Published online 22 December 2022

NAR Cancer, 2022, Vol. 4, No. 4 **1**
https://doi.org/10.1093/narcan/zcac043

Pancan-MNVQTLdb: systematic identification of multi-nucleotide variant quantitative trait loci in 33 cancer types

Dongyang Wang[1,†], Wen Cao[2,†], Wenqian Yang[1,†], Weiwei Jin[1], Haohui Luo[1], Xiaohui Niu[1,*] and Jing Gong[1,3,*]

[1]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, Hubei 430074, China, [2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China and [3]College of Biomedicine and Health, Huazhong Agricultural University, Wuhan, Hubei 430070, China

Received September 20, 2022; Revised November 22, 2022; Editorial Decision December 10, 2022; Accepted December 13, 2022

ABSTRACT

Multi-nucleotide variants (MNVs) are defined as clusters of two or more nearby variants existing on the same haplotype in an individual. Recent studies have identified millions of MNVs in human populations, but their functions remain largely unknown. Numerous studies have demonstrated that single-nucleotide variants could serve as quantitative trait loci (QTLs) by affecting molecular phenotypes. Therefore, we propose that MNVs can also affect molecular phenotypes by influencing regulatory elements. Using the genotype data from The Cancer Genome Atlas (TCGA), we first identified 223 759 unique MNVs in 33 cancer types. Then, to decipher the functions of these MNVs, we investigated the associations between MNVs and six molecular phenotypes, including coding gene expression, miRNA expression, lncRNA expression, alternative splicing, DNA methylation and alternative polyadenylation. As a result, we identified 1 397 821 *cis*-MNVQTLs and 402 381 *trans*-MNVQTLs. We further performed survival analysis and identified 46 173 MNVQTLs associated with patient overall survival. We also linked the MNVQTLs to genome-wide association studies (GWAS) data and identified 119 762 MNVQTLs that overlap with existing GWAS loci. Finally, we developed Pancan-MNVQTLdb (http://gong_lab.hzau.edu.cn/mnvQTLdb/) for data retrieval and download. Pancan-MNVQTLdb will help decipher the functions of MNVs in different cancer types and be an important resource for genetic and cancer research.

INTRODUCTION

Sequencing technologies have rapidly advanced our understanding of human genetic variants (1). Among human genetic variants, single-nucleotide variants (SNVs) are the most common type (2), and genome-wide association studies (GWAS) have identified thousands of SNVs associated with numerous traits and diseases (3–5). However, genetic studies showed that only a tiny proportion of the heritability of quantitative traits was explained by SNVs, especially for complex diseases, and missing heritability remains to be explored (6), indicating that other types of genetic variants may also contribute to the heritability.

Multi-nucleotide variants (MNVs) are defined as clusters of two or more nearby variants existing on the same haplotype in an individual (7). When nearby variants are within the same codon, the amino acid changes caused by MNVs may differ from either of the separate SNVs (7). Currently, most existing variant identification tools often misannotate MNVs as individual SNVs with incorrect function predictions, which probably hampers scientific research and clinical practice (8). For example, Srinivasan *et al*. (8) analyzed genotype data from The Cancer Genome Atlas (TCGA) (9) and found that a large amount of MNVs was misannotated as SNVs, and half of these MNVs were annotated with novel protein functions relative to SNVs. In addition, another study (10) found significant enrichment of MNVs in genes associated with diagnosing developmental disorders, indicating that MNVs could be causal variants in diseases. However, the functions of MNVs, especially those in noncoding regions, remain largely unknown.

Previous studies have shown that single-nucleotide polymorphisms (SNPs, common variants of SNVs) could affect regulatory elements, thereby influencing gene expression (11) and other molecular phenotypes (12–15). For example, Beesley *et al*. (16) reported that the risk allele of rs61938093

*To whom correspondence should be addressed. Tel: +86 027 87285085; Email: gong.jing@mail.hzau.edu.cn
Correspondence may also be addressed to Xiaohui Niu. Tel: +86 027 87285085; Email: niuxiaoh@mail.hzau.edu.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

© The Author(s) 2022. Published by Oxford University Press on behalf of NAR Cancer.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

potentially decreased the expression level of *NTN4* by reducing the activity of the *NTN4* promoter in breast cancer, indicating that *NTN4* may be a potential mediator for breast cancer risk. Another study (17) reported rs12203592 as a hotspot of melanocyte *trans*-meQTLs. This SNP as a *cis*-eQTL was related to *IRF4* and 131 candidate target CpGs, which were enriched in *IRF4* binding sites, indicating that *IRF4* may play an important role in the regulatory network. There were also databases providing comprehensive annotations about the SNP functions (18,19). Considering that both SNVs and MNVs can alter DNA sequences, we propose that MNVs can also affect molecular phenotypes by influencing regulatory elements, such as methylation, miRNA binding and alternative splicing. Over the past few years, other researchers and we have reported many databases on the association of SNPs and molecular phenotypes (11,20–27). However, there is currently no systematic analysis of the associations between MNVs and molecular phenotypes.

The TCGA data portal (9) provides multi-omic data (e.g. genomics, transcriptomics) and clinical data of 33 cancer types generated from >9100 samples. Using this valuable resource, we first identified 223 759 unique MNVs in 33 cancer types using the genotype data. To better understand the impact of these MNVs on molecular phenotypes, we investigated the associations between MNVs and six currently available molecular phenotypes. Finally, we developed Pancan-MNVQTLdb, a database that allows users to query, search and visualize quantitative trait loci (QTLs) by cancer type, molecular phenotype and variants. Pancan-MNVQTLdb also incorporates GWAS and clinical data to help users to understand the functions of MNVs in cancers.

