpretation of the clinical relevance of tumor mutations to promote the development of precision medicine. CIViC is organized in a gene-focus manner: each evidence record is associated with a specific gene, variant, disease, and clinical action. Detailed variant information is related to therapy, prognostics, diagnostics, and predisposition for cancer. Notably, CIViC has a clinical-evidence-based variant classification, ranging from an established clinical practice to inferential evidence, and uses combined levels of evidence to predict the variant actionability.

Most of the variant type mentioned in the databases above are single nucleotide variation (SNV), small insertion or deletion, or limited structural variation (SV). However, many SVs such as gene fusions and copy number variations (CNVs) are also taking important roles as therapeutic targets and prognostic markers in numerous types of tumors. Owing to the increased amounts of gene fusion and CNV data, several databases have been developed. The commonly used databases for gene fusions are ChimerDB 3.0, FusionCancer, Fusion Gene annotation Database (FusionGDB) and FusionHub. ChimerDB 3.0 was built as a knowledge base of fusion genes and consists of three modules with different functions. ChimerKB mainly stores fusion genes that are compiled from well-known public resources with experimental evidence. ChimerPub provides fusion gene information from published literature in PubMed. ChimerSeq collects nonpublished or novel fusion candidates based on computational analyses of transcriptome sequencing data, but its reliability still needs to be verified [87]. FusionGDB is the first knowledge base to annotate the function of fusion genes across multiple tumor types. This database not only provides multiple annotation results of fusion genes for researchers but also provides useful information on fusion proteins and offers fusion-gene-related drugs and diseases. This makes FusionGDB a useful resource in precision medicine [88]. FusionHub serves as a unified platform that annotates and visualizes a large number of gene fusions by collecting information from 14 fusion gene datasets from the literature and 10 public databases. FusionHub can also provide the siRNA designing for a given fusion gene sequence [89].

For CNV, the database of genomic variants (DGV), Copy Number Variations across Vertebrate Genomes (CNVVdb), Copy Number Variation in Disease Database (CNVD) and CaSNP database are the most commonly used. Each of them has advantages and limitations. DGV was built as a comprehensive catalog of CNVs, but it only covers aberrants confirmed in healthy individual cohorts and cannot be used to study the effects of CNVs on cancer [90]. CNVVdb ascertains the CNV information across 16 vertebrate genomics using pairwise sequence alignment based on the Blastz algorithm. However, its limitations are that the relationship information between CNVs and tumors is not included in this database, and it still remains controversial with regard to the reliability of prediction results produced by sequence algorithms [91]. CNVD aims to become the most comprehensive and reliable CNV database. Its uniqueness lies in its text-mining-based method to manually extract and collect CNV information [92]. CaSNP is the largest SNP array-based repository for storing and visualizing CNV data for 34 different types of cancers, and can be a valuable tool to analyze the correlation between CNV data and a specific genome location across different types of cancers. However, this database only collects the CNV data based on the SNP-array method; data from other experiment types

(such as NGS) are not included [93].

Additionally, it is crucial to match patients to optimal clinical trials based on their genomic profiles [94,95]. Several clinical trial designs such as basket and umbrella trials have been suggested to test the benefit of somatic alterations in guiding treatment selection [96]. It is noteworthy that the clinical trials can be "genotype-selected" or "genotype-relevant" [97]. A "genotype-selected" trial is considered when a specific genomic alteration is required to be eligible for a trial. For example, erlotinib serves as first-line therapy in *EGFR* mutation-positive patients (NCT01250119). The "genotype-relevant" trial is a clinical trial that targets a specific gene product or the downstream signaling pathway of the genomic alteration. Although it is difficult for oncologists to determine the association between genomic alterations and genotype-relevant clinical trials owing to rapid developments in the biomedical field, some measures have been taken to address this need [98]. Clinical trial information can be obtained by linking the above-mentioned personalized cancer medicine databases and by using the largest database of clinical trials, ClinicalTrials.gov (CTG) [99]. CTG was established to help patients obtain supported clinical trial information related to the treatment of their conditions and improve their outcomes. Clinicians can obtain the matched trials by inquiring the disease, drug name, or country. They can then carefully read the medical information including the trial's purpose, interventions, and eligibility criteria, as well as the organizational information including the timeframes, sponsors, and participating centers, and choose the optional trials that the patients can participate in. However, it has been revealed that trial descriptions in CTG are extremely difficult to read, and that more studies should focus on improving its readability [100].

Some drug-associated databases can also help to provide targeted drug information or preclinical evidence. The therapeutic target database (TTD) is designed to supply detailed information on treatment targets and relevant guideline-approved clinical trial and investigational drugs. This is highly useful in promoting focused drug discovery efforts and pharmaceutical investigations for the most relevant and validated targets [101,102]. DrugBank is a database that offers comprehensive information on drug functions, formulations, mechanisms, and metabolisms [103]. The primary function of the database is to be a comprehensive and fully searchable drug resource, connecting the sequence, structure, and mechanistic data of drug molecules with their drug targets. The abundant and high-quality drug information contained in DrugBank renders it one of the most widely used reference drug resources in the world [104]. Unlike other online drug resources, DrugBank contains many crosslinks with other bioinformatics and medical databases; therefore, it can support higher-level database searching and function selection. The Pharmacogenomics Knowledgebase (PharmGKB) is a publicly available database that summarizes information on the different responses of human genetic alterations to drugs [105]. It provides very detailed clinically associated information such as dosing guidelines, annotated drug labels, actionable gene-drug associations, and genotype-phenotype relationships, enabling users to apply their pharmacogenomics knowledge under a personalized medicine background [105,106]. The Genomics of Drug Sensitivity in Cancer (GDSC) is a valuable database that provides information on the molecular markers of drug response and drug sensitivity in cancer cells to develop novel and rationally designed cancer treatment methods [107]. The key features of GDSC include a comprehensive synthesis of large-scale genomic and cell-line anticancer drug sensitivity datasets to identify the therapeutic biomarkers for subsequent preclinical validation and to promote molecular biomarker discovery for drug reaction [107].

3.3.3. Mining information associated with rare or novel somatic mutations

It may be more complex when WES or WGS is sequenced because the possibility of encountering rare or novel is greater. In this scenario, clinical evidence is either absent or insufficient to support the routine clinical implementation of a "standard of care." Therefore, it may be useful to use the cancer genome information databases to excavate and obtain variant-associated information in patients' tumor samples. The information contained in these knowledge bases can be compiled from large-scale genomic datasets, shared data from the scientific community, and published literature.

The launch of the TCGA Pilot Project established a comprehensive human cancer genomic profile. This project catalogs and discovers abundant cancer-associated genome alterations in more than 10,000 tissue samples from more than 30 types of tumors, providing a snapshot of cancer drivers and disease-specific genetic backgrounds, and enabling researchers to enrich their current knowledge [4,108–113]. Similar to the TCGA project, ICGC is also a multidisciplinary, multi-institutional collaborative effort project that uses next-generation sequencing technology to systematically and comprehensively characterize 500 tumor genomes in 50 different cancer types and subtypes to detect a wide range of somatic mutations with different variant types [5,114]. Publicly available cancer genomic datasets and multi-dimensional analysis on diverse platforms will enable scientists to better understand tumor biology and accelerate studies on the discovery of novel tumor mutations, resulting in the enhanced accuracy of diagnoses and improved treatments. The AACR Project GENIE was built as an international data-sharing platform that focuses on linking cancer genomic sequencing data with clinical treatments and outcomes of multiphase clinical-grade cancer patients treated at worldwide institutions. This project makes sure that all of the data is available publicly to the entire scientific community, aiming to make researchers better understand clinical actionability across cancer mutations, identify novel therapeutic targets, and improve clinical decision-making [35,115]. This project offers insight into the common mutations of cancers and makes it possible for researchers to draw statistical conclusions between cancer-related gene mutation and its role in cancer development and treatment based on large populations of cancer patients. In addition, it can also provide the important information on rare mutations. Compared with TCGA, which includes only tumor information from patients' cancer diagnoses, this project enriches the sample data from the late-stage tumors. This expands the universe of patients who are available for study because it includes all the cancer patients sequenced at the member institutions and is not limited to those who sequenced at diagnosis or as part of a clinical trial. In addition, the AACR Project GENIE contains more patient's medical information and enables clinicians to gain a detailed understanding of the entire disease progression information from the initial diagnosis to the final prognosis.

Some cancer genome database platforms, such as cBioPortal and IntOGen, have been developed based on large-scale cancer genomics datasets, and different functions have been developed according to various demands. cBioPortal provides access to the TCGA,ICGC and AACR Project GENIE data portals, and allows for the interactive exploration of custom datasets by accessing the OncoPrinter or MuttationMapper web tools. Users can determine gene mutations in different samples within cancer research, compare the variant frequency data in multiple cancer studies, and collect all associated genomic alterations in a single tumor sample. The platform also contains functions for biological pathway exploration, survival analysis, mutual exclusivity analysis, and data download. Cross-cancer queries enable the variant frequency and mutation data for individual genes or combinations of genes for different tumor types to be assessed. Patient View collects comprehensive tumor-associated information such as clinical characteristics, details about gene alterations, targeted drugs, and relevant clinical trials [24,25]. This portal significantly reduces the barriers between genomic data and cancer researchers and allows for genomic profiles to be obtained quickly from large-scale cancer genomics projects and translated into useful clinical evidence. The IntOGen database is the first web platform that integrates high-throughput data from genome-wide experiments to analyze and summarize genes and mutations. Their relevant signaling-pathways are identified in more

Table 2
Database resources in variant interpretation for tumor somatic mutation detection.

Database	Website	Supported by	Database characteristics	Tumor type	Therapeutic implication	Classification criteria	Data resources	Utilization	Limitations	Ref
COSMIC	https://cancer.sanger.ac.uk/cosmic/	The Wellcome Trust Sanger Institute	Somatic mutation-focus manner	All types	Approved drugs, clinical trials and literatures	Score-based FATHMM-MKL algorithm	PubMed	Collect somatic mutations from numerous public resources and easily access in graphical, tabulated and downloadable ways.	The detailed information in the key genes is limited; the summarized drug information is unable to obtain directly.	[22,79–82]
My Cancer Genome	www.mycancergenome.org	VICC	Disease-organized approach	22 tumor types	Approved drugs and clinical trials	NA	Guidelines from FDA,NCCN and ClinicaTrials.gov	Provide general information for the selected mutation and the corresponding therapeutic treatment options	Covering a small amount of mutations with strong clinical significance; the therapeutic implication cannot be updated timely.	[31,85]
Personalized Cancer Therapy	www.personalizedcancertherapy.org	MDA	Gene-centric manner	All types	FDA-approved drugs and clinical trials	PCT level-based drug classification	Guidelines from FDA,-ClinicalTrials.gov, and PubMed	Provide variant basic information in a visual manner and summarize the available therapeutic treatment options	Focus on DNA-based alteration	[32,86]
OncoKB	https://oncokb.org	MSK	Gene or alteration-focus manner	56 tumor types	Standard of care drugs	OncoKB level-based gene classification	Guidelines from FDA, NCCN, ASCO, other disease-specific expert and advocacy group recommendations, ClinicalTrials.gov and PubMed.	Provide the general variant information, classify variant based on OncoKB-levels of evidence, provide basic standard of care drug options	Limited gene or variant information and only provide the FDA-approved drugs, not included the clinical trials information	[34]
CIViC	https://civicdb.org	VICC	Gene-focus manner	All types	Standard of care drugs and clinical trials	CIViC star-based evidence level	PubMed	Provide a free platform for disseminating knowledge and actively discuss the clinical implications of cancer genome alterations	Limited number of -variants-	[33]

(*continued on next page*)

Table 2 (*continued*)

Database	Website	Supported by	Database characteristics	Tumor type	Therapeutic implication	Classification criteria	Data resources	Utilization	Limitations	Ref
TCGA	https://cancergenome.nih.gov	NCI and NHGRI	Tumor type-focus manner	33 tumor types	NA	NA	Independent projects worldwide	Generate comprehensive, multi-dimensional maps of the key genomic changes in different types of cancers.	Limited analysis for new discoveries	[4,23,108–113]
ICGC	https://icgc.org/	Participating nations	Tumor type-focus manner	50 tumor types	NA	NA	TCGA and the Sanger Cancer Genome Project	Analyze more than 25,000 cancer genomes at the genomic, epigenomic and transcriptomic levels to provide a comprehensive information of the somatic mutations in the broad range of tumors.	Limited projects of the rare cancers	[5,114]
The AACR Project GENIE	https://www.aacr.org/RESEARCH/RESEARCH/PAGES/AACR-PROJECT-GENIE.ASPX	DFCI,GRCC,JHU,MDA,MSK,N-KI,UHN,VICC	Tumor-type focus manner	All tumor types	Approved drugs and clinical trials	NA	Datasets from eight member institutions	Integrate cancer genomic sequencing data and patient's clinical treatments at worldwide institutions and share all the data to the entire scientific community; Develop the harmonized standards for sharing genomic and clinical data	Merely collect genomic data from targeted-sequencing platforms, not include the sequencing data from WES or WGS genomic platforms	[35,115]
cBioPortal	http://cbioportal.org	MSK	Tumor-type focus manner	20 tumor types	FDA-approved drugs and clinical trials	NA	TCGA,ICGC, UCSC,OncoKB and intOGen	Visualize, analyze, and download the large-scale cancer genomics data sets	NA	[24,25]
IntOGen	http://www.intogen.org/	IRB	Mutation-focus manner	28 tumor types	NA	NA	TCGA,ICGC, and independent projects	Summarize somatic mutations, genes and pathways involved in tumorigenesis	The use of non-HGVS standard variant nomenclature	[26,116,117]

(*continued on next page*)

Table 2 (*continued*)

Database	Website	Supported by	Database characteristics	Tumor type	Therapeutic implication	Classification criteria	Data resources	Utilization	Limitations	Ref
Clinicaltrials.gov	https://clinicaltrials.gov	NLM	Clinical trial-focus manner	All types	Clinical trials	NA	PubMed	Summarize information on publicly or privately supported clinical studies on a wide range of diseases and help patients find trials related to the treatment of their condition and improve their outcomes	The trial description is extremely difficult to read, more works should focus on its readability improvement.	[99,100]
Therapeutic Target Database	http://bidd.nus.edu.sg/group/ttd/ttd.asp	BIDD and IDRB	Gene-centric or drug-centric	NA	Approved, clinical trial and investigational drugs	NA	PubMed	Provide information about known and explored therapeutic protein and nucleic acid targets,the targeted disease, pathway information and the corresponding drugs directed at each of these targets	Lack of newly derived data and novel treatment strategies -	[101,102]
DrugBank	www.drugbank.ca	CIHR	Drug data focus manner	NA	FDA/ Health Canada/ EMA-approved drugs, clinical trials	NA	Therapeutic Targets Database/ PharmGKB/ UNIPROT/PubMed	Provide comprehensive information on drug function, formulation, mechanism and metabolism	Lack of newly drug associated information -	[103,104]
PharmGKB	https://www.pharmgkb.org	NIH and NIGMS	Pharmacogenomics or pharmacogenetics information focus manner	NA	Give detailed clinically relevant information,such as dosing guidelines, drug labels and clinical annotations	PharmGKB level-based evidence classification	PubMed	Aid researchers in understanding how variation in a person's genetic makeup affects how he or she responds to a drug and to identify consistent genetic variant-drug response interactions	Lack of the analysis on rare variations	[105,106]

(*continued on next page*)

Table 2 (*continued*)

Database	Website	Supported by	Database characteristics	Tumor type	Therapeutic implication	Classification criteria	Data resources	Utilization	Limitations	Ref
Genomics of Drug Sensitivity in Cancer	https://www.cancerrxgene.org	The Wellcome Sanger Institute and the Center for Molecular Therapeutics, Massachusetts General Hospital Cancer Center	Tumor cell lines and their associated drug susceptibility information focus manner	Almost 700 cancer cell lines	Give the information on drug sensitivity in cancer cells and molecular markers of drug response	NA	COSMIC	Give the information on drug sensitivity in cancer cells and molecular markers of drug response	Limited drug sensitivity data on tumor cell lines	[107]

Abbreviation: VICC, Vanderbilt-Ingram Cancer Centre; MDA, The university of Texas MD Anderson Cancer Center; MSK, Memorial Sloan Kettering Cancer Center; NCI, National Center Institute; NHGRI, The National Human Genome Research Institute; DFCI, Dana-Farber Cancer Institute; GRCC, Institute Gustave Roussy; JHU, Johns Hopkins Sidney Kimmel Comprehensive Cancer Center; NKI, Netherlands Cancer Institute; UHN, Princess Margaret Cancer Centre; IRB, Institute for Research Biomedicine; NLM, National Library of Medicine; BIDD, Bioinformatics and Drug Design group; IDRB, The Innovative Drug Research and Bioinformatics Group; CIHR, Canadian Institutes of Health Research; NIH, National Institutes of Health; NIGMS, National Institute of General Medical Sciences; FDA, Food and Drug Administration; WES, whole-exome sequencing; WGS, whole-genome sequencing; ASCO, American Society of Clinical Oncology; NCCN, National Comprehensive Cancer Network.

than 4000 exomes from 13 types of tumors [26,116,117]. The significance of IntOGen is in identifying driver alterations across tumor types and presenting the results of tumor genomes analyzed by different mutation-calling workflows. Additionally, IntOGen can be used to support cancer researchers in somatic mutations ranking, with the final objective of providing better clinical decision-making. In addition, many in silico tools are frequently used to predict changes in the structure and function of proteins owing to gene alterations, especially when interpreting novel or rare somatic mutations. Typical algorithms used to predict the effect of missense mutations on protein function include PolyPhen2 [118], SIFT [119], MutationTaster [120], and GERP ++ [121]. Typically used methods for splice site prediction include the Human Splicing Finder [122], NetGene2 [123], and GeneSplicer [124].