MATERIALS AND METHODS

MNV identification, genotyping and quality control

We first obtained the genotype data detected using the Affymetrix SNP 6.0 array from the TCGA data portal (9) and imputed autosomal variants for all samples of 33 cancer types using IMPUTE2 (1000 Genomes Phase 3 as the reference panel) (28). The imputation was performed in the two-step procedure provided by IMPUTE2. After imputation, we performed quality control to select high-quality SNPs with (i) imputation confidence score ≥0.4, (ii) minor allele frequency (MAF) ≥5%, (iii) missing rate <5% and (iv) Hardy–Weinberg equilibrium $P > 1e{-}6$.

After SNP imputation, MNV identification was performed using the strategy described by Wang *et al*. (7). Specifically, we used SHAPEIT4 (29) to perform computational phasing in each cancer type to determine the chromosome from which the allele comes. Next, Hail (version 0.2.11, https://hail.is/docsh/0.2/api.html) was used to identify MNVs from the phased genotype. First, the genotype data were converted to a matrix. Second, all the variants with the homozygote reference allele were filtered. Next, MNV calling was conducted from SNPs with distances ≤10 (window_by_locus = 10). The reason for selecting MNVs within 10 bp is that when the length of window exceeds 10, the accuracy of phase determination will be reduced precipitously (7). Finally, the aggregate function was used to obtain the site-level information.

According to the MNV definition, an MNV was defined as multiple variants in the same haplotype. We used '0', '1' and '2', representing the number of haplotypes containing an MNV, to encode the MNV genotypes. For example, an MNV with two SNPs (A>a and B>b) has 16 different combinations of SNP genotypes (Figure 1A). Individuals with (a|b, a|b) are counted as '2', and individuals with (A|B, a|b), (A|b, a|b), (a|B, a|b), (a|b, A|B), (a|b, A|b) or (a|b, a|B) are counted as '1', while others are counted as '0'. The coding of polynucleotides is similar. Using this method, we generated an MNV genotype matrix for each cancer type.

Finally, the quality controls of MNVs were performed. We excluded MNVs with low allele frequency (MAF < 1%) in QTL identification. Additionally, for each MNV, we requested that the heterozygous genotype group contain at least three samples; all the MNVs that did not meet this condition were also excluded.

Molecular phenotype data collection and processing

Six molecular phenotypes were used for MNVQTL mapping. They are miRNA expression, lncRNA expression, coding gene expression, alternative splicing, alternative polyadenylation (APA) and DNA methylation (Figure 1B). After obtaining the data from the TCGA data portal (9), we performed quality control for these data.

For the quality control of coding genes and lncRNAs, we downloaded the gene expression profile from TCGA (9) and excluded all the genes with an extremely low expression [median fragment per kilobase million (FPKM) <0.01] for downstream analysis. Then, genes were classified into coding genes and lncRNAs according to the annotation from ENCODE (version v36) (20,21). These coding genes and lncRNAs were separately used for the eQTL and lncRNA-eQTL mapping.

For the quality control of miRNAs, we downloaded the miRNA sequencing data from TCGA (9) and excluded the miRNAs with a median transcription per million (TPM) <0.01.

Percent spliced index (PSI) is a commonly used metric for quantifying splicing events. PSI is defined as the ratio of reads indicating the presence of a transcript element versus the total number of reads covering the splicing event. After downloading the PSI data from the TCGASpliceSeq (30), we excluded the PSIs with a missing rate >0.1 or on sex chromosomes.

APA events were often quantified by the percentage of distal polyA site usage index (PDUI), and PDUI is used to represent the relative usage rate of transcripts on the distal polyA site. The PDUI data for apaQTL analysis were downloaded from the TC3A data portal (31). We excluded APA events with a missing rate >0.1 or a standard deviation <5% (23).

The DNA methylation of TCGA (9) was generated by the Illumina Infinium HumanMethylation450 BeadChip array. We downloaded these data from the TCGA data portal (9) and filtered the sites according to the following criteria: (i) on sex chromosomes; (ii) mapping to multiple locations on the genome; (iii) containing known SNP on CpG sites; and (iv) beta value with a missing rate >0.05.