Although some of the interpretative databases have similar functions, differences still exist. Detailed information about the differences between the databases is listed in Table 2. The differences and limited variant information in each database render them complementary. In addition, it is advisable to integrate different levels of clinical evidence from guidelines, different types of databases, and the literature to obtain the most comprehensive and accurate variant-associated information. This can contribute to accurate variant classification and provide correct and comprehensive targeted drug information.

According to the categories of clinical evidence mentioned in the AMP-ACMG-AMP guidelines [82], we recommend that the basic composition of database resources that realizes uniform variant interpretation should contain the following aspects. 1) Consensus and professional database. This evidence can be originate from the FDA and professional guidelines such as those of the National Comprehensive Cancer Network (NCCN). 2) Variant interpretative databases. Evidence can be collected from variant interpretative databases such as MCG, PCT, OncoKB, and CIViC, which can provide information based on well-organized studies with consensus, clinical trials, and case reports. However, the variants included in the variant interpretative databases may be limited and are always hot-spots or have strong clinical significance. Therefore, databases from other levels of evidence should complementary them. 3) Variant catalog databases. Evidence can be collected from variant catalog databases such as COSMIC. This covers large amounts of existing somatic mutation information, which can complement with the variant interpretative databases, but the targeted drug information included may be insufficient. 4) Clinical trial database. NCCN demonstrates that the best management for any patient with cancer is in a clinical trial [125]. Evidence from databases such as CTG and PubMed can provide inclusion criteria for clinical trial information or small studies, and can provide supplement drug information to give clinicians more opportunities for treatment options. 5) Drug-associated database. Evidence collected from drug-associated databases can help supplement not only relevant targeted drug information but also preclinical evidence, which may not be included in variant interpretative or catalog databases. 6) Databases or web platforms based on large-scale genomic datasets. Evidence can be collected from large-scale genomic datasets such as TCGA, ICGC, the AACR Project GENIE, and some cancer genome database platforms such as cBioPortal and IntOGen, which can excavate the relevant mutation information and determine rare or novel somatic alterations from large datasets.

A schematic overview of the database composition in the variant interpretation is shown in Fig. 3. It is noteworthy that the databases mentioned in our review are principal types that should be included in the variant interpretation process. However, our review did not determine the specific databases that should be used. More types of databases will be established and included to increase the available clinical evidence and drug information with the development of precision medicine. The application of this interpretative database combination strategy will help achieve consistent variant interpretation results.

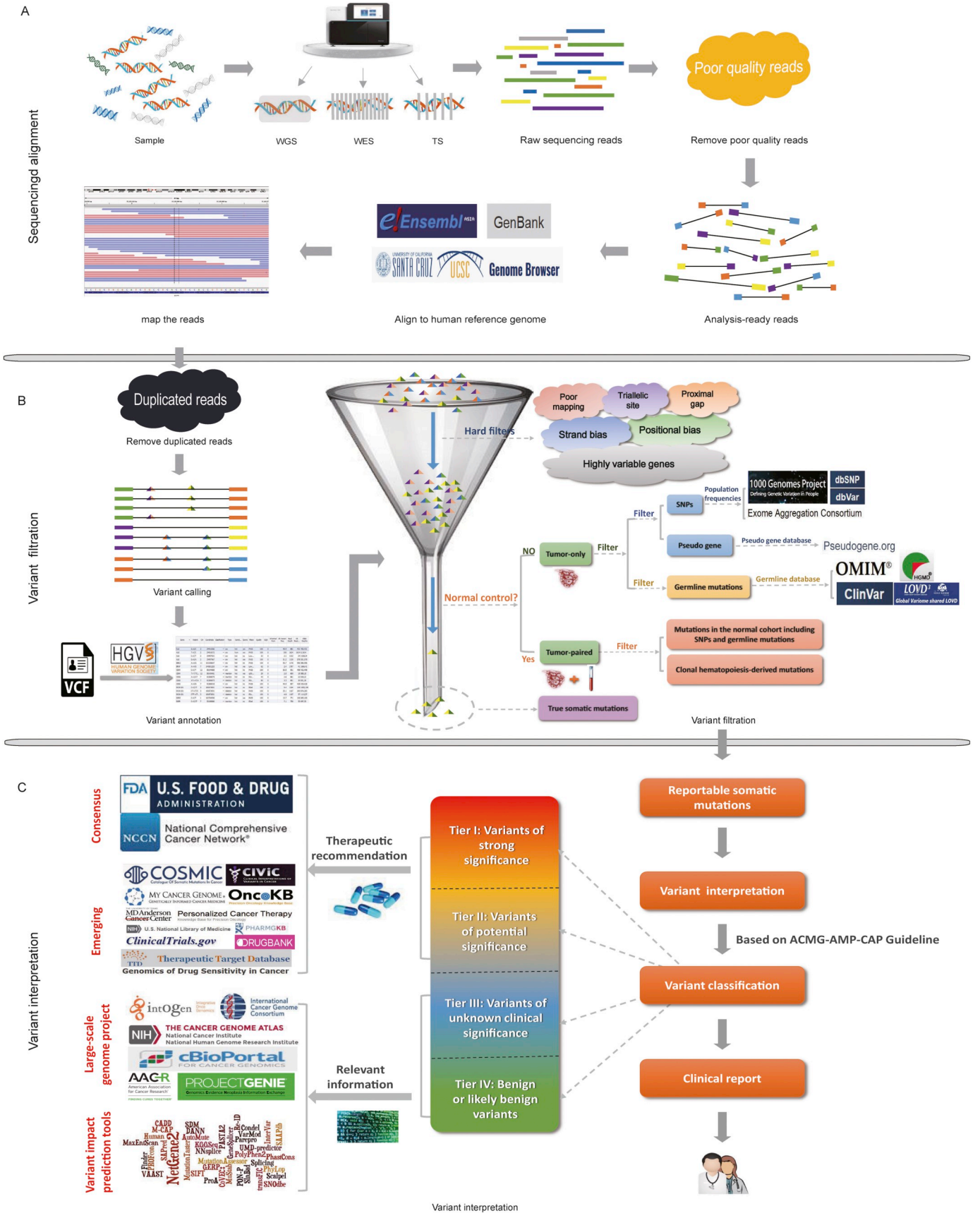

Fig. 2. A schematic illustration of the used databases for three main processes in NGS-based somatic mutation detection. (A) The databases used in the sequence alignment process. (B) The databases used in the variant filtration process. (C) The databases used in the variant interpretation process.

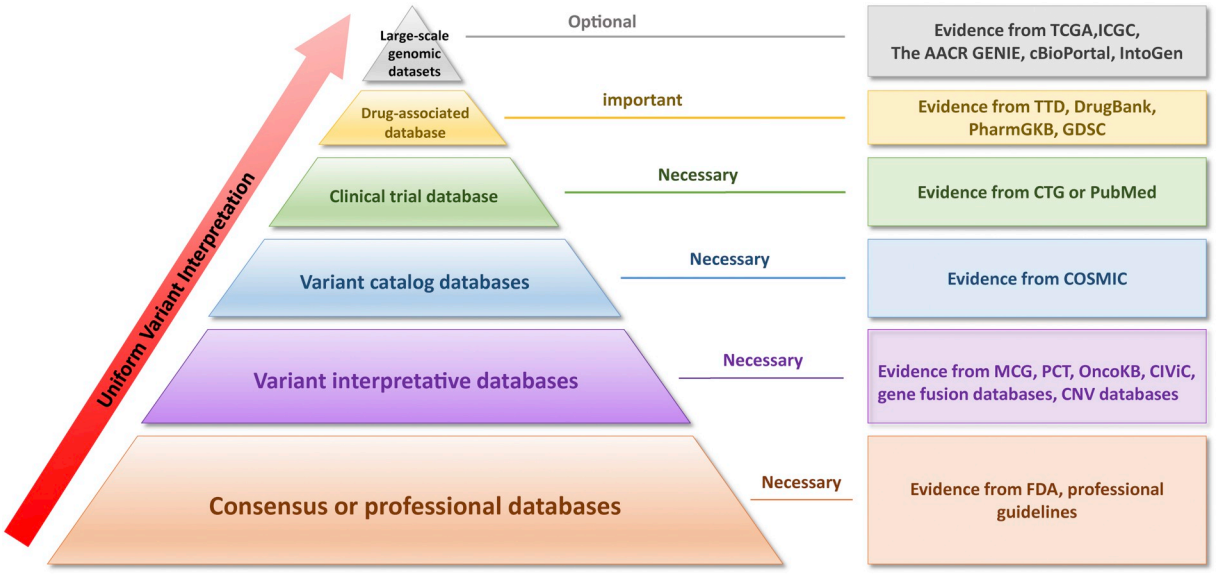

Fig. 3. A schematic diagram of database composition in the variant interpretation process. Different levels of evidence complement each other to achieve the unification of the variant interpretation process.

4. Conclusion and future directions

The rapid development of genomic profile detection and the emergence of numerous targeted therapeutics have enabled precision oncology. Various types of databases were utilized to translate experimental findings into clinical evidence, thereby allowing clinicians to carefully assess and validate the clinical actionability of detected variations in tumors. However, many problems still exist when using these public databases. Most databases have no authoritative evaluation of the accuracy or innovation of the variant data, resulting in incorrect data, duplicated data, and unclear data classification. Additionally, owing to the wide variety of databases and the large amount of variant information, it is currently impossible to incorporate all of the information into one database. Therefore, it is important to understand the characteristics and limitations of each used database. Each database should be used in a complementary manner based on its own strengths to effectively provide variant information and improve the quality and applicability of clinical reports. Herein, we introduced the typically used databases during sequence alignment, variant filtration, and variant interpretation, and compared the differences between the databases with similar functions. We subsequently determined the limitations of every type of database and provided corresponding solutions. Furthermore, we presented an overview diagram to illustrate the database used in the entire NGS-based somatic mutation detection pipeline (shown in Fig. 2).

With the increased use of NGS-based multigene molecular detection in clinical practice, many laboratories have continuously accumulated mutation information from their detected tumor samples and established internal databases for better conducting variant analysis and interpretation. Using the SNP information internal database as an example, the laboratory gradually collected SNP information through detected tumor samples and established an internal database. When the number of detected tumor samples was sufficient, the collected SNP information represented a basic SNP level in this region, and laboratories could use this internal database to quickly filter the typical SNPs. Additionally, establishing an internal database of targeted drug information that stores and expands the corresponding drug information for a specific somatic mutation found in detected tumor samples can help laboratories to quickly retrieve the relevant drug information when the same somatic mutation is detected. However, this database must be updated in a timely manner so that clinicians can obtain the most accurate and comprehensive information.

In the future, large scientific organizations should strive to create a global and comprehensive database that includes multiple levels of clinical evidence from different types of databases to minimize the time in querying multiple sources for each variant. The database must be searched and queried based on the HGVS nomenclature and exhibit the functions of automatic annotation and updating. Except for the submission or mining of all available variant data from public databases, the accredited laboratories can also submit mutation information from their internal databases. Establishing such a high-quality database can provide researchers with a convenient data analysis service and data-sharing platform, and lay the foundation for revealing tumor molecular mechanisms. Furthermore, the variant information in the database can be used to clarify its clinical significance and provide relevant targeted drug or clinical trial information that can aid patients in treatment to achieve the goal of precision oncology.

Acknowledgments

This work was supported by a grant from the Fund for Beijing Hospital Nova Project BJ-2018-136 (Rui Zhang), National Natural Science Foundation of China grant 81601848 (Rui Zhang) and National Natural Science Foundation of China grant 81772273 (Jinming Li).

Declaration of Competing Interest

None.

References

[1] L. Schwartzberg, E.S. Kim, D. Liu, D. Schrag, Precision Oncology: who, how, what, when, and when not? Am. Soc. Clin. Oncol. Educ. Book 37 (2017) 160–169.
[2] E.A. Ashley, Towards precision medicine, Nat. Rev. Genet. 17 (2016) 507–522.
[3] C. Kumar-Sinha, A.M. Chinnaiyan, Precision oncology in the age of integrative genomics, Nat. Biotechnol. 36 (2018) 46–60.
[4] N. Cancer Genome Atlas Research, J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart, The Cancer Genome Atlas Pan-Cancer analysis project, Nat. Genet. 45 (2013) 1113–1120.
[5] C. International Cancer Genome, T.J. Hudson, W. Anderson, A. Artez, A.D. Barker, C. Bell, R.R. Bernabe, M.K. Bhan, F. Calvo, I. Eerola, D.S. Gerhard, A. Guttmacher, M. Guyer, F.M. Hemsley, J.L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusada, D.P. Lane, F. Laplace, L. Youyong, G. Nettekoven, B. Ozenberger, J. Peterson, T.S. Rao, J. Remacle, A.J. Schafer, T. Shibata, M.R. Stratton, J.G. Vockley, K. Watanabe, H. Yang, M.M. Yuen, B.M. Knoppers, M. Bobrow, A. Cambon-

Thomsen, L.G. Dressler, S.O. Dyke, Y. Joly, K. Kato, K.L. Kennedy, P. Nicolas, M.J. Parker, E. Rial-Sebbag, C.M. Romeo-Casabona, K.M. Shaw, S. Wallace, G.L. Wiesner, N. Zeps, P. Lichter, A.V. Biankin, C. Chabannon, L. Chin, B. Clement, E. de Alava, F. Degos, M.L. Ferguson, P. Geary, D.N. Hayes, T.J. Hudson, A.L. Johns, A. Kasprzyk, H. Nakagawa, R. Penny, M.A. Piris, R. Sarin, A. Scarpa, T. Shibata, M. van de Vijver, P.A. Futreal, H. Aburatani, M. Bayes, D.D. Botwell, P.J. Campbell, X. Estivill, D.S. Gerhard, S.M. Grimmond, I. Gut, M. Hirst, C. Lopez-Otin, P. Majumder, M. Marra, J.D. McPherson, H. Nakagawa, Z. Ning, X.S. Puente, Y. Ruan, T. Shibata, M.R. Stratton, H.G. Stunnenberg, H. Swerdlow, V.E. Velculescu, R.K. Wilson, H.H. Xue, L. Yang, P.T. Spellman, G.D. Bader, P.C. Boutros, P.J. Campbell, P. Flicek, G. Getz, R. Guigo, G. Guo, D. Haussler, S. Heath, T.J. Hubbard, T. Jiang, S.M. Jones, Q. Li, N. Lopez-Bigas, R. Luo, L. Muthuswamy, B.F. Ouellette, J.V. Pearson, X.S. Puente, V. Quesada, B.J. Raphael, C. Sander, T. Shibata, T.P. Speed, L.D. Stein, J.M. Stuart, J.W. Teague, Y. Totoki, T. Tsunoda, A. Valencia, D.A. Wheeler, H. Wu, S. Zhao, G. Zhou, L.D. Stein, R. Guigo, T.J. Hubbard, Y. Joly, S.M. Jones, A. Kasprzyk, M. Lathrop, N. Lopez-Bigas, B.F. Ouellette, P.T. Spellman, J.W. Teague, G. Thomas, A. Valencia, T. Yoshida, K.L. Kennedy, M. Axton, S.O. Dyke, P.A. Futreal, D.S. Gerhard, C. Gunter, M. Guyer, T.J. Hudson, J.D. McPherson, L.J. Miller, B. Ozenberger, K.M. Shaw, A. Kasprzyk, L.D. Stein, J. Zhang, S.A. Haider, J. Wang, C.K. Yung, A. Cros, Y. Liang, S. Gnaneshan, J. Guberman, J. Hsu, M. Bobrow, D.R. Chalmers, K.W. Hasel, Y. Joly, T.S. Kaan, K.L. Kennedy, B.M. Knoppers, W.W. Lowrance, T. Masui, P. Nicolas, E. Rial-Sebbag, L.L. Rodriguez, C. Vergely, T. Yoshida, S.M. Grimmond, A.V. Biankin, D.D. Bowtell, N. Cloonan, A. de Fazio, J.R. Eshleman, D. Etemadmoghadam, B.B. Gardiner, J.G. Kench, A. Scarpa, R.L. Sutherland, M.A. Tempero, N.J. Waddell, P.J. Wilson, J.D. McPherson, S. Gallinger, M.S. Tsao, P.A. Shaw, G.M. Petersen, D. Mukhopadhyay, L. Chin, R.A. DePinho, S. Thayer, L. Muthuswamy, K. Shazand, T. Beck, M. Sam, L. Timms, V. Ballin, Y. Lu, J. Ji, X. Zhang, F. Chen, X. Hu, G. Zhou, Q. Yang, G. Tian, L. Zhang, X. Xing, X. Li, Z. Zhu, Y. Yu, J. Yu, H. Yang, M. Lathrop, J. Tost, P. Brennan, I. Holcatova, D. Zaridze, A. Brazma, L. Egevard, E. Prokhortchouk, R.E. Banks, M. Uhlen, A. Cambon-Thomsen, J. Viksna, F. Ponten, K. Skryabin, M.R. Stratton, P.A. Futreal, E. Birney, A. Borg, A.L. Borresen-Dale, C. Caldas, J.A. Foekens, S. Martin, J.S. Reis-Filho, A.L. Richardson, C. Sotiriou, H.G. Stunnenberg, G. Thoms, M. van de Vijver, L. Van't Veer, F. Calvo, D. Birnbaum, H. Blanche, P. Boucher, S. Boyault, C. Chabannon, I. Gut, J.D. Masson-Jacquemier, M. Lathrop, I. Pauporte, X. Pivot, A. Vincent-Salomon, E. Tabone, C. Theillet, G. Thomas, J. Tost, I. Treilleux, F. Calvo, P. Bioulac-Sage, B. Clement, T. Decaens, F. Degos, D. Franco, I. Gut, M. Gut, S. Heath, M. Lathrop, D. Samuel, G. Thomas, J. Zucman-Rossi, P. Lichter, R. Eils, B. Brors, J.O. Korbel, A. Korshunov, P. Landgraf, H. Lehrach, S. Pfister, B. Radlwimmer, G. Reifenberger, M.D. Taylor, C. von Kalle, P.P. Majumder, R. Sarin, T.S. Rao, M.K. Bhan, A. Scarpa, P. Pederzoli, R.A. Lawlor, M. Delledonne, A. Bardelli, A.V. Biankin, S.M. Grimmond, T. Gress, D. Klimstra, G. Zamboni, T. Shibata, Y. Nakamura, H. Nakagawa, J. Kusada, T. Tsunoda, S. Miyano, H. Aburatani, K. Kato, A. Fujimoto, T. Yoshida, E. Campo, C. Lopez-Otin, X. Estivill, R. Guigo, S. de Sanjose, M.A. Piris, E. Montserrat, M. Gonzalez-Diaz, X.S. Puente, P. Jares, A. Valencia, H. Himmelbauer, V. Quesada, S. Bea, M.R. Stratton, P.A. Futreal, P.J. Campbell, A. Vincent-Salomon, A.L. Richardson, J.S. Reis-Filho, M. van de Vijver, G. Thomas, J.D. Masson-Jacquemier, S. Aparicio, A. Borg, A.L. Borresen-Dale, C. Caldas, J.A. Foekens, H.G. Stunnenberg, L. van't Veer, D.F. Easton, P.T. Spellman, S. Martin, A.D. Barker, L. Chin, F.S. Collins, C.C. Compton, M.L. Ferguson, D.S. Gerhard, G. Getz, C. Gunter, A. Guttmacher, M. Guyer, D.N. Hayes, E.S. Lander, B. Ozenberger, R. Penny, J. Peterson, C. Sander, K.M. Shaw, T.P. Speed, P.T. Spellman, J.G. Vockley, D.A. Wheeler, R.K. Wilson, T.J. Hudson, L. Chin, B.M. Knoppers, E.S. Lander, P. Lichter, L.D. Stein, M.R. Stratton, W. Anderson, A.D. Barker, C. Bell, M. Bobrow, W. Burke, F.S. Collins, C.C. Compton, R.A. DePinho, D.F. Easton, P.A. Futreal, D.S. Gerhard, A.R. Green, M. Guyer, S.R. Hamilton, T.J. Hubbard, O.P. Kallioniemi, K.L. Kennedy, T.J. Ley, E.T. Liu, Y. Lu, P. Majumder, M. Marra, B. Ozenberger, J. Peterson, A.J. Schafer, P.T. Spellman, H.G. Stunnenberg, B.J. Wainwright, R.K. Wilson, H. Yang, International network of cancer genome projects, Nature 464 (2010) 993–998.