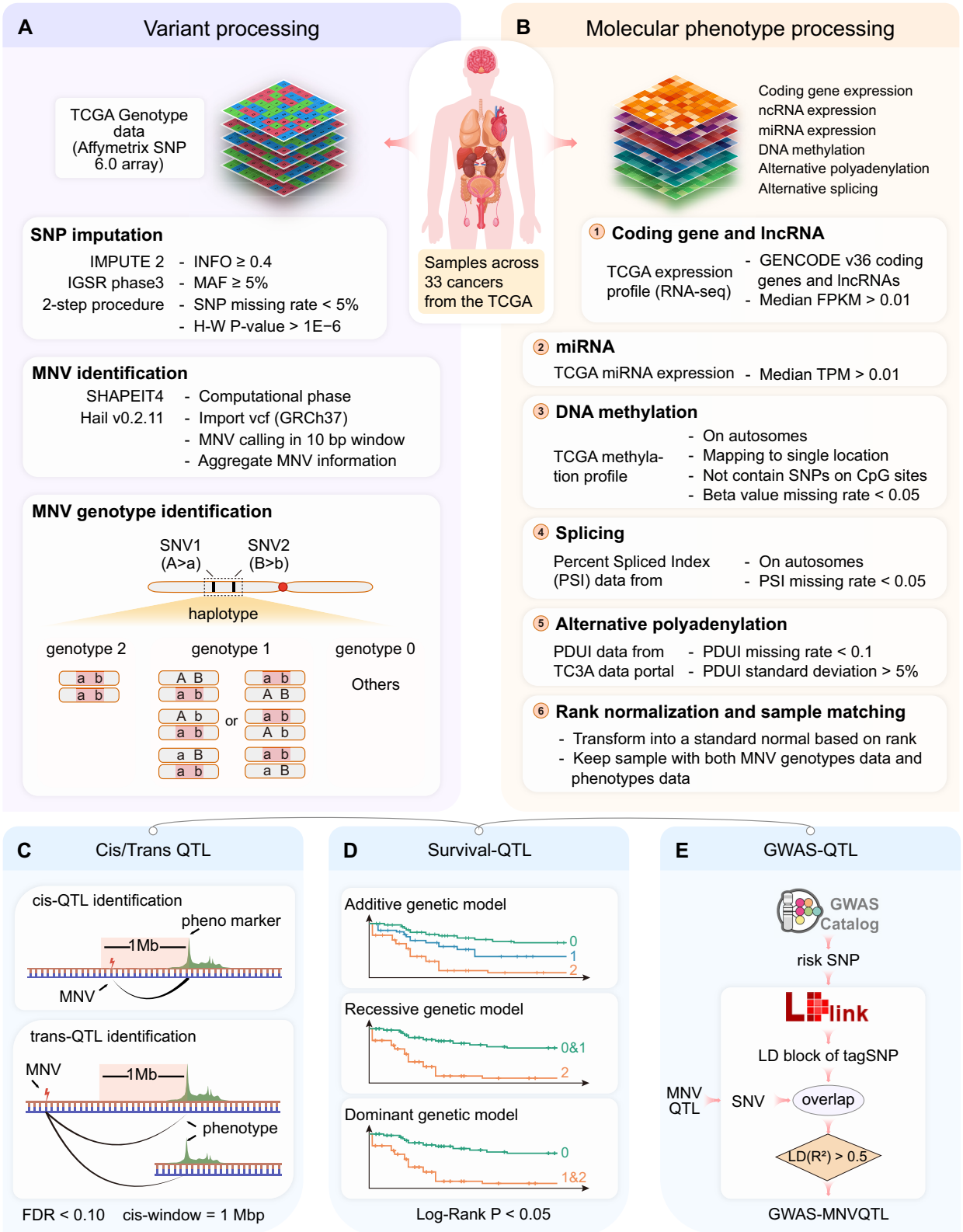

Figure 1. The pipeline of Pancan-MNVQTLdb. (**A**) Processing of variant data. (**B**) Processing of phenotype data. (**C**) Identification of *cis*- and *trans*-MNVQTLs. (**D**) Identification of survival-QTLs. Three genetic models were used to test the associations between MNVs and sample survival time. (**E**) Identification of GWAS-QTLs.

Covariates

Studies have proved that factors affecting molecular phenotypes may reduce the power of QTL identification (32,33). Adjusting covariates in QTL identification is necessary to minimize confounding effects (34). Three types of covariates were included in this study: (i) the top five principal components in genotype data to minimize the effect of population structure on molecular phenotypes (34); (ii) PEER factors equivalent to 25% of the sample size but not >100 generated by PEER software (35) to minimize the potential batch effects and other confounders in the quantitative data; and (iii) basic clinical information, including age, gender and tumor stage, to minimize the effects of clinical status on molecular phenotypes. We calculated these covariates for each cancer type separately.

Identification of MNVQTLs

For each molecular phenotype of each cancer, genotype data, molecular phenotype indexes and covariates were processed into three matrix files with the matched samples. QTL analysis was performed using the linear regression model in the Matrix eQTL R package (36). MNVs with false discovery rates (FDRs) <0.1 were defined as MNVQTLs. In addition, a window of 1 Mb was used to distinguish *cis*-MNVQTLs from *trans*-MNVQTLs (Figure 1C).

Identification of survival-related MNVQTLs

To identify MNVQTLs associated with disease prognosis, we obtained clinical information, including the survival time of all donors from TCGA (9). We applied the log-rank test to analyze the associations between survival time and MNV genotype. We used three models in this study, i.e. additive, dominant and recessive. In the additive model, we divided the donors into three groups based on the three genotypes of each MNV and tested the difference in survival time among the three groups. In the dominant model, donors with genotype '1' and donors with genotype '2' were merged into one group and tested against the other. In the recessive model, donors with genotype '1' and donors with genotype '0' were merged into one group and tested against the other. The Kaplan–Meier (KM) plot was used for visualization (Figure 1D).

Identification of GWAS-related MNVQTLs

Due to linkage disequilibrium, GWAS identified trait-related variants that are often tag variants rather than causal ones. So far, causal variants in most GWAS loci are still unclear. Thus, to facilitate the fine mapping of functional variants, we integrated the MNVQTLs with GWAS risk loci. First, we downloaded all the GWAS risk variants from the latest release of the NHGRI-EBI GWAS Catalog (37). Then, LDlink (https://ldlink.nci.nih.gov/) (38) was used to calculate the R^2 between each MNV and each GWAS locus. For each MNVQTL, if any SNV in the MNVQTL is located in the LD region ($R^2 > 0.5$) of the GWAS tagSNP, we define it as a GWAS-related MNVQTL (Figure 1E).

DATABASE CONSTRUCTION AND CONTENT

Data summary of Pancan-MNVQTLdb

In Pancan-MNVQTLdb, the multi-omic data of >9100 patient samples (Table 1, Figure 2A) from TCGA (9) were used to perform MNVQTL analysis. The sample sizes of 33 cancer types ranged from 35 (cholangiocarcinoma, CHOL) to 1091 (breast invasive carcinoma, BRCA) (Table 1). From imputed SNP genotype data, we identified 223 759 unique MNVs, ranging from 62 514 in BRCA to 138 731 in acute myeloid leukemia (LAML), with an average of 113 049 MNVs per cancer type (Table 1).