[6] G.M. Frampton, A. Fichtenholtz, G.A. Otto, K. Wang, S.R. Downing, J. He, M. Schnall-Levin, J. White, E.M. Sanford, P. An, J. Sun, F. Juhn, K. Brennan, K. Iwanik, A. Maillet, J. Buell, E. White, M. Zhao, S. Balasubramanian, S. Terzic, T. Richards, V. Banning, L. Garcia, K. Mahoney, Z. Zwirko, A. Donahue, H. Beltran, J.M. Mosquera, M.A. Rubin, S. Dogan, C.V. Hedvat, M.F. Berger, L. Pusztai, M. Lechner, C. Boshoff, M. Jarosz, C. Vietz, A. Parker, V.A. Miller, J.S. Ross, J. Curran, M.T. Cronin, P.J. Stephens, D. Lipson, R. Yelensky, Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing, Nat. Biotechnol. 31 (2013) 1023–1031.

[7] S. Roychowdhury, M.K. Iyer, D.R. Robinson, R.J. Lonigro, Y.M. Wu, X. Cao, S. Kalyana-Sundaram, L. Sam, O.A. Balbin, M.J. Quist, T. Barrette, J. Everett, J. Siddiqui, L.P. Kunju, N. Navone, J.C. Araujo, P. Troncoso, C.J. Logothetis, J.W. Innis, D.C. Smith, C.D. Lao, S.Y. Kim, J.S. Roberts, S.B. Gruber, K.J. Pienta, M. Talpaz, A.M. Chinnaiyan, Personalized oncology through integrative high-throughput sequencing: a pilot study, Sci. Transl. Med. 3 (2011) 111ra121.

[8] R.R. Singh, K.P. Patel, M.J. Routbort, N.G. Reddy, B.A. Barkoh, B. Handal, R. Kanagal-Shamanna, W.O. Greaves, L.J. Medeiros, K.D. Aldape, R. Luthra, Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes, J. Mol. Diagn. 15 (2013) 607–622.

[9] E. Privman, O. Penn, T. Pupko, Improving the performance of positive selection inference by filtering unreliable alignment regions, Mol. Biol. Evol. 29 (2012) 1–5.

[10] K.B. Hwang, I.H. Lee, H. Li, D.G. Won, C. Hernandez-Ferrer, J.A. Negron, S.W. Kong, Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings, Sci. Rep. 9 (2019) 3219.

[11] J. Durtschi, R.L. Margraf, E.M. Coonrod, K.C. Mallempati, K.V. Voelkerding, VarBin, a novel method for classifying true and false positive variants in NGS data, BMC Bioinformatics 14 (Suppl. 13) (2013) S2.

[12] M.A. Field, V. Cho, T.D. Andrews, C.C. Goodnow, Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies, PLoS One 10 (2015) e0143199.

[13] F. Sanger, G.M. Air, B.G. Barrell, N.L. Brown, A.R. Coulson, C.A. Fiddes, C.A. Hutchison, P.M. Slocombe, M. Smith, Nucleotide sequence of bacteriophage phi X174 DNA, Nature 265 (1977) 687–695.

[14] F. Sanger, S. Nicklen, A.R. Coulson, DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. U. S. A. 74 (1977) 5463–5467.

[15] D.A. Benson, M.S. Boguski, D.J. Lipman, J. Ostell, B.F. Ouellette, B.A. Rapp, D.L. Wheeler, GenBank, Nucleic Acids Res. 27 (1999) 12–17.

[16] D.G. Wang, J.B. Fan, C.J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M.S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T.J. Hudson, R. Lipshutz, M. Chee, E.S. Lander, Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome, Science 280 (1998) 1077–1082.

[17] < dbSNP the NCBI database of genetic variation.pdf > .

[18] C. International HapMap, The international HapMap Project, Nature 426 (2003) 789–796.

[19] C. International Human Genome Sequencing, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931–945.

[20] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, M. Clamp, The Ensembl genome database project, Nucleic Acids Res. 30 (2002) 38–41.

[21] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.

[22] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, R. Wooster, The COSMIC (catalogue of somatic mutations in Cancer) database and website, Br. J. Cancer 91 (2004) 355–358.

[23] K. Tomczak, P. Czerwinska, M. Wiznerowicz, The Cancer genome atlas (TCGA): an immeasurable source of knowledge, Contemp. Oncol. (Pozn) 19 (2015) A68–A77.

[24] E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, S.O. Sumer, B.A. Aksoy, A. Jacobsen, C.J. Byrne, M.L. Heuer, E. Larsson, Y. Antipin, B. Reva, A.P. Goldberg, C. Sander, N. Schultz, The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Cancer Disc. 2 (2012) 401–404.

[25] J. Gao, B.A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S.O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, N. Schultz, Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, Sci. Signal. 6 (2013) pl1.

[26] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M.P. Schroeder, A. Jene-Sanz, A. Santos, N. Lopez-Bigas, IntOGen-mutations identifies cancer drivers across tumor types, Nat. Methods 10 (2013) 1081–1082.

[27] C. Genomes Project, G.R. Abecasis, D. Altshuler, A. Auton, L.D. Brooks, R.M. Durbin, R.A. Gibbs, M.E. Hurles, G.A. McVean, A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061–1073.

[28] C. Genomes Project, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (2015) 68–74.

[29] K.J. Karczewski, B. Weisburd, B. Thomas, M. Solomonson, D.M. Ruderfer, D. Kavanagh, T. Hamamsy, M. Lek, K.E. Samocha, B.B. Cummings, D. Birnbaum, C. The Exome Aggregation, M.J. Daly, D.G. MacArthur, The ExAC browser: displaying reference data information from over 60 000 exomes, Nucleic Acids Res. 45 (2017) D840–D845.

[30] Y. Kobayashi, S. Yang, K. Nykamp, J. Garcia, S.E. Lincoln, S.E. Topper, Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation, Genome Med. 9 (2017) 13.

[31] S.V. Kusnoor, T.Y. Koonce, M.A. Levy, C.M. Lovly, H.M. Naylor, I.A. Anderson, C.M. Micheel, S.C. Chen, F. Ye, N.B. Giuse, My cancer genome: evaluating an educational model to introduce patients and caregivers to precision medicine information, AMIA Jt Summits Transl. Sci. Proc. 2016 (2016) 112–121.

[32] K.C. Kurnit, A.M. Bailey, J. Zeng, A.M. Johnson, M.A. Shufean, L. Brusco, B.C. Litzenburger, N.S. Sanchez, Y.B. Khotskaya, V. Holla, A. Simpson, G.B. Mills, J. Mendelsohn, E. Bernstam, K. Shaw, F. Meric-Bernstam, "Personalized cancer therapy": a publicly available precision oncology resource, Cancer Res. 77 (2017) e123–e126.

[33] M. Griffith, N.C. Spies, K. Krysiak, J.F. McMichael, A.C. Coffman, A.M. Danos, B.J. Ainscough, C.A. Ramirez, D.T. Rieke, L. Kujan, E.K. Barnell, A.H. Wagner, Z.L. Skidmore, A. Wollam, C.J. Liu, M.R. Jones, R.L. Bilski, R. Lesurf, Y.Y. Feng, N.M. Shah, M. Bonakdar, L. Trani, M. Matlock, A. Ramu, K.M. Campbell, G.C. Spies, A.P. Graubert, K. Gangavarapu, J.M. Eldred, D.E. Larson, J.R. Walker, B.M. Good, C. Wu, A.I. Su, R. Dienstmann, A.A. Margolin, D. Tamborero, N. Lopez-Bigas, S.J. Jones, R. Bose, D.H. Spencer, L.D. Wartman, R.K. Wilson, E.R. Mardis, O.L. Griffith, CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer, Nat Genet 49 (2017) 170–174.

[34] D. Chakravarty, J. Gao, S.M. Phillips, R. Kundra, H. Zhang, J. Wang, J.E. Rudolph, R. Yaeger, T. Soumerai, M.H. Nissan, M.T. Chang, S. Chandarlapaty, T.A. Traina, P.K. Paik, A.L. Ho, F.M. Hantash, A. Grupe, S.S. Baxi, M.K. Callahan, A. Snyder,

P. Chi, D. Danila, M. Gounder, J.J. Harding, M.D. Hellmann, G. Iyer, Y. Janjigian, T. Kaley, D.A. Levine, M. Lowery, A. Omuro, M.A. Postow, D. Rathkopf, A.N. Shoushtari, N. Shukla, M. Voss, E. Paraiso, A. Zehir, M.F. Berger, B.S. Taylor, L.B. Saltz, G.J. Riely, M. Ladanyi, D.M. Hyman, J. Baselga, P. Sabbatini, D.B. Solit, N. Schultz, OncoKB: A Precision Oncology Knowledge Base, JCO Precis Oncol 2017 (2017).

[35] C.M. Micheel, S.M. Sweeney, M.L. LeNoue-Newton, F. Andre, P.L. Bedard, J. Guinney, G.A. Meijer, B.J. Rollins, C.L. Sawyers, N. Schultz, K.R.M. Shaw, V.E. Velculescu, M.A. Levy, A.P.G. Consortium, American Association for Cancer Research Project genomics evidence neoplasia information exchange: from inception to first data release and beyond-lessons learned and member institutions' perspectives, JCO Clin, Cancer Inform. 2 (2018) 1–14.

[36] C. Trapnell, S.L. Salzberg, How to map billions of short reads onto genomes, Nat. Biotechnol. 27 (2009) 455–457.

[37] R. Bao, L. Huang, J. Andrade, W. Tan, W.A. Kibbe, H. Jiang, G. Feng, Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing, Cancer Inform. 13 (2014) 67–82.

[38] D.A. Benson, M.S. Boguski, D.J. Lipman, J. Ostell, B.F. Ouellette, GenBank, Nucleic Acids Res. 26 (1998) 1–7.

[39] D.M. Church, V.A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.C. Chen, R. Agarwala, W.M. McLaren, G.R. Ritchie, D. Albracht, M. Kremitzki, S. Rock, H. Kotkiewicz, C. Kremitzki, A. Wollam, L. Trani, L. Fulton, R. Fulton, L. Matthews, S. Whitehead, W. Chow, J. Torrance, M. Dunn, G. Harden, G. Threadgold, J. Wood, J. Collins, P. Heath, G. Griffiths, S. Pelan, D. Grafham, E.E. Eichler, G. Weinstock, E.R. Mardis, R.K. Wilson, K. Howe, P. Flicek, T. Hubbard, Modernizing reference genome assemblies, PLoS Biol 9 (2011) e1001091.

[40] D.T. Cheng, T.N. Mitchell, A. Zehir, R.H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z.Y. Liu, H.H. Won, S.N. Scott, A.R. Brannon, C. O'Reilly, J. Sadowska, J. Casanova, A. Yannes, J.F. Hechtman, J. Yao, W. Song, D.S. Ross, A. Oultache, S. Dogan, L. Borsu, M. Hameed, K. Nafa, M.E. Arcila, M. Ladanyi, M.F. Berger, Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology, J. Mol. Diagn. 17 (2015) 251–264.

[41] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.

[42] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359.

[43] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.

[44] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment, Bioinformatics 25 (2009) 1966–1967.

[45] D.B. Costa, Kinase inhibitor-responsive genotypes in EGFR mutated lung adenocarcinomas: moving past common point mutations or indels into uncommon kinase domain duplications and rearrangements, Transl. Lung Cancer Res. 5 (2016) 331–337.

[46] C. Gilissen, A. Hoischen, H.G. Brunner, J.A. Veltman, Disease gene identification strategies for exome sequencing, Eur. J. Hum. Genet. 20 (2012) 490–497.

[47] I. Lappalainen, J. Lopez, L. Skipper, T. Hefferon, J.D. Spalding, J. Garner, C. Chen, M. Maguire, M. Corbett, G. Zhou, J. Paschall, V. Ananiev, P. Flicek, D.M. Church, DbVar and DGVa: public archives for genomic structural variation, Nucleic Acids Res. 41 (2013) D936–D941.

[48] W. Song, S.A. Gardner, H. Hovhannisyan, A. Natalizio, K.S. Weymouth, W. Chen, I. Thibodeau, E. Bogdanova, S. Letovsky, A. Willis, N. Nagan, Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification, Genet. Med. 18 (2016) 850–854.

[49] A.A. Mitchell, M.E. Zwick, A. Chakravarti, D.J. Cutler, Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns, Bioinformatics 20 (2004) 1022–1032.

[50] L. Musumeci, J.W. Arthur, F.S. Cheung, A. Hoque, S. Lippman, J.K. Reichardt, Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies, Hum. Mutat. 31 (2010) 67–73.

[51] H.S. Kim, J.D. Minna, M.A. White, GWAS meets TCGA to illuminate mechanisms of cancer predisposition, Cell 152 (2013) 387–389.

[52] O. Fletcher, R.S. Houlston, Architecture of inherited susceptibility to common cancer, Nat. Rev. Cancer 10 (2010) 353–361.

[53] H. Jung, T. Bleazard, J. Lee, D. Hong, Systematic investigation of cancer-associated somatic point mutations in SNP databases, Nat. Biotechnol. 31 (2013) 787–789.

[54] A.J. Mighell, N.R. Smith, P.A. Robinson, A.F. Markham, Vertebrate pseudogenes, FEBS Lett. 468 (2000) 109–114.

[55] E.S. Balakirev, F.J. Ayala, Pseudogenes: are they "junk" or functional DNA? Annu. Rev. Genet. 37 (2003) 123–151.

[56] Z. Zhang, P.M. Harrison, Y. Liu, M. Gerstein, Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome, Genome Res. 13 (2003) 2541–2558.