Six molecular phenotypes were used for MNVQTL mapping (Table 2). For apaQTL identification, we downloaded the PDUI data from the TC3A data portal (31), containing PDUI values for an average of 4144 APA events per cancer among 32 TCGA cancer types. For eQTL, lncRNA-eQTL and miRNA-eQTL identification, we obtained gene expression profiles from the TCGA data portal (9). After filtering out the genes with low expression (median FPKM < 0.01 for eQTL and lncRNA-eQTL, and average TPM < 0.01 for miRNA-eQTL), an average of 16 686 coding genes, 13 159 lncRNA genes and 723 miRNA genes per cancer were retained. For sQTL identification, we downloaded the PSI data from the TCGASpliceSeq database (30). After filtering out the splicing array probes with a missing rate >0.1 or located on sex chromosomes, an average of 34 751 splicing array probes per cancer remained, ranging from 24 708 in UCEC to 43 938 in ESCA. For meQTL, we obtained the methylation data from TCGA (9), and an average of 380 563 methylation probes per cancer type were retained for meQTL analyses.

cis- and *trans*-MNVQTLs in Pancan-MNVQTLdb

For each molecular phenotype of each cancer type, the associations of MNV–molecular trait pairs were tested for *cis*- and *trans*-QTL mapping with a cutoff of FDR < 0.1 (Table 3). In apaQTL analysis, we identified 31 415 *cis*-apaQTLs and 878 *trans*-apaQTLs. For coding genes, lncRNA and miRNA expression QTLs, we identified 222 039 MNV–coding gene pairs, 191 332 MNV–lncRNA pairs and 1421 MNV–miRNA pairs in *cis*-eQTL analysis. In *trans*-eQTL analysis, we identified 996 MNV–coding gene pairs, 16 697 MNV–lncRNA pairs and 154 MNV–miRNA pairs. In sQTL analysis, 266 486 *cis*-sQTLs and 357 383 *trans*-sQTLs were identified. In the meQTL analysis, there were 685 128 *cis*-meQTLs and 26 273 *trans*-meQTLs identified.

Survival and GWAS-associated MNVQTLs

To identify MNVQTLs associated with cancer prognosis, we associated the MNVQTL genotypes with the survival time of patients downloaded from TCGA (9) and identified 46 173 MNVQTLs significantly (log-rank $P < 0.05$) related to overall survival. Specifically, we separately identified 31 484, 25 728 and 28 481 survival-associated MNVQTLs in the additive, recessive and dominant models. To further identify MNVQTLs associated with complex human diseases or traits, we mapped MNVQTL results into

Table 1. Sample sizes and the numbers of MNVs in Pancan-MNVQTLdb

Cancer type	No. of MNVs	Sample size in QTL analyses					
		apaQTL	eQTL	lncRNA-eQTL	miRNA-eQTL	meQTL	sQTL
ACC	86 722	77	77	77	77	76	77
BLCA	106 344	408	406	406	400	407	406
BRCA	62 514	1091	1091	1091	1059	784	1090
CESC	108 279	299	299	299	242	298	250
CHOL	98 469	36	36	36	36	35	36
COAD	114 555	285	285	285	282	275	285
DLBC	127 570	48	48	48	46	47	48
ESCA	115 155	184	165	165	168	183	180
GBM	118 365	150	149	149	149	50	150
HNSC	108 197	518	500	500	489	517	499
KICH	97 672	66	66	66	66	65	66
KIRC	118 716	525	526	526	507	315	527
KIRP	128 472	290	290	290	289	273	290
LAML	138 731	122	107	107	113	121	122
LGG	121 213	515	513	513	499	514	514
LIHC	104 191	369	369	369	363	368	369
LUAD	111 666	511	514	514	504	455	512
LUSC	92 112	500	500	500	469	370	500
MESO	126 960	87	87	87	82	86	87
OV	70 947	291	265	265	289	8	300
PAAD	133 398	178	178	178	178	177	178
PCPG	123 425	178	178	178	175	177	178
PRAD	127 053	494	494	494	476	493	494
READ	117 161	–	94	94	91	91	94
SARC	102 014	258	258	258	255	257	258
SKCM	128 871	103	103	103	97	103	103
STAD	109 256	414	376	376	409	371	414
TGCT	127 357	150	150	150	148	149	149
THCA	129 070	503	502	502	497	502	503
THYM	130 258	120	120	120	120	119	120
UCEC	129 128	176	176	176	175	171	176
UCS	93 616	56	56	56	56	55	56
UVM	123 160	80	80	80	77	79	80

GWAS regions and identified 119 762 MNVQTLs overlapping with GWAS loci.

Database construction of Pancan-MNVQTLdb

All results were stored in the MongoDB database (version 3.4.2). A user-friendly web interface, Pancan-MNVQTLdb (http://gong_lab.hzau.edu.cn/mnvQTLdb/), was constructed based on Flask (version 1.0.3) framework to support data browsing, searching and downloading. The database was running on Apache2 web server (version 2.4.18) and was compatible with modern browsers.

The function and usage of Pancan-MNVQTLdb

Pancan-MNVQTLdb provides a user-friendly web interface (http://gong_lab.hzau.edu.cn/mnvQTLdb/) for users to browse, visualize, search and download different types of MNVQTLs.

We provided a quick entry on the 'Home' page for users to conveniently access each module (Figure 2A). We also offered an aggregation search option for users. By entering an MNV, a gene or a genomic region, users could obtain integrated search results, including the information for all associations of the query MNV in each QTL type (Figure 2B). In the separate query pages for each QTL type (Figure 2C), users could obtain a table containing detailed information on query results on these pages (Figure 2D). In that table, users could also get the boxplot showing the distribution of the molecular phenotype in each genotype group (Figure 2E). The result table contained eight standard columns, i.e. cancer type, QTL type, MNV_ID and five additional columns of the association statistics, which are beta value, standard error, Pearson correlation, *P*-value and FDR. Other columns varied according to the molecular phenotype, showing the specific information of the corresponding molecular phenotype. For example, on the 'eQTL' page, the columns for molecular phenotype would be gene position, gene name and gene ID.

Similar to the 'QTL' pages, users could obtain a table containing the survival analysis results on the 'survival-QTL' page, including the analysis model, samples, median survival time and the log-rank *P*-value. Users could obtain a KM curve plot showing the difference in survival time among different genotype groups (Figure 2F). On the 'GWAS-QTL' page, users could get a table of MNVQTLs overlapping the LD regions of GWAS risk variants and the LD information between MNVs and GWAS risk variants, including R^2, *D*-prime and the associated traits.

In Pancan-MNVQTLdb, datasets for six types of MNVQTLs, survival-QTLs and GWAS-QTLs are freely available from the 'Download' page. Additionally, users can find the definition of MNV and a tutorial for using this database on the 'Help' page. Pancan-MNVQTLdb is open to any

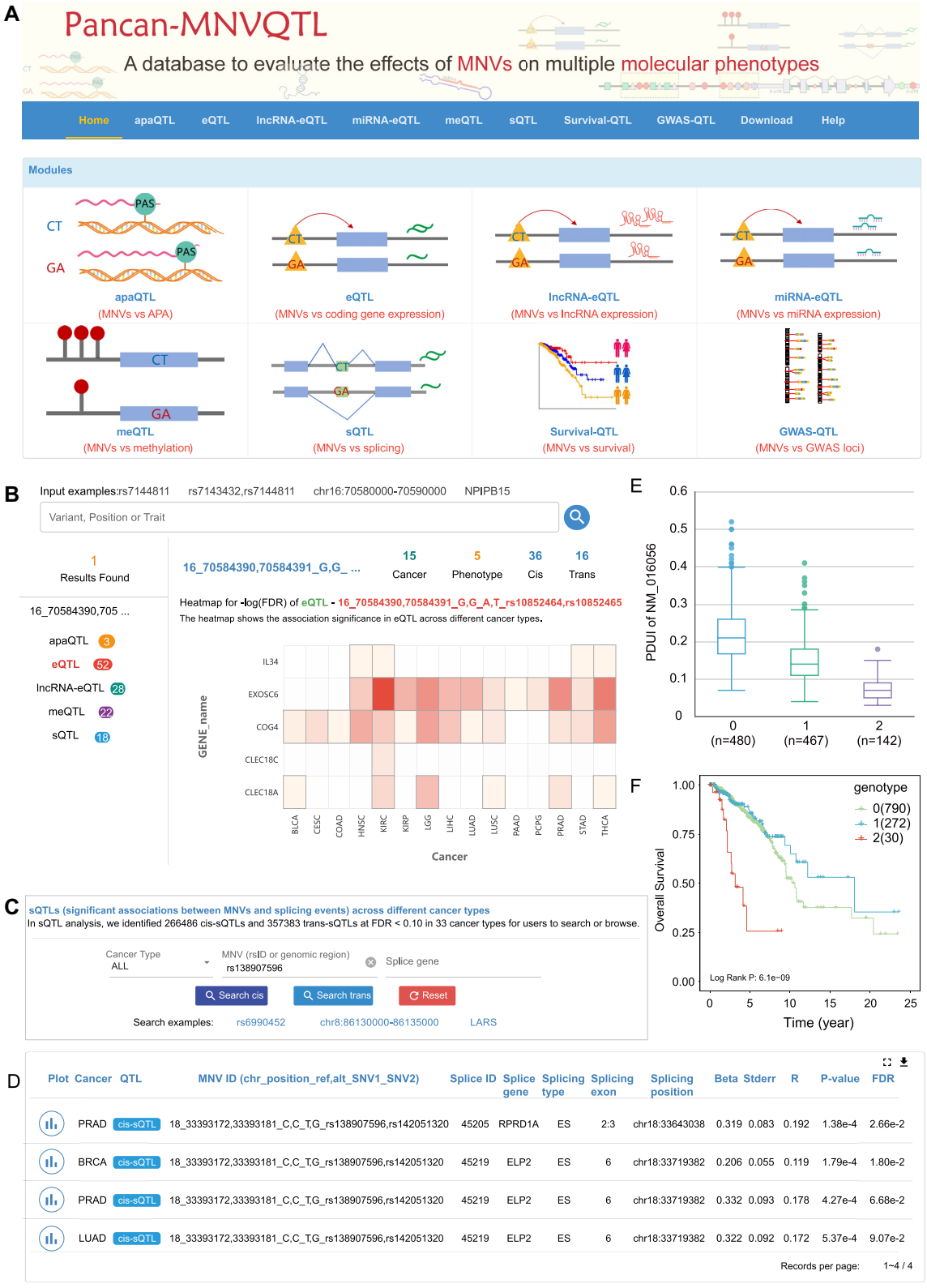

Figure 2. The interface of Pancan-MNVQTLdb. (**A**) Browser bar in Pancan-MNVQTLdb and main modules in Pancan-MNVQTLdb, including 'apaQTL', 'eQTL', 'lncRNA-eQTL', 'miRNA-eQTL', 'meQTL', 'sQTL', 'Survival-QTL', 'GWAS-QTL' and 'Download' modules. (**B**) Example of the aggregate search in the Pancan-MNVQTLdb. (**C**) The query page of the specific MNVQTL section (sQTL is shown). (**D**) Search results of sQTL dataset. (**E**) An example of the boxplot provided on the search result page. (**F**) An example of a KM plot provided on the survival-QTL result page.

Table 2. The numbers of molecular phenotypes in Pancan-MNVQTLdb

	Number of molecular phenotypes in QTL analyses					
Cancer type	APA event	Coding gene	lncRNA gene	miRNA gene	Methylation probe	Splice event
ACC	3115	16 098	10 387	744	407 747	26 621
BLCA	3781	16 564	11 962	761	372 097	32 126
BRCA	5380	16 777	13 315	646	368 273	38 429
CESC	3269	16 509	12 147	732	371 369	33 444
CHOL	3565	16 602	12 069	702	379 193	31 209
COAD	3357	16 553	12 127	665	387 590	27 467
DLBC	3659	16 095	11 354	733	381 312	26 278
ESCA	4511	17 317	18 932	695	373 181	43 938
GBM	5354	17 108	14 979	–	380 049	38 905
HNSC	4647	16 608	11 587	741	384 470	35 649
KICH	4478	16 467	12 767	638	387 055	39 172
KIRC	4907	16 846	14 727	587	375 521	39 697
KIRP	4356	16 544	12 665	670	385 983	33 439
LAML	3755	16 558	18 745	516	395 506	29 805
LGG	5252	16 953	14 201	771	384 927	41 897
LIHC	3128	15 895	9583	733	359 912	26 211
LUAD	4472	16 944	13 813	738	384 051	37 237
LUSC	5127	17 106	14 345	723	375 056	39 641
MESO	4000	16 611	12 179	712	375 122	36 011
OV	6175	17 123	16 412	727	384 253	41 416
PAAD	4467	17 114	13 181	688	375 660	39 105
PCPG	3697	16 456	11 809	779	389 173	34 322
PRAD	4705	16 848	12 918	610	378 419	37 655
READ	–	16 638	12 423	681	387 226	29 275
SARC	3911	16 349	11 226	645	359 148	33 923
SKCM	4180	16 267	11 027	796	376 260	34 943
STAD	6979	17 314	18 383	680	361 142	41 434
TGCT	4617	17 591	14 397	993	381 647	35 759
THCA	520	16 487	12 939	742	394 672	39 755
THYM	3774	16 758	13 177	951	377 828	33 235
UCEC	2589	16 801	12 326	748	385 968	24 708
UCS	3734	17 173	13 264	833	388 295	32 023
UVM	3150	15 550	8885	754	390 464	32 068

feedback with the email address provided at the bottom of the 'Help' page.

CONCLUSION AND FUTURE DIRECTIONS

In Pancan-MNVQTLdb, we identified MNVs in 33 human cancers. By associating them with six molecular phenotypes, we found considerable local and distal associations between MNVs and these regulation-related phenotypes. Furthermore, to further understand the potential function of these MNVs, we tested the association of MNVs with overall patient survival time and examined the association of MNVs with disease-related variants. As a result, we found that many MNVs are significantly associated with cancer prognosis or complex disease. To facilitate easy access to the abundant data we generated, we developed Pancan-MNVQTLdb, a comprehensive resource of MNVs associated with multiple molecular phenotypes. To the best of our knowledge, this is the first database systematically evaluating the effects of the MNVs on molecular phenotypes in multiple cancer types.

While Pancan-MNVQTLdb provides comprehensive data on associations between MNVs and multiple molecular phenotypes, there are some limitations in this study. First, we did not analyze the effects of rare MNVs in our QTL analyses. Rare variants usually need large sample size populations to analyze their effect, but most cancer types in TCGA only have ∼400 samples. Rare MNVs (MAF < 0.01) often only have <3 samples with homozygote alternative alleles, which would lead to high false positives in QTL analysis. Thus, we did not analyze rare MNVs in this study. In the future, we will try to analyze the functions of MNVs by integrating large sample-size data and designing possible algorithms specific to rare variants. Second, MNV is a newly discovered variant type with limited methods explicitly designed for MNVs, which poses an unprecedented challenge to deciphering the function of MNVs, especially distinguishing whether MNVs or SNPs caused the effect on molecular phenotype. Further studies, especially studies with precise fine mapping and biological experiments, are needed to help clarify the causative polymorphism in these MNVs.

Nevertheless, recent studies have suggested that MNVs may have a different impact from single variants or even be more harmful. Therefore, investigating the functions of MNVs will advance the understanding of genetic variants and provide opportunities to bridge the knowledge gap from multiple variants in sequence to phenotypes. Currently, MNV studies are still much less than SNV studies. In the future, with the increasing number of genotype data and molecular phenotype data from large consortium projects, we will consider updating the Pancan-MNVQTLdb database and maintaining it as a valuable resource for the genetic and cancer research community.

Table 3. The numbers of *cis*- and *trans*-MNVQTLs in Pancan-MNVQTLdb

Cancer type	Number of *cis*-QTLs						Number of *trans*-QTLs					
	apaQTL	eQTL	lncRNA-eQTL	miRNA-eQTL	meQTL	sQTL	apaQTL	eQTL	lncRNA-eQTL	miRNA-eQTL	meQTL	sQTL
ACC	114	141	229	0	2232	675	16	2	34	0	14	22 773
BLCA	1077	5347	6079	48	26 367	8751	40	47	562	2	744	2471
BRCA	2952	15 489	16 787	88	47 013	19 038	56	86	1621	17	3088	2377
CESC	736	3285	3839	62	16 453	7286	47	22	482	0	232	4919
CHOL	1	1	9	0	0	9	6	0	0	0	1	5361
COAD	613	5655	4822	72	17 381	8244	38	59	521	0	424	9
DLBC	41	19	61	2	81	304	0	0	0	0	0	722
ESCA	774	680	1285	15	7199	6641	1	0	193	0	65	15 154
GBM	1082	1741	2146	0	261	6434	0	4	146	0	0	18 697
HNSC	2222	11 658	9403	155	60 066	13 965	26	63	938	0	1393	1871
KICH	165	232	341	2	694	1155	2	0	7	0	2	34 663
KIRC	2344	26 167	19 662	82	46 489	20 699	57	88	1508	15	942	3006
KIRP	889	8863	6735	66	22 815	10 428	75	15	662	2	536	8894
LAML	170	1423	1681	7	6962	1434	0	4	206	0	98	29 418
LGG	2932	25 826	22 946	133	61 047	24 944	78	44	1597	26	2366	3185
LIHC	640	5289	5079	54	20 238	5953	52	7	499	3	410	4115
LUAD	2124	12 168	11 834	100	51 644	14 479	77	63	925	6	1194	2368
LUSC	2105	9021	10 029	97	28 432	14 615	28	58	965	12	620	3858
MESO	242	335	416	3	2572	2127	13	0	99	5	7	9418
OV	1681	1982	2699	26	0	7734	2	16	328	0	0	147
PAAD	757	3472	3682	12	19 525	6769	29	7	341	5	214	28 545
PCPG	569	2753	3113	7	11 098	5925	65	9	262	1	303	16 834
PRAD	2987	27 584	19 529	101	88 604	20 328	83	120	1571	13	10 564	4133
READ	0	497	674	15	3550	2429	0	0	101	0	17	7633
SARC	522	1791	2462	14	8823	6858	2	4	249	1	101	8662
SKCM	385	296	328	0	3457	2549	63	0	98	2	13	28 096
STAD	1684	5809	5336	60	25 021	8765	2	37	505	0	412	1655
TGCT	431	1906	2673	46	7884	5620	0	0	200	1	215	24 744
THCA	318	39 298	22 771	133	80 511	23 126	1	238	1570	35	2006	7460
THYM	354	1972	3088	14	10 193	4457	2	0	236	8	183	9923
UCEC	240	749	1056	5	4687	2878	15	3	202	0	75	3119
UCS	41	26	20	0	410	262	1	0	4	0	0	28 274
UVM	223	564	518	2	3419	1605	1	0	65	0	34	14 879

DATA AVAILABILITY

Pancan-MNVQTLdb is freely available to the public without registration or login requirements at http://gong_lab.hzau.edu.cn/mnvQTLdb/.

FUNDING

National Key Research and Development Program of China [2021YFF0703703 to J.G.]; National Natural Science Foundation of China (NSFC) [31970644 to J.G.]; Natural Science Foundation of Hubei Province [2021CFB404 to X.H.N.]; Huazhong Agricultural University Scientific & Technological Self-innovation Foundation [11041810351 to J.G., 2662022XXYJ008 to X.H.N.].
Conflict of interest statement. None declared.

REFERENCES

1. Ding,L., Wendl,M.C., McMichael,J.F. and Raphael,B.J. (2014) Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.*, **15**, 556–570.
2. Shastry,B.S. (2002) SNP alleles in human disease and evolution. *J. Hum. Genet.*, **47**, 561–566.
3. Dehghan,A. (2018) Genome-wide association studies. *Methods Mol. Biol.*, **1793**, 37–49.
4. Wu,C., Miao,X., Huang,L., Che,X., Jiang,G., Yu,D., Yang,X., Cao,G., Hu,Z., Zhou,Y. *et al.* (2011) Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nat. Genet.*, **44**, 62–66.
5. Gallagher,M.D. and Chen-Plotkin,A.S. (2018) The post-GWAS era: from association to function. *Am. J. Hum. Genet.*, **102**, 717–730.
6. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
7. Wang,Q., Pierce-Hoffman,E., Cummings,B.B., Alföldi,J., Francioli,L.C., Gauthier,L.D., Hill,A.J., O'Donnell-Luria,A.H., Genome,Aggregation Database Production Team, Genome,Aggregation Database Consortium *et al.* (2020) Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat. Commun.*, **11**, 2539.
8. Srinivasan,S., Kalinava,N., Aldana,R., Li,Z., van Hagen,S., Rodenburg,S.Y.A., Wind-Rotolo,M., Qian,X., Sasson,A.S., Tang,H. *et al.* (2021) Misannotated multi-nucleotide variants in public cancer genomics datasets lead to inaccurate mutation calls with significant implications. *Cancer Res.*, **81**, 282–288.
9. Hutter,C. and Zenklusen,J.C. (2018) The Cancer Genome Atlas: creating lasting value beyond its data. *Cell*, **173**, 283–285.
10. Kaplanis,J., Akawi,N., Gallone,G., McRae,J.F., Prigmore,E., Wright,C.F., Fitzpatrick,D.R., Firth,H.V., Barrett,J.C., Hurles,M.E. *et al.* (2019) Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. *Genome Res.*, **29**, 1047–1056.
11. Gong,J., Mei,S., Liu,C., Xiang,Y., Ye,Y., Zhang,Z., Feng,J., Liu,R., Diao,L., Guo,A.-Y. *et al.* (2018) PancanQTL: systematic identification of *cis*-eQTLs and *trans*-eQTLs in 33 cancer types. *Nucleic Acids Res.*, **46**, D971–D976.
12. Do,H. and Kim,W. (2018) Roles of oncogenic long non-coding RNAs in cancer development. *Genomics Inform.*, **16**, e18.

13. Guo,Z., Zhu,H., Xu,W., Wang,X., Liu,H., Wu,Y., Wang,M., Chu,H. and Zhang,Z. (2020) Alternative splicing related genetic variants contribute to bladder cancer risk. *Mol. Carcinog.*, **59**, 923–929.
14. Li,M., Li,Y.-P., Deng,H.-L., Wang,M.-Q., Chen,Y., Zhang,Y.-F., Wang,J. and Dang,S.-S. (2021) DNA methylation and SNP in IFITM3 are correlated with hand, foot and mouth disease caused by enterovirus 71. *Int. J. Infect. Dis.*, **105**, 199–208.
15. Okumura,K., Saito,M., Isogai,E., Tokunaga,Y., Hasegawa,Y., Araki,K. and Wakabayashi,Y. (2022) Functional polymorphism in Pak1-3′ untranslated region alters skin tumor susceptibility by alternative polyadenylation. *J. Invest. Dermatol.*, **142**, 2323–2333.
16. Beesley,J., Sivakumaran,H., Moradi Marjaneh,M., Shi,W., Hillman,K.M., Kaufmann,S., Hussein,N., Kar,S., Lima,L.G., Ham,S. *et al.* (2020) eQTL colocalization analyses identify NTN4 as a candidate breast cancer risk gene. *Am. J. Hum. Genet.*, **107**, 778–787.
17. Zhang,T., Choi,J., Dilshat,R., Einarsdóttir,B.Ó., Kovacs,M.A., Xu,M., Malasky,M., Chowdhury,S., Jones,K., Bishop,D.T. *et al.* (2021) Cell-type-specific meQTLs extend melanoma GWAS annotation beyond eQTLs and inform melanocyte gene-regulatory mechanisms. *Am. J. Hum. Genet.*, **108**, 1631–1646.
18. Pan,Q., Liu,Y.-J., Bai,X.-F., Han,X.-L., Jiang,Y., Ai,B., Shi,S.-S., Wang,F., Xu,M.-C., Wang,Y.-Z. *et al.* (2021) VARAdb: a comprehensive variation annotation database for human. *Nucleic Acids Res.*, **49**, D1431–D1444.
19. Huang,D., Zhou,Y., Yi,X., Fan,X., Wang,J., Yao,H., Sham,P.C., Hao,J., Chen,K. and Li,M.J. (2022) VannoPortal: multiscale functional annotation of human genetic variants for interrogating molecular mechanism of traits and diseases. *Nucleic Acids Res.*, **50**, D1408–D1416.
20. Gong,J., Wan,H., Mei,S., Ruan,H., Zhang,Z., Liu,C., Guo,A.-Y., Diao,L., Miao,X. and Han,L. (2019) Pancan-meQTL: a database to systematically evaluate the effects of genetic variants on methylation in human cancer. *Nucleic Acids Res.*, **47**, D1066–D1072.
21. Tian,J., Wang,Z., Mei,S., Yang,N., Yang,Y., Ke,J., Zhu,Y., Gong,Y., Zou,D., Peng,X. *et al.* (2019) CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer. *Nucleic Acids Res.*, **47**, D909–D916.
22. Li,J., Xue,Y., Amin,M.T., Yang,Y., Yang,J., Zhang,W., Yang,W., Niu,X., Zhang,H.-Y. and Gong,J. (2020) ncRNA-eQTL: a database to systematically evaluate the effects of SNPs on non-coding RNA expression across cancer types. *Nucleic Acids Res.*, **48**, D956–D963.
23. Yang,Y., Zhang,Q., Miao,Y.-R., Yang,J., Yang,W., Yu,F., Wang,D., Guo,A.-Y. and Gong,J. (2020) SNP2APA: a database for evaluating effects of genetic variants on alternative polyadenylation in human cancers. *Nucleic Acids Res.*, **48**, D226–D232.
24. Xin,J., Du,M., Jiang,X., Wu,Y., Ben,S., Zheng,R., Chu,H., Li,S., Zhang,Z. and Wang,M. (2021) Systematic evaluation of the effects of genetic variants on PIWI-interacting RNA expression across 33 cancer types. *Nucleic Acids Res.*, **49**, 90–97.
25. Wang,D., Wu,X., Jiang,G., Yang,J., Yu,Z., Yang,Y., Yang,W., Niu,X., Tang,K. and Gong,J. (2022) Systematic analysis of the effects of genetic variants on chromatin accessibility to decipher functional variants in non-coding regions. *Front. Oncol.*, **12**, 1035855.
26. Zhang,Z., Luo,M., Li,Q., Liu,Y., Lussier,C., Zhang,J., Ye,Y., Guo,A.-Y. and Han,L. (2022) Genetic, pharmacogenomic, and immune landscapes of enhancer RNAs across human cancers. *Cancer Res.*, **82**, 785–790.
27. Tian,J., Cai,Y., Li,Y., Lu,Z., Huang,J., Deng,Y., Yang,N., Wang,X., Ying,P., Zhang,S. *et al.* (2021) CancerImmunityQTL: a database to systematically evaluate the impact of genetic variants on immune infiltration in human cancer. *Nucleic Acids Res.*, **49**, D1065–D1073.
28. Howie,B.N., Donnelly,P. and Marchini,J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.
29. O'Connell,J., Gurdasani,D., Delaneau,O., Pirastu,N., Ulivi,S., Cocca,M., Traglia,M., Huang,J., Huffman,J.E., Rudan,I. *et al.* (2014) A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.*, **10**, e1004234.
30. Ryan,M., Wong,W.C., Brown,R., Akbani,R., Su,X., Broom,B., Melott,J. and Weinstein,J. (2016) TCGASpliceSeq: a compendium of alternative mRNA splicing in cancer. *Nucleic Acids Res.*, **44**, D1018–1022.
31. Feng,X., Li,L., Wagner,E.J. and Li,W. (2018) TC3A: The Cancer 3′ UTR Atlas. *Nucleic Acids Res.*, **46**, D1027–D1030.
32. Kang,H.M., Ye,C. and Eskin,E. (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**, 1909–1925.
33. Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, 1724–1735.
34. GTEx,Consortium (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
35. Stegle,O., Parts,L., Piipari,M., Winn,J. and Durbin,R. (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, **7**, 500–507.
36. Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.
37. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A., Morales,J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.
38. Machiela,M.J. and Chanock,S.J. (2015) LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, **31**, 3555–3557.