[57] J.E. Karro, Y. Yan, D. Zheng, Z. Zhang, N. Carriero, P. Cayting, P. Harrrison, M. Gerstein, Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation, Nucleic Acids Res. 35 (2007) D55–D60.

[58] Q. Zhang, Using pseudogene database to identify lineage-specific genes and pseudogenes in humans and chimpanzees, J. Hered. 105 (2014) 436–443.

[59] Z. Zhang, N. Carriero, D. Zheng, J. Karro, P.M. Harrison, M. Gerstein, PseudoPipe: an automated pseudogene identification pipeline, Bioinformatics 22 (2006) 1437–1439.

[60] P.D. Stenson, M. Mort, E.V. Ball, K. Evans, M. Hayden, S. Heywood, M. Hussain, A.D. Phillips, D.N. Cooper, The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies, Hum. Genet. 136 (2017) 665–677.

[61] P.D. Stenson, M. Mort, E.V. Ball, K. Shaw, A. Phillips, D.N. Cooper, The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine, Hum. Genet. 133 (2014) 1–9.

[62] P.D. Stenson, E.V. Ball, M. Mort, A.D. Phillips, K. Shaw, D.N. Cooper, The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution, Curr. Protoc. Bioinforma. 39 (2012) 1.13.1–1.13.20.

[63] M.C. Lopes, C. Joyce, G.R. Ritchie, S.L. John, F. Cunningham, J. Asimit, E. Zeggini, A combined functional annotation score for non-synonymous variants, Hum. Hered. 73 (2012) 47–51.

[64] X. Liu, C. Wu, C. Li, E. Boerwinkle, dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs, Hum. Mutat. 37 (2016) 235–241.

[65] J.J. Johnston, L.G. Biesecker, Databases of genomic variation and phenotypes: existing resources and future needs, Hum. Mol. Genet. 22 (2013) R27–R31.

[66] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res. 33 (2005) D514–D517.

[67] J.S. Amberger, C.A. Bocchini, F. Schiettecatte, A.F. Scott, A. Hamosh, OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders, Nucleic Acids Res. 43 (2015) (D789–798).

[68] M.J. Landrum, J.M. Lee, G.R. Riley, W. Jang, W.S. Rubinstein, D.M. Church, D.R. Maglott, ClinVar: public archive of relationships among sequence variation and human phenotype, Nucleic Acids Res. 42 (2014) D980–D985.

[69] M.J. Landrum, J.M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, D.R. Maglott, ClinVar: public archive of interpretations of clinically relevant variants, Nucleic Acids Res. 44 (2016) D862–D868.

[70] I.F. Fokkema, J.T. den Dunnen, P.E. Taschner, LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach, Hum. Mutat. 26 (2005) 63–68.

[71] I.F. Fokkema, P.E. Taschner, G.C. Schaafsma, J. Celli, J.F. Laros, J.T. den Dunnen, LOVD v.2.0: the next generation in gene variant databases, Hum. Mutat. 32 (2011) 557–563.

[72] M.S. Brose, P. Volpe, M. Feldman, M. Kumar, I. Rishi, R. Gerrero, E. Einhorn, M. Herlyn, J. Minna, A. Nicholson, J.A. Roth, S.M. Albelda, H. Davies, C. Cox, G. Brignell, P. Stephens, P.A. Futreal, R. Wooster, M.R. Stratton, B.L. Weber, BRAF and RAS mutations in human lung cancer and melanoma, Cancer Res. 62 (2002) 6997–7000.

[73] V.M. Raymond, S.W. Gray, S. Roychowdhury, S. Joffe, A.M. Chinnaiyan, D.W. Parsons, S.E. Plon, G. Clinical Sequencing Exploratory Research Consortium Tumor Working, Germline findings in tumor-only sequencing: points to consider for clinicians and laboratories, J. Natl. Cancer Inst. 108 (2016).

[74] S. Jones, V. Anagnostou, K. Lytle, S. Parpart-Li, M. Nesselbush, D.R. Riley, M. Shukla, B. Chesnick, M. Kadan, E. Papp, K.G. Galens, D. Murphy, T. Zhang, L. Kann, M. Sausen, S.V. Angiuoli, L.A. Diaz Jr., V.E. Velculescu, Personalized genomic analyses for cancer mutation discovery and interpretation, Sci. Transl. Med. 7 (2015) (283ra253).

[75] L. Busque, J.P. Patel, M.E. Figueroa, A. Vasanthakumar, S. Provost, Z. Hamilou, L. Mollica, J. Li, A. Viale, A. Heguy, M. Hassimi, N. Socci, P.K. Bhatt, M. Gonen, C.E. Mason, A. Melnick, L.A. Godley, C.W. Brennan, O. Abdel-Wahab, R.L. Levine, Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis, Nat. Genet. 44 (2012) 1179–1181.

[76] R.L. Bowman, L. Busque, R.L. Levine, Clonal Hematopoiesis and evolution to hematopoietic malignancies, Cell Stem Cell 22 (2018) 157–170.

[77] C.C. Coombs, N.K. Gillis, X. Tan, J.S. Berg, M. Ball, M.E. Balasis, N.D. Montgomery, K.L. Bolton, J.S. Parker, T.E. Mesa, S.J. Yoder, M.C. Hayward, N.M. Patel, K.L. Richards, C.M. Walko, T.C. Knepper, J.T. Soper, J. Weiss, J.E. Grilley-Olson, W.Y. Kim, H.S. Earp 3rd, R.L. Levine, E. Papaemmanuil, A. Zehir, D.N. Hayes, E. Padron, Identification of clonal Hematopoiesis mutations in solid tumor patients undergoing unpaired next-generation sequencing assays, Clin. Cancer Res. 24 (2018) 5918–5924.

[78] R.N. Ptashkin, D.L. Mandelker, C.C. Coombs, K. Bolton, Z. Yelskaya, D.M. Hyman, D.B. Solit, J. Baselga, M.E. Arcila, M. Ladanyi, L. Zhang, R.L. Levine, M.F. Berger, A. Zehir, Prevalence of clonal hematopoiesis mutations in tumor-only clinical genomic profiling of solid Tumors, JAMA Oncol. 4 (2018) 1589–1593.

[79] S.A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C.G. Cole, S. Ward, E. Dawson, L. Ponting, R. Stefancsik, B. Harsha, C.Y. Kok, M. Jia, H. Jubb, Z. Sondka, S. Thompson, T. De, P.J. Campbell, COSMIC: somatic cancer genetics at high-resolution, Nucleic Acids Res. 45 (2017) D777–D783.

[80] S.A. Forbes, N. Bindal, S. Bamford, C. Cole, C.Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J.W. Teague, P.J. Campbell, M.R. Stratton, P.A. Futreal, COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in Cancer, Nucleic Acids Res. 39 (2011) D945–D950.

[81] S.A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C.Y. Kok, M. Jia, T. De, J.W. Teague, M.R. Stratton, U. McDermott, P.J. Campbell, COSMIC: exploring the world's knowledge of somatic mutations in human cancer, Nucleic Acids Res. 43 (2015)

D805–D811.
[82] S.A. Forbes, D. Beare, N. Bindal, S. Bamford, S. Ward, C.G. Cole, M. Jia, C. Kok, H. Boutselakis, T. De, Z. Sondka, L. Ponting, R. Stefancsik, B. Harsha, J. Tate, E. Dawson, S. Thompson, H. Jubb, P.J. Campbell, COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer, Curr. Protoc. Hum. Genet. 91 (2016) (10 11 11–10 11 37).
[83] T.H. Carr, R. McEwen, B. Dougherty, J.H. Johnson, J.R. Dry, Z. Lai, Z. Ghazoui, N.M. Laing, D.R. Hodgson, F. Cruzalegui, S.J. Hollingsworth, J.C. Barrett, Defining actionable mutations for oncology therapeutic development, Nat. Rev. Cancer 16 (2016) 319–329.
[84] M.M. Li, M. Datto, E.J. Duncavage, S. Kulkarni, N.I. Lindeman, S. Roy, A.M. Tsimberidou, C.L. Vnencak-Jones, D.J. Wolff, A. Younes, M.N. Nikiforova, Standards and Guidelines for the interpretation and reporting of sequence variants in Cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists, J. Mol. Diagn. 19 (2017) 4–23.
[85] A.D. Taylor, C.M. Micheel, I.A. Anderson, M.A. Levy, C.M. Lovly, The path(way) less traveled: a pathway-oriented approach to providing information about precision Cancer medicine on my cancer genome, Transl. Oncol. 9 (2016) 163–165.
[86] E.I. Dumbrava, F. Meric-Bernstam, Personalized cancer therapy-leveraging a knowledge base for clinical decision-making, Cold Spring Harb. Mol. Case Stud. 4 (2018).
[87] M. Lee, K. Lee, N. Yu, I. Jang, I. Choi, P. Kim, Y.E. Jang, B. Kim, S. Kim, B. Lee, J. Kang, S. Lee, ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining, Nucleic Acids Res. 45 (2017) D784–D789.
[88] P. Kim, X. Zhou, FusionGDB: fusion gene annotation DataBase, Nucleic Acids Res. 47 (2019) D994–D1004.
[89] P. Panigrahi, A. Jere, K. Anamika, FusionHub: a unified web platform for annotation and visualization of gene fusion events in human cancer, PLoS One 13 (2018) e0196588.
[90] J.R. MacDonald, R. Ziman, R.K. Yuen, L. Feuk, S.W. Scherer, The database of genomic variants: a curated collection of structural variation in the human genome, Nucleic Acids Res. 42 (2014) D986–D992.
[91] F.C. Chen, Y.Z. Chen, T.J. Chuang, CNVVdb: a database of copy number variations across vertebrate genomes, Bioinformatics 25 (2009) 1419–1421.
[92] F. Qiu, Y. Xu, K. Li, Z. Li, Y. Liu, H. DuanMu, S. Zhang, Z. Li, Z. Chang, Y. Zhou, R. Zhang, S. Zhang, C. Li, Y. Zhang, M. Liu, X. Li, CNVD: text mining-based copy number variation in disease database, Hum. Mutat. 33 (2012) E2375–E2381.
[93] Q. Cao, M. Zhou, X. Wang, C.A. Meyer, Y. Zhang, Z. Chen, C. Li, X.S. Liu, CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data, Nucleic Acids Res. 39 (2011) D968–D974.
[94] S. Roychowdhury, A.M. Chinnaiyan, Advancing precision medicine for prostate cancer through genomics, J. Clin. Oncol. 31 (2013) 1866–1873.
[95] A.P. Venook, M.E. Arcila, A.B. Benson 3rd, D.A. Berry, D.R. Camidge, R.W. Carlson, T.K. Choueiri, V. Guild, G.P. Kalemkerian, R. Kurzrock, C.M. Lovly, A.E. McKee, R.J. Morgan, A.J. Olszanski, M.W. Redman, V. Stearns, J. McClure, M.L. Birkeland, NCCN working group report: designing clinical trials in the era of multiple biomarkers and targeted therapies, J. Natl. Compr. Cancer Netw. 12 (2014) 1629–1649.
[96] L.L. Siu, B.A. Conley, S. Boerner, P.M. LoRusso, Next-generation sequencing to guide Clinical trials, Clin. Cancer Res. 21 (2015) 4536–4544.
[97] F. Meric-Bernstam, A. Johnson, V. Holla, A.M. Bailey, L. Brusco, K. Chen, M. Routbort, K.P. Patel, J. Zeng, S. Kopetz, M.A. Davies, S.A. Piha-Paul, D.S. Hong, A.K. Eterovic, A.M. Tsimberidou, R. Broaddus, E.V. Bernstam, K.R. Shaw, J. Mendelsohn, G.B. Mills, A decision support framework for genomically informed investigational cancer therapy, J. Natl. Cancer Inst. 107 (2015).
[98] A.M. Bailey, Y. Mao, J. Zeng, V. Holla, A. Johnson, L. Brusco, K. Chen, J. Mendelsohn, M.J. Routbort, G.B. Mills, F. Meric-Bernstam, Implementation of biomarker-driven cancer therapy: existing tools and remaining gaps, Discov. Med. 17 (2014) 101–114.
[99] V. Huser, J.J. Cimino, Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials, PLoS One 8 (2013) e68409.
[100] D.T. Wu, D.A. Hanauer, Q. Mei, P.M. Clark, L.C. An, J. Proulx, Q.T. Zeng, V.G. Vydiswaran, K. Collins-Thompson, K. Zheng, Assessing the readability of ClinicalTrials.gov, J. Am. Med. Inform. Assoc. 23 (2016) 269–275.
[101] X. Chen, Z.L. Ji, Y.Z. Chen, TTD: therapeutic target database, Nucleic Acids Res. 30 (2002) 412–415.
[102] X. Liu, F. Zhu, X. Ma, L. Tao, J. Zhang, S. Yang, Y. Wei, Y.Z. Chen, The therapeutic target database: an internet resource for the primary targets of approved, clinical trial and experimental drugs, Expert Opin. Ther. Targets 15 (2011) 903–912.
[103] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901–D906.
[104] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (2018) D1074–D1082.
[105] T.E. Klein, R.B. Altman, PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base, Pharmacogenomics J. 4 (2004) 1.
[106] M. Whirl-Carrillo, E.M. McDonagh, J.M. Hebert, L. Gong, K. Sangkuhl, C.F. Thorn, R.B. Altman, T.E. Klein, Pharmacogenomics knowledge for personalized medicine, Clin. Pharmacol. Ther. 92 (2012) 414–417.
[107] W. Yang, J. Soares, P. Greninger, E.J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J.A. Smith, I.R. Thompson, S. Ramaswamy, P.A. Futreal, D.A. Haber, M.R. Stratton, C. Benes, U. McDermott, M.J. Garnett, Genomics of drug sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res. 41 (2013) D955–D961.
[108] N. Cancer Genome Atlas Research, Comprehensive genomic characterization of squamous cell lung cancers, Nature 489 (2012) 519–525.
[109] N. Cancer Genome Atlas Research, Comprehensive genomic characterization defines human glioblastoma genes and core pathways, Nature 455 (2008) 1061–1068.
[110] N. Cancer Genome Atlas Research, Integrated genomic analyses of ovarian carcinoma, Nature 474 (2011) 609–615.
[111] N. Cancer Genome Atlas, Comprehensive molecular characterization of human colon and rectal cancer, Nature 487 (2012) 330–337.
[112] N. Cancer Genome Atlas, Comprehensive molecular portraits of human breast tumours, Nature 490 (2012) 61–70.
[113] C. Kandoth, M.D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J.F. McMichael, M.A. Wyczalkowski, M.D.M. Leiserson, C.A. Miller, J.S. Welch, M.J. Walter, M.C. Wendl, T.J. Ley, R.K. Wilson, B.J. Raphael, L. Ding, Mutational landscape and significance across 12 major cancer types, Nature 502 (2013) 333–339.
[114] J. Zhang, J. Baran, A. Cros, J.M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, A. Kasprzyk, International cancer genome consortium data portal–a one-stop shop for cancer genomics data, Database (Oxford) 2011 (2011) (bar026).
[115] A.P.G. Consortium, AACR project GENIE: powering precision medicine through an international Consortium, Cancer Disc. 7 (2017) 818–831.
[116] G. Gundem, C. Perez-Llamas, A. Jene-Sanz, A. Kedzierska, A. Islam, J. Deu-Pons, S.J. Furney, N. Lopez-Bigas, IntOGen: integration and data mining of multidimensional oncogenomic data, Nat. Methods 7 (2010) 92–93.
[117] C. Perez-Llamas, G. Gundem, N. Lopez-Bigas, Integrative cancer genomics (IntOGen) in Biomart, Database (Oxford) 2011 (2011) (bar039).
[118] I. Adzhubei, D.M. Jordan, S.R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2, Curr. Protoc. Hum. Genet. 76 (2013) 7.20.1–7.20.41.
[119] N.L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, P.C. Ng, SIFT web server: predicting effects of amino acid substitutions on proteins, Nucleic Acids Res. 40 (2012) W452–W457.
[120] J.M. Schwarz, D.N. Cooper, M. Schuelke, D. Seelow, MutationTaster2: mutation prediction for the deep-sequencing age, Nat. Methods 11 (2014) 361–362.
[121] E.V. Davydov, D.L. Goode, M. Sirota, G.M. Cooper, A. Sidow, S. Batzoglou, Identifying a high fraction of the human genome to be under selective constraint using GERP + +, PLoS Comput. Biol. 6 (2010) e1001025.
[122] F.O. Desmet, D. Hamroun, M. Lalande, G. Collod-Beroud, M. Claustres, C. Beroud, Human splicing finder: an online bioinformatics tool to predict splicing signals, Nucleic Acids Res. 37 (2009) e67.
[123] S. Brunak, J. Engelbrecht, S. Knudsen, Prediction of human mRNA donor and acceptor sites from the DNA sequence, J. Mol. Biol. 220 (1991) 49–65.
[124] M. Pertea, X. Lin, S.L. Salzberg, GeneSplicer: a new computational method for splice site prediction, Nucleic Acids Res. 29 (2001) 1185–1190.
[125] National Comprehensive Cancer Network. (NCCN) Clinical Practice guidelines in oncology. Non-Small Cell Lung Cancer, Version 4. 2019. Accessed on 02 June 2019. Available online: https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf