O'Brien *et al. Genome Medicine* (2018) 10:16
https://doi.org/10.1186/s13073-018-0522-9

Genome Medicine

RESEARCH Open Access

Weak sharing of genetic association signals in three lung cancer subtypes: evidence at the SNP, gene, regulation, and pathway levels

Timothy D. O'Brien[1,2,3], Peilin Jia[3], Neil E. Caporaso[4], Maria Teresa Landi[4] and Zhongming Zhao[1,2,3,5*]

Abstract

Background: There are two main types of lung cancer: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC has many subtypes, but the two most common are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). These subtypes are mainly classified by physiological and pathological characteristics, although there is increasing evidence of genetic and molecular differences as well. Although some work has been done at the somatic level to explore the genetic and biological differences among subtypes, little work has been done that interrogates these differences at the germline level to characterize the unique and shared susceptibility genes for each subtype.

Methods: We used single-nucleotide polymorphisms (SNPs) from a genome-wide association study (GWAS) of European samples to interrogate the similarity of the subtypes at the SNP, gene, pathway, and regulatory levels. We expanded these genotyped SNPs to include all SNPs in linkage disequilibrium (LD) using data from the 1000 Genomes Project. We mapped these SNPs to several lung tissue expression quantitative trait loci (eQTL) and enhancer datasets to identify regulatory SNPs and their target genes. We used these genes to perform a biological pathway analysis for each subtype.

Results: We identified 8295, 8734, and 8361 SNPs with moderate association signals for LUAD, LUSC, and SCLC, respectively. Those SNPs had $p < 1 \times 10^{-3}$ in the original GWAS or were within LD ($r^2 > 0.8$, Europeans) to the genotyped SNPs. We identified 215, 320, and 172 disease-associated genes for LUAD, LUSC, and SCLC, respectively. Only five genes (*CHRNA5*, *IDH3A*, *PSMA4*, *RP11-650 L12.2*, and *TBC1D2B*) overlapped all subtypes. Furthermore, we observed only two pathways from the Kyoto Encyclopedia of Genes and Genomes shared by all subtypes. At the regulatory level, only three eQTL target genes and two enhancer target genes overlapped between all subtypes.

Conclusions: Our results suggest that the three lung cancer subtypes do not share much genetic signal at the SNP, gene, pathway, or regulatory level, which differs from the common subtype classification based upon histology. However, three (*CHRNA5, IDH3A,* and *PSMA4*) of the five genes shared between the subtypes are well-known lung cancer genes that may act as general lung cancer genes regardless of subtype.

Keywords: GWAS, eQTL, Enhancer, Lung cancer subtype, Functional genomics, Pathway analysis

* Correspondence: zhongming.zhao@uth.tmc.edu
[1]Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN, USA
[2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
Full list of author information is available at the end of the article

Background

Lung cancer is the second most commonly occurring cancer in the United States and is responsible for the most cancer-related deaths for both men and women, excluding data from skin cancer [1]. Although environmental risk factors such as smoking have major contributions to lung cancer development [2], there is also a genetic component, and heritability estimates of genetic risk for lung cancer range from 8% to 14% [3, 4]. Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are the two main histological types of lung cancer [5]. The two main subtypes of NSCLC are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). LUAD and LUSC comprise the vast majority of newly reported lung cancer cases, while SCLC comprises only a small subset (~15%) [6]. These subtypes differ in their location within the lung as well as the cell type of origin [7] and, therefore, may have different underlying disease etiologies. LUAD is the most researched lung cancer subtype, and studies have identified genomic alterations and actionable mutations [8–11]. Additionally, genomic alterations have been discovered in LUSC [12–14] and SCLC [15, 16]. Although it was discovered from these studies that few somatically mutated genes overlap all three subtypes, most of these studies focused on somatic mutations. Few studies have expanded the analysis to the germline genome.

In 2014, Hoadley et al. performed an integrative analysis to cluster cancers using 12 different cancer types from The Cancer Genome Atlas (TCGA) project [17]. They discovered that LUAD is a separate cluster and is distinct from LUSC, which clusters with the other squamous-like cancer types. In 2016, Campbell et al. compared somatic genomic alterations of LUAD and LUSC using over 1000 combined somatic tumor tissue samples [18]. They found that only six mutated genes overlapped both subtypes and that each subtype shared only about 25% of copy number alterations. Their work supports the conclusion that both subtypes are very distinct diseases. Common germline variation associated with lung cancer has also been studied for more than one subtype using genome-wide association studies (GWASs) [19–22].

Several GWASs have discovered common genetic variation associated with lung cancer risk [19–32]. However, few studies used data for all three subtypes [19, 20, 23, 26, 28, 30, 31]. Additionally, most of these findings did not reach the stringent genome-wide significance for a GWAS ($p < 5 \times 10^{-8}$), and most of the genome-level significant single-nucleotide polymorphisms (SNPs) were located within non-coding regions of the genome, making it difficult to infer the underlying mechanism of the significant variants that could cause disease. Recent studies have shown that these marginally significant SNPs found from GWASs within non-coding regions of the genome may function in regulatory roles [33, 34]. Therefore, these results can be used to obtain a set of regulated genes to investigate and compare the similarity of the three lung cancer subtypes at the germline gene level and at the regulation level.

In this study, we first selected a set of SNPs with moderate association signals ($p < 1 \times 10^{-3}$) from the summary results of a prior GWAS that covered three lung cancer subtypes (LUAD, LUSC, and SCLC). Then, we identified and compared regulatory variants associated with the three subtypes of lung cancer, as well as their target genes. We used these results to investigate the similarity of the subtypes at the SNP, gene, regulatory, and pathway levels. We first remapped these SNPs to an updated human genome reference (hg19) and expanded them using linkage disequilibrium (LD) patterns from a European population. We used this final set of SNPs to examine several lung tissue expression quantitative trait loci (eQTL) and enhancer datasets for evidence of a regulatory function for each SNP and identified their target genes. We compared the target genes of these regulatory SNPs and observed that only five genes overlapped all three subtypes. We also observed a weak overlap among all three subtypes across all comparisons. Through this analysis, we identified many genes that might have an important association with lung cancer for each specific subtype. Follow-up studies on these genes may lead to a better understanding not only of the genes themselves, but also the underlying biology that differentiates these subtypes of lung cancer. Our results provide insights into the distinct genetic components among the three lung cancer subtypes.

Methods

GWAS dataset

We previously performed a multi-site GWAS for lung cancer in a European population and analyzed each sample by lung cancer subtype for the National Cancer Institute's GWAS for lung cancer (more details are available in the original publication [20]). Briefly, this GWAS for lung cancer used cases and controls from four different studies: Environment and Genetics in Lung Cancer Etiology (EAGLE), Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC), Prostate, Lung, Colon, Ovary (PLCO) screening trial, and Cancer Prevention Study II (CPS-II). After the quality control of the genotyping results, there remained 5739 cases and 5848 controls of European ancestry and 515,922 SNPs. The analysis was stratified by lung cancer subtype with 1730 LUAD cases, 1400 LUSC cases, 678 SCLC cases, and 5848 shared controls. It used unconditional logistic regression. We used the full set of significant lung cancer SNPs ($p < 1 \times 10^{-3}$) separated by subtype for this analysis.

Genomic annotation of SNPs

The online web tool SNP Nexus [35] (http://snp-nexus.org/) was used to annotate the genomic location of the significant SNPs by lung cancer subtype based on the NCBI36/hg18 genome assembly. We used the UCSC hg18 gene definitions for the genomic annotation of each region.

Conversion of hg18 SNPs to hg19 SNPs

The results from the lung cancer GWAS were originally generated using coordinates from the hg18 reference of the human genome. We converted these SNPs to hg19 coordinates using the online tool Remap from the National Center for Biotechnology Information (NCBI) with default settings (http://www.ncbi.nlm.nih.gov/genome/tools/remap). This conversion allowed us to map the SNPs to the regulatory annotation information, which were based on hg19 coordinates.

We used these updated hg19 coordinates for the SNPs to obtain the updated SNP rsID numbers using dbSNP data (build 142) from NCBI to account for any SNPs that may have been merged between assemblies.

Identification of SNPs in LD with the genotyped SNPs

For each SNP, we retrieved all other SNPs in a 1-Mb region both upstream and downstream from the SNP site using Tabix [36] (version 0.2.5). We obtained the SNP data from the European super population group from the 1000 Genomes Phase III data (v5.20120502). Vcftools [37] (version 0.1.12b) was used to convert the Tabix vcf files to the plink-tped file format. Then we used the 1000 Genomes data for each SNP and applied PLINK [38] (version 1.07) to identify the final set of SNPs that were in the same LD with the tagging SNPs using $r^2 > 0.8$ with 1 Mb upstream and downstream of the SNP. The LD results from PLINK were combined for every SNP and any SNPs in LD that were duplicated across all SNP sets were removed.

Randomization for overlapping SNPs

All LD-based pruning of SNPs was performed using the PLINK formatted 1000 Genomes Phase III European dataset. To identify a set of more independent SNPs (more independent and not purely independent) in each subtype, SNPs with $p < 1 \times 10^{-3}$ from the GWAS summary results were extracted for LUAD, LUSC, and SCLC, and PLINK was used to prune out a set of SNPs with no strong linkage using the `indep-pairwise` function. The r^2 used for all LD trimming was 0.5. These results were used to identify the new overlap of SNPs between LUAD, LUSC, and SCLC. To trim the background set of SNPs for the randomization, the entire set of SNPs genotyped and reported for each subtype was imported into PLINK with the same function and options.

The same number of SNPs for each subtype were randomly selected 10,000 times from the background pool of SNPs without strong linkage from the genotype chip in R. For each random selection, we determined the number of overlapping SNPs to identify the level of overlap that may occur by chance.

Genotype-Tissue Expression eQTLs

The full set of significant human-tissue-specific eQTLs version 6 (V6) was downloaded from the Genotype-Tissue Expression (GTEx) website (https://www.gtexportal.org) on 22 February 2016. The eQTLs were identified using linear regression with the tool Matrix eQTL [39] with a ±1-Mb region around the transcription start site in each individual tissue that had >70 samples. The significance of the eQTLs was determined by empirical p values using permutations followed by a Storey false discovery rate. The eQTLs with a q value ≤5% were considered significant.

We also downloaded the full set of all multi-tissue eQTLs for nine different tissue types for the pilot phase of the GTEx Project on 11 June 2015. This file contained eQTLs discovered using two different methods, the University of Chicago model [40] and the University of North Carolina model [41], which are fully explained in the respective publications. Additionally, a file was included that contained the average between both methods including calculated posterior probabilities for every gene–SNP pair titled `res_final_amean_com_genes_com_snps_all.txt`. The whole SNP set (including LD SNPs) was used to detect eQTLs in this dataset. We plotted the distribution of posterior probabilities of all the eQTLs found using the SNPs and defined an eQTL as significant if its posterior probability was >80% (Additional file 1: Figure S1). We removed all duplicated genes in each subtype to obtain the final GTEx set of genes.

Lung tissue eQTLs from the Hao et al. study

Hao et al. [42] investigated how genetic variation affects gene expression levels in human lung tissues. They used this dataset to interrogate SNPs associated with asthma. The authors used lung tissue and blood from more than 1000 patients across three cohorts to identify a set of eQTLs in lung tissue. We downloaded the entire set of cis-eQTLs identified from this study with the false discovery rate at 10%. We removed the target genes without annotated gene names. We also merged duplicate probes that specified the same genes into a single gene.

FANTOM5 transcribed enhancers

The FANTOM consortium aims to identify and assign regulatory function to the mammalian genome. Part of this comprehensive project is to identify all transcribed enhancers and promoters in multiple human cell lines and tissue types. The entire set of permissive enhancers found in the FANTOM5 data was downloaded in bed file format from http://enhancer.binf.ku.dk/presets/permissive_enhancers.bed on 26 August 2015. The gene-report function was used in PLINK to search for any SNPs that were located within permissive enhancer regions. SNPs that were located within each enhancer region were then matched with the set of correlated expressed promoters for FANTOM5 enhancer transcription start sites downloaded from http://enhancer.binf.ku.dk/presets/enhancer_tss_associations.bed on 25 August 2015.

IM-PET predicted enhancers

He et al. [43] developed a novel approach for identifying the target genes of histone-derived enhancers using a random forest classifier. The authors used this tool, Integrated Methods for Predicting Enhancer Targets (IM-PET), to define a set of enhancer target genes for 12 different cell types. We used the results for two lung cell types, IMR90 and NHLF, for our analysis. We used bedtools [44] version 2.17.0 to identify lung cancer SNPs within the enhancer regions that had an associated target gene. To remove non-expressed genes, we filtered the results to remove target genes with reads per kilobase per million mapped reads of 0. The enhancer targets were originally formatted as Ensembl-defined transcripts, so we converted them to gene symbols using the BioMart tool from Ensembl using the archived site pertaining to genome assembly GRCh37.p13 [45].

Identification of independent loci for the identified germline genes

biomaRt [46] was used to annotate the genomic locations for the germline-regulated genes discovered from each dataset for each subtype using gene start and stop coordinates from Ensembl gene definitions using genome build GRCh37.3. Genomic locations that were not defined from Ensembl were manually annotated using NCBI's Gene online web resource https://www.ncbi.nlm.nih.gov/gene. The function `cluster` from bedtools [44] was used to cluster the genes into independent 1-Mb regions.

Pathway enrichment analysis

The final set of germline-regulated genes was uploaded to the WebGestalt online resource [47]. The hypergeometric test was used for enrichment with specific pathways followed by Benjamini and Hochberg multiple test correction [48].

GWAS Catalog SNPs

We downloaded all SNPs from the GWAS Catalog using the search term "lung cancer" on 13 January 2016. We removed SNPs where the initial or replication population was other than European. We also removed SNPs that were reported in the original lung cancer report [20] because they were used in our analyses.

Principal component analysis of TCGA germline genotyped SNPs

TCGA germline genotype level 2 data for six cancer types (LUAD, LUSC, head and neck squamous cell carcinoma, bladder urothelial carcinoma, glioblastoma multiforme, and lower grade glioma) were downloaded from the legacy archive of the data portal of the National Cancer Institute's Genomic Data Commons using its data transfer tool after obtaining permission from the database of Genotypes and Phenotypes (dbGaP). The normal blood samples were extracted from these sets to use in the analyses. These normal blood samples were then filtered to exclude non-white and white Hispanic samples as defined by the clinical data from TCGA. These birdseed genotype formatted files were then altered for use in PLINK as follows. First, all low-confidence calls (confidence > 0.1) were initially recoded as −9 in place of the 0,1,2 allele birdseed conventions. Second, each sample was merged together for each cancer type to obtain a matrix of the number of genotyped SNPs (906,600) times the number of samples. Third, the hg19/b37 Affymetrix mapping file was downloaded from the Affymetrix website and merged with the probe IDs in the birdseed matrices. Fourth, the Affymetrix annotation file was used to generate a PLINK format map file. Fifth, the alleles in the birdseed files were recoded to their appropriate bases according to the Affymetrix annotation file. After converting the files to PLINK format, all samples across all cancer types were merged together into one matrix using PLINK's merge-list function. Triallelic SNPs were further removed to obtain the final merged genotype matrix.

To run the principal component analysis (PCA) function in PLINK, the SNPs were first filtered by LD r^2 0.5 using 1000 Genomes Phase III European only data downloaded from the VEGAS2 website. After an LD trim, the PCA function was run on all samples. After visualization of the top two principal components, we determined a set of outliers from multiple cancer types and ran the PCA function again after removing these outliers.

Analysis of the overlap of SNPs and gene sets

The R package UpSetR [49] was used to make the plots of the overlapping SNPs and gene sets.

Results

Description of data

We obtained SNPs ($p < 1 \times 10^{-3}$) for three lung cancer subtypes, LUAD, LUSC, and SCLC, from the results of the National Cancer Institute's GWAS for lung cancer [20].

This GWAS utilized cases and controls from four smaller studies: Environment and Genetics in Lung Cancer Etiology (EAGLE), Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC), Prostate, Lung, Colon, Ovary (PLCO) screening trial, and Cancer Prevention Study II (CPS-II) nutrition cohort. Table 1 shows the total number of cases genotyped for each subtype, the total number of SNPs discovered by selection criterion ($p < 1 \times 10^{-3}$), and the distribution of their locations within the genome. We also show the overlap of SNPs per subtype in Additional file 1: Figure S2A. We found that, like many GWASs for various disease types, only 2–3% of variants were located within coding regions of the genome.

Weak sharing of genetic association signals

A direct comparison using genotyped SNPs revealed that ten SNPs at $p < 1\times 10^{-3}$ overlapped among all three subtypes (Additional file 1: Figures S2A and S3). To determine if this overlap of SNPs was different from what would be expected by chance, we conducted a randomization test through random resampling of the genotyped SNPs on the GWAS chip, which did not require individual genotyping data (see "Methods"). To avoid a randomization analysis that could be biased due to SNPs in strong LD, we pruned the original 515,922 SNPs as genotyped on the chip (Illumina HumanHap 550) that passed a quality check to obtain a set of SNPs that are assumed unrelated or weakly related at $r^2 = 0.5$ (see "Methods"). These SNPs (234,859 after LD pruning) served as the pool for our randomization test. We provide the details of this SNP selection in Additional file 1: Figure S3. After LD pruning of these SNPs, we discovered that only one SNP, compared to ten SNPs from the original list, overlapped across all three lung cancer subtypes (Additional file 1: Figure S2C). This one SNP was rs578776, on chromosome 15 in the 3' untranslated region of *CHRNA3,* in the chr.15q25 locus known to be associated with different histology subtypes of lung cancer [50]. For the randomization test, we randomly chose the same number of SNPs for each subtype after LD pruning 10,000 times. Each time, we compared the three randomly selected size-matched sets of SNPs representing three subtypes and recorded the overlapping SNPs. After the 10,000 randomization trials, we observed ten times that one SNP overlapped among the three random sets of SNPs (10/10,000), while in the remaining 9990 sets, no overlap was observed (Additional file 1: Table S1). We observed no instances of an overlap greater than one SNP. Given the large number of SNPs in the pool, it was expected that there would not be many overlapping SNPs. Thus, the discovery of one overlapping SNP among the three lung cancer subtypes is likely within random expectation with a chance of 0.001 according to our randomization test. Therefore, we conclude that there is no strong evidence that the one overlapping SNP we observed is higher than randomly expected.

Table 1 Summary of data from lung cancer genome-wide association studies

Subtype	Sample size	SNPs ($p < 1 \times 10^{-3}$)				
		Total SNPs	Coding	Intron	UTR	Intergenic
LUAD	1730	544	13	228	7	296
LUSC	1400	598	18	299	16	265
SCLC	678	558	14	247	10	287

LUAD lung adenocarcinoma, *LUSC* lung squamous cell carcinoma, *SCLC* small cell lung cancer, *SNP* single-nucleotide polymorphism, *UTR* untranslated region

SNP expansion

To obtain a comprehensive annotation of the SNPs, we expanded our SNP list to include those that are in strong LD with SNPs from the GWAS results at $p < 1\times 10^{-3}$. We mapped all SNPs from genome build hg18 to genome build hg19 using the NCBI tool Remap (http://www.ncbi.nlm.nih.gov/genome/tools/remap) and obtained updated SNP rsID numbers using data from dbSNP build 142. We expanded the initial set of genotyped SNPs to include all SNPs in LD within 1 Mb of the genotyped SNP based on data from the European population super group from Phase III of the 1000 Genomes Project [51]. Table 2 outlines the results of this SNP expansion for each subtype. After we removed duplicated SNPs within each subtype, we found 8295 SNPs associated with LUAD, 8734 with LUSC, and 8361 with SCLC, among which 167 SNPs overlapped between all three subtypes (Additional file 1: Figure S2B). We next used the final subtype-specific sets of SNPs from our LD expansion (Table 2) for the subsequent interrogation of regulatory function (Fig. 1).

Lung tissue eQTLs

We first utilized three lung eQTL datasets to annotate the SNPs. The first lung eQTL dataset was retrieved from the GTEx project [52]. Using this dataset, we found 1297 SNPs for LUAD, 1429 for LUSC, and 1171 for SCLC (Fig. 2a) that acted as eQTLs using a set of precompiled significant lung-tissue-specific eQTLs from GTEx. To explore all eQTLs for the lung, including non-tissue-specific eQTLs, we used a second set of eQTLs identified using a multi-tissue model from GTEx (Additional file 1: Figure S1). We combined the single

Table 2 Summary of SNP results and LD expansion

	LUAD	LUSC	SCLC
Number of SNPs (GWAS, $p < 10^{-3}$)	544	598	558
Number of SNPs (LD, r^2 0.8, within 1 Mb)	14,312	16,021	13,104
Number of final SNPs	8295	8734	8361

LD linkage disequilibrium, *LUAD* lung adenocarcinoma, *LUSC* lung squamous cell carcinoma, *SCLC* small cell lung cancer, *SNP* single-nucleotide polymorphism

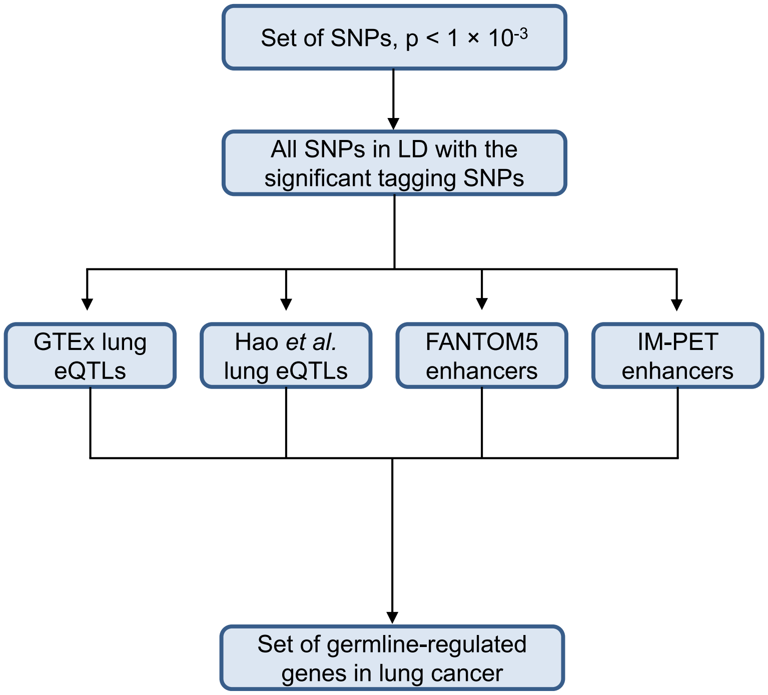
Set of SNPs, p < 1 × 10-3
All SNPs in LD with the significant tagging SNPs
GTEx lung eQTLs
Hao *et al.* lung eQTLs
FANTOM5 enhancers
IM-PET enhancers
Set of germline-regulated genes in lung cancer

Fig. 1 Pipeline to identify germline-regulated genes from the SNPs of the genome-wide association study. The start of our pipeline utilizes moderately associated SNPs ($p < 1 \times 10^{-3}$) from the genome-wide association study for three lung cancer subtypes. The outcome is a set of germline-regulated genes in each lung cancer subtype. eQTL expression quantitative trait loci, GTEx Genotype-Tissue Expression project, LD linkage disequilibrium, SNP single-nucleotide polymorphism

and multi-tissue eQTLs represented by the SNPs to form the final set of GTEx eQTLs. Many of these eQTL SNPs were within strong LD of each other and controlled the expression of the same target gene, so we collapsed all eQTLs to the specific genes they control. As illustrated in Fig. 2b, we found a total of 71 genes for LUAD, 108 for LUSC, and 67 for SCLC. Three genes overlapped from one unique locus in all three subtypes (*CHRNA5, PSMA4,* and *RP11-650 L12.2*). *CHRNA5* is in the nicotinic acetylcholine region that has well-known associations with lung cancer [19, 24, 25] and smoking [53, 54], while *PSMA4* is also associated with lung cancer [55, 56].

We examined a third set of lung tissue eQTLs generated from a meta-analysis that used lung tissue samples from three different recruitment sites (not including GTEx data) [42]. We refer to this eQTL dataset as the Hao et al. eQTLs. We found 25 SNPs for LUAD, 34 for LUSC, and 16 for SCLC that acted as eQTLs (Fig. 2c). We reduced the number of eQTLs to unique target genes (see "Methods") and found no genes that overlapped all three subtypes, no genes that overlapped LUAD and SCLC, two genes that overlapped LUSC and SCLC in one genomic region (*MYL4* and *RPRML*), and one gene (*IREB2*) that overlapped the two NSCLC subtypes (Fig. 2d). *IREB2* has been previously reported to be associated with both chronic obstructive pulmonary disease and lung cancer, and a recent study suggests a stronger association for lung cancer than chronic obstructive pulmonary disease [57].

Finding transcribed enhancers and their target genes

We next examined SNPs located within enhancer regions of the genome that had associated target genes. We used data from the Functional Annotation of the Mammalian Genome (FANTOM) collaborative project [58] that identified transcribed enhancer regions of the genome known as eRNAs using the Cap Analysis of Gene Expression (CAGE) technology [59]. We used this permissive set of enhancers and their corresponding transcribed target genes from the Promoter Enhancer Slider Selector Tool (PrESSTo) website [58, 60]. We found that the number of genes that were targeted by the enhancers was 45 for LUAD, 104 for LUSC, and 43 for SCLC (Fig. 3a). We removed duplicated genes in each subtype and found no overlap for these enhancer target genes among all three subtypes (Fig. 3b). We also observed no overlap among LUAD and SCLC or SCLC and LUSC. However, we did find five target genes from two genomic loci that overlapped LUAD and LUSC (*EPB49, LGI3, LPCAT1, NPM2,* and *PHYHIP*).

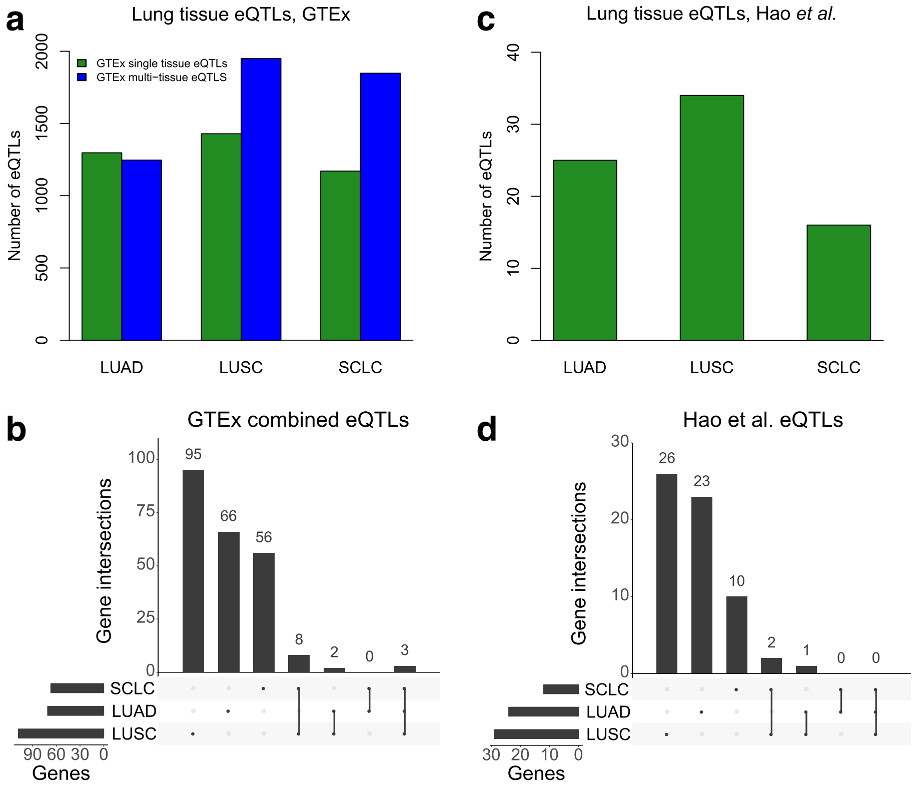

Fig. 2 Lung tissue eQTLs in three lung cancer subtypes. **a** Total number of significant eQTLs found in each lung cancer subtype using lung-tissue-specific data (*q* value ≤5%) and multi-tissue data (posterior probability >0.8) from GTEx. **b** UpSetR plot shows the overlap of genes discovered from the GTEx eQTLs. For each lung cancer subtype, we obtained the final gene set by collapsing all SNPs from (**a**) into genes. **c** Total number of eQTLs (false discovery rate < 10%) found in the lung-tissue-specific dataset from Hao et al. [42]. **d** UpSetR plot shows the overlap of genes based on Hao et al. eQTLs. Duplicate genes were removed from **c** for this comparison. eQTL expression quantitative trait loci, GTEx Genotype-Tissue Expression project, LUAD lung adenocarcinoma, LUSC lung squamous cell carcinoma, SCLC small cell lung cancer

Finding epigenetically defined enhancers and their predicted target genes

To find SNPs located within epigenetically defined enhancers, we used a dataset that defined enhancers using histone modifications such as H3K4me1 [61] and H3K27ac [62]. Specifically, we used the results from a newly developed software tool, IM-PET, that uses specific histone marks to identify enhancers and other data types to predict their targets using a sophisticated random forest classifier [43]. We found more than 100 enhancer targets in all subtypes across two lung-related cell lines (IMR90 and NHLF) (Fig. 3c). These enhancer targets are mRNA transcripts. Therefore, for a comparison similar to that used for the previous datasets, we collapsed all transcripts into single genes (see "Methods"). We merged genes found across both cell lines and removed duplicated genes within subtypes. Only two genes from one unique locus overlapped all subtypes (*ID3HA and TBC1D2B)* (Fig. 3d). *IDH3A* encodes an enzyme in the metabolic tricarboxylic acid (TCA) cycle that is frequently altered in cancer cells [63].

Final set of germline-regulated genes and comparison to original study

We collected all the genes identified by all of the above methods, removed duplicated genes within subtypes, and referred to this final collection of genes as germline-regulated genes (Additional file 1: Tables S2–S4). Only five genes were shared by all of the subtypes: *CHRNA5*, *IDH3A*, *PSMA4, RP11-650 L12.2,* and *TBC1D2B* (Fig. 4a). Although we found five unique genes, these genes are all located together in one unique genomic region on 15q25 and probably represent only one unique signal. We also compared the genes found across all of the different methods per subtype. Interestingly, we found very little overlap in the target genes identified between the different methods in each subtype. This trend is consistent across all three subtypes (Fig. 5).

A common approach used to report genes that may be associated with SNPs found from a GWAS is to report genes that are the closest in proximity upstream or downstream of the genotyped SNP. Therefore, we next verified that our approach to determine target genes

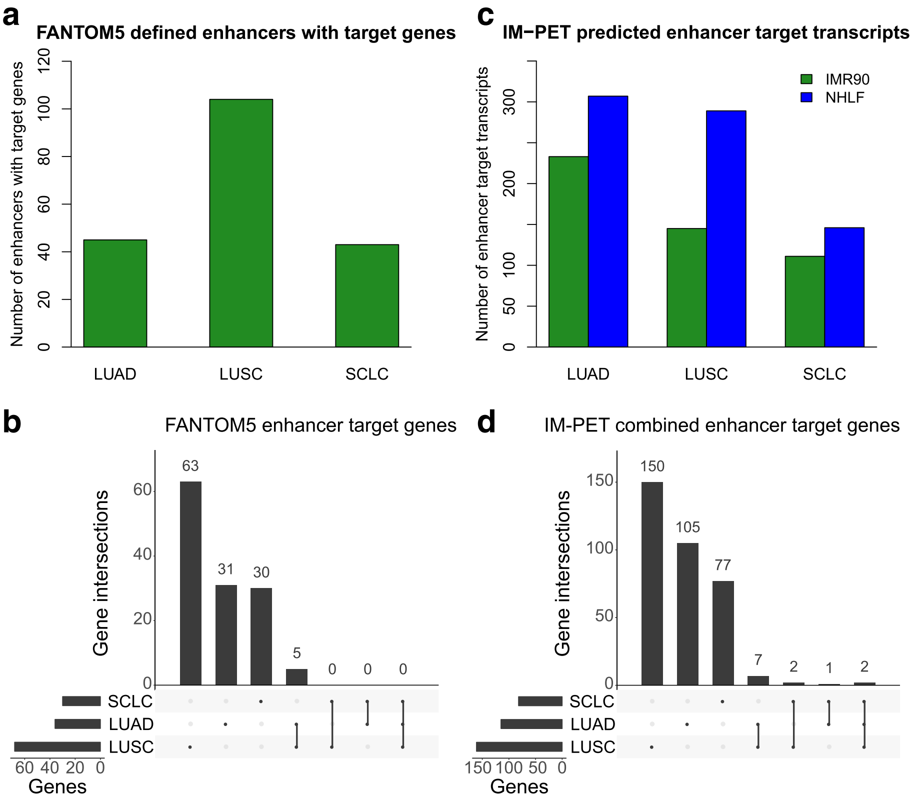

Fig. 3 Comparison of the SNPs located within the enhancer regions and their target genes. **a** Total number of enhancer target genes identified by FANTOM5. **b** UpSetR plot shows the overlap of FANTOM5 enhancer target genes by subtype. **c** Total number of enhancer target transcripts identified by IM-PET for two lung-related cell lines. **d** UpSetR plot shows the overlap of the lung cancer predicted enhancer target genes for IMR90 and NHLF identified by IM-PET. LUAD lung adenocarcinoma, LUSC lung squamous cell carcinoma, SCLC small cell lung cancer, SNP single-nucleotide polymorphism

from GWAS SNPs identified a different set of genes than the genes originally reported using the closest gene approach in the original study [20]. For this comparison, we ran the same analysis described above and used the same set of SNPs reported in the original paper's supplemental tables. We found that only ~25% of the germline-regulated genes that we found using our approach were reported in the original GWAS publication (Additional file 1: Figure S4A).

We further applied our approach to analyze the data from the GWAS Catalog and obtained a set of SNPs for a matched European population type from the GWAS Catalog [64] using the search term "lung cancer" (see "Methods"). After removing the SNPs from our original study, we identified 17 SNPs to run through our pipeline. Additional file 1: Table S5 shows the results from this analysis. We ran the SNPs through the pipeline and identified six germline-regulated genes from the GWAS Catalog SNPs: *CHRNA5*, *CLPTM1L*, *PSMA4*, *RP11-650 L12.2*, *TP63*, and *ZSCAN29*. We examined the overlap between these genes and our germline-regulated genes by lung cancer subtype in the above subsections. There was a strong overlap (67%) between the genes in at least one subtype from our analysis and the target genes associated with lung cancer from the GWAS Catalog (Additional file 1: Figure S4B).

Pathway enrichment analysis of germline-regulated genes

To gain a deeper understanding into the biology driven by these germline-regulated genes, we performed biological pathway enrichment analysis of the genes in each subtype. We used the web-based tool, WEB-based Gene Set Analysis Toolkit (WebGestalt) [47], to identify significantly enriched pathways with the set of germline-regulated genes for each subtype using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. The pathways enriched in each subtype are listed in Additional file 1: Tables S6–S8. We found that all three subtypes had genes enriched in the metabolic pathways and proteasome pathways (Fig. 4c). We note that many of the pathways found for LUSC represent only one genomic locus (HLA region, chromosome 6p21), which contains the same sets of genes (Additional file 1: Table S7).

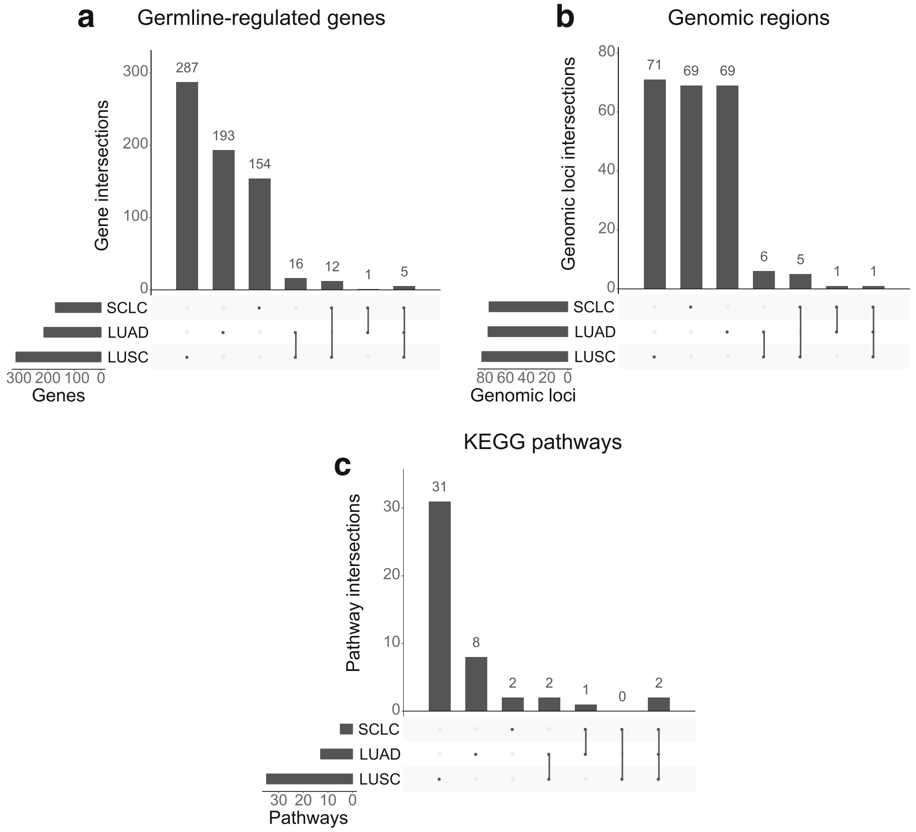

Fig. 4 Comparison of the germline-regulated genes, independent genomic loci, and enriched biological pathways by subtype. **a** UpSetR plot shows the overlap of germline-regulated genes identified in the present study for the three lung cancer subtypes. **b** UpSetR plot shows the overlap of independent genomic loci that represent the genes shown in (**a**). **c** UpSetR plot shows the overlap of pathways from the Kyoto Encyclopedia of Genes and Genomes enriched with the germline-regulated genes. KEGG Kyoto Encyclopedia of Genes and Genomes, LUAD lung adenocarcinoma, LUSC lung squamous cell carcinoma, SCLC small cell lung cancer

Identification of overlapping genes with independent signals

Our approach to identify germline-regulated genes included LD expansion of the preliminary set of GWAS SNPs. Therefore, there were genes that were identified in the same regions of the genome that potentially share the same genetic signal. To determine the number of independent signals we identified in our analysis, we clustered the genes within 1 Mb of each other on each chromosome into a single unique signal. We performed this clustering of genes for all the comparisons that were done at the gene level. We have listed these results for each data source as independent genomic loci in Additional file 1: Tables S9–S15. These results supported the same conclusion from our main analyses.

Principal component analysis of TCGA germline genotyped SNPs

We also expanded our analysis beyond the GWAS SNPs to determine the degree of genetic sharing using data from TCGA. Importantly, TCGA did not study SCLC, so we were limited to data generated for LUAD and LUSC. We obtained all germline genotyping data from normal blood samples for six cancer types (LUAD, LUSC, head and neck squamous cell carcinoma, bladder urothelial carcinoma, glioblastoma multiforme, and brain lower grade glioma) from TCGA's data portal (https://portal.gdc.cancer.gov/legacy-archive/search/f). To avoid any genetic influence that may occur due to population differences, we limited our samples to TCGA's defined white population for our analysis. We ran a PCA on these six cancer types to identify the degree of similarity at the germline genetic level (Additional file 1: Figure S5). Our results indicated that the samples for LUAD and LUSC are no closer to each other spatially than other cancer types. Additionally, several of the samples for LUAD and LUSC are in locations of the plot with other cancer types. These results agree with our regulatory analysis and suggest that LUAD and LUSC do not share much in common with each other at the germline genetic level.

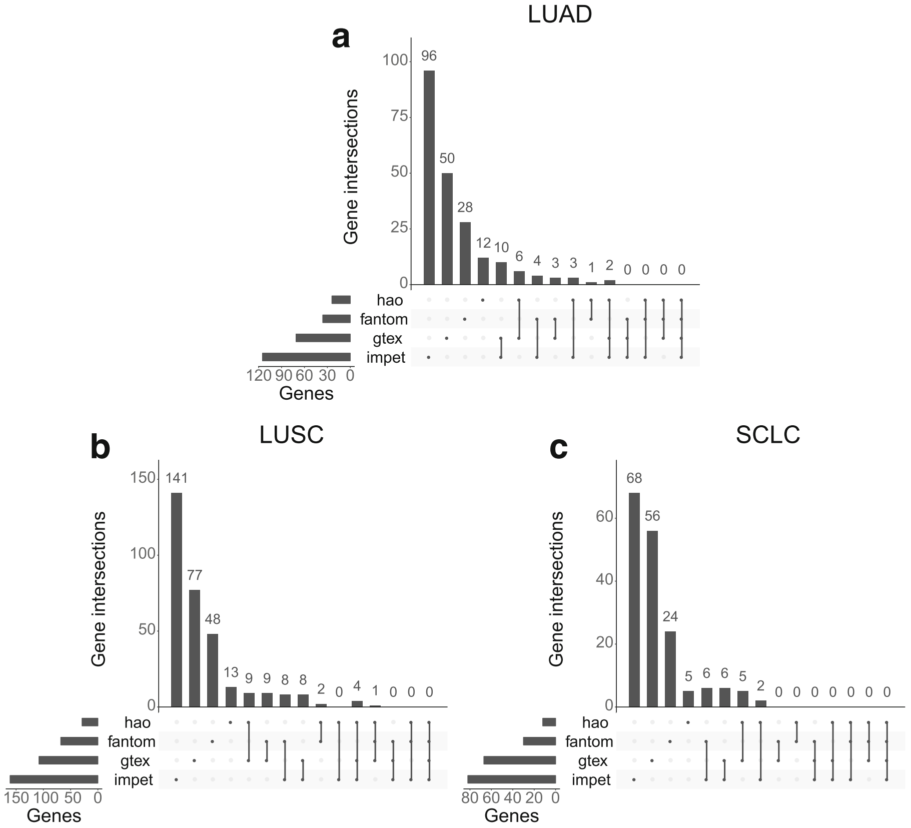

Fig. 5 Comparison of the final germline-regulated genes in each subtype separated by the different data sources. UpSetR plots show the overlap between genes found from each data source for **a** lung adenocarcinoma (LUAD), **b** lung squamous cell carcinoma (LUSC), and **c** small cell lung cancer (SCLC). GTEx Genotype-Tissue Expression project, LUAD lung adenocarcinoma, LUSC lung squamous cell carcinoma, SCLC small cell lung cancer

Discussion

Understanding the genetic risk factors for any cancer type is important in uncovering the underlying biology of the disease. For example, if a SNP is unique to one subtype and acts as an eQTL for a gene involved in a cancer-related pathway, somatic alterations in that gene or other genes in the same pathway can be investigated to understand the development of that subtype. It is also possible that somatic mutations can act in concert with expression-altering SNPs in driving the tumor and would not have the same effect on growth advantage in the absence of the SNP.

Additionally, a specific understanding of the regulatory roles that common genetic variants play in the development of lung cancer subtypes is an important research question because the majority of common variants that increase cancer risk are located within non-coding regions and most likely act as regulators of gene expression. An improved understanding of the carcinogenesis process may provide indications for biomarkers for risk prediction and therapeutic strategies. This is particularly important for SCLC, which is typically diagnosed at a late stage and for which there are not many therapeutic options. To address these questions, we performed a detailed analysis of common genetic variants (SNPs) associated with three subtypes of lung cancer (LUAD, LUSC, and SCLC).

We used marginally significant GWAS results ($p < 1 \times 10^{-3}$) to search for regulatory roles for common variants associated with LUAD, LUSC, and SCLC. We expanded this set of results to include all SNPs in LD with the genotyped SNPs using data from the 1000 Genomes Phase III project. This expansion resulted in ~15,000 more SNPs to test per subtype that may be acting as the actual causal variant [65]. We used a diverse set of regulatory data to identify SNPs that were within regulatory regions of the genome that had an identified target gene. Overall, we found a very small overlap between all three subtypes at the SNP, gene, pathway, and regulatory levels. Of note, we found a similar lack of overlap between all subtypes on all levels when we used SNPs with $p < 1 \times 10^{-4}$.

It is worth highlighting that three (*CHRNA5, IDH3A,* and *PSMA4)* out of the five genes shared in all three subtypes of lung cancer have been previously reported to be associated with lung cancer. *CHRNA5* has strong implications in its association with lung cancer [19, 24, 25].

CHRNA5 encodes a nicotinic acetylcholine receptor (nAChR). nAChRs are a class of ligand-gated ion channels that are activated by the neurotransmitter acetylcholine to allow the flow of ions across a cell membrane [66]. There is still an ongoing debate about *CHRNA5*'s role in lung cancer risk versus its risk for lung cancer through nicotine addiction [67], but finding this gene in all three subtypes of lung cancer, which have biological and environmental differences, suggests it may play a direct role in lung cancer risk. *IDH3A* encodes an isocitrate dehydrogenase (IDH). IDHs are important enzymes in the regulation of the TCA cycle [68]. Additionally, *IDH3A* promotes tumor growth by activating hypoxia-inducible factor 1 (HIF-1) and promotes the stability of HIF-1 in participating in angiogenesis and is also associated with poor survival in lung cancer [69]. *IDH3A* also acts in the conversion of metabolism that occurs with cancer fibroblasts [70]. *PSMA4* encodes a subunit of the proteasome. Experimental studies have shown that *PSMA4* mRNA is increased in lung tumor versus normal samples and plays major roles in cell proliferation using data from lung carcinoma cell lines [71]. Another gene, *RP11-650 L12.2*, is a non-coding antisense RNA that has not been well characterized. However, one recent study by Jin et al. [72] found a variant in the promoter region of *RP11-650 L12.2* that is associated with risk of colorectal cancer. This finding, in addition to its association with all three subtypes of lung cancer, warrants future experimental studies of this gene. The final gene shared by all subtypes, *TBC1D2B*, is a protein-coding gene that may have GTPase activity and may have a role in autophagy [73].

In addition to the five overlapping genes above, our pathway enrichment analysis revealed two biological pathways shared in the three subtypes. Among them, all three subtypes shared metabolic pathways and the proteasome pathways. Metabolic pathways are frequently modified in cancer to provide the over-proliferating cells with the required nutrients [74, 75]. The proteasome pathway has several links to cell growth in several cancer types [76]. Although these two pathways have a strong relevance to cancer, they are also associated with other disease types due to their components acting in many biological processes. We also observed that the oxidative phosphorylation pathway was significantly enriched in LUSC (Benjamini–Hochberg adjusted p = 0.0317). It is interesting to find this pathway dysregulated in the germline genome, because it has strong associations in the transition from oxidative phosphorylation to the less efficient aerobic glycolysis, known as the Warburg effect, which occurs in cancer cell proliferation [77]. Although the Warburg effect may be attributable to glycolysis inhibiting a still active oxidative phosphorylation pathway, this result suggests that commonly occurring variants in LUSC may lead to some disruption in the oxidative phosphorylation pathway that makes this process easier to arrest or inhibit and enhance cell proliferation after some somatic disruption in somatic lung tissue. We also found several cancer-related pathways in LUSC such as pathways in cancer and prostate cancer, and many signaling pathways associated with cancer. We discovered that the focal adhesion pathway was significantly enriched with genes from SCLC (Benjamini–Hochberg adjusted p = 0.0275). This is an intriguing finding because this process is involved in the epithelial–mesenchymal transition, which is important in cancer metastasis [78, 79]. In summary, this pathway-based evidence suggests both shared subtype and unique subtype associations.

There are several limitations to this study. First, we utilized a set of marginally significant SNPs. Although previous studies [80, 81] have shown that this is a practical approach, this may have resulted in some false positive SNPs in our study. Second, we did not impute the GWAS data to obtain a larger set of SNPs for the analysis. This would have resulted in more SNPs that could have been tested for significance. We will integrate such SNPs in future analyses. Third, we used the q value cutoff identified by GTEx for eQTL significance or non-significance. However, it is possible that there are subtle changes to gene expression from SNPs in the genome and therefore, we may be unintentionally adding or removing SNPs that subtly act in this manner by using a strict predefined cutoff value. Fourth, to validate our results, we were limited to a small set of SNPs reported in the GWAS Catalog because we focused only on SNPs specifically found in one population. While we observed a strong overlap (67%), it would have been better to include a larger set of SNPs for better power of confirming our pipeline. Another limitation of our study is that we may have discovered several different genes that may represent only one unique signal because we used SNPs in LD for our analysis. For example, if we found five genes that were shared by all subtypes, but these genes were clustered in one genomic location, these fives genes may represent only a single unique signal. To account for this potential bias, we separated the genes into unique signals to give a better understanding of the overlap of the subtypes while still including all discovered germline-regulated genes.

Conclusions

In summary, we used common genetic variants found in three lung cancer subtypes to interrogate the similarity between them at four biological levels (SNP, gene, regulatory, and pathway levels). We found very little overlap between the three subtypes at these levels. At the most basic level (SNPs), we observed less than 1% overlap between the subtypes. Similarly, we found only five genes (from one independent genomic locus) that overlap in all three subtypes, representing <1% of the genes we examined. Three of these five genes (*CHRNA5, IDH3A,* and *PSMA4*) are well-known lung cancer genes. We observed the same trend at the pathway level and found only two KEGG pathways overlapped the subtypes. At the regulatory level, we discovered that many of the enhancer target genes and eQTL target genes are unique to each subtype. Not much work has been done comparing all three subtypes at the somatic level, but recent work interrogating the differences between LUAD and LUSC concluded similarly that there was little overlap between these two subtypes at the molecular level in somatic tumor tissue [18]. Overall, this study provides some important insight into the genetic architecture of three subtypes of lung cancer.

Additional file

Additional file 1: Table S1. Randomization results for overlapping SNPs. **Table S2.** Final germline-regulated genes for LUAD. **Table S3.** Final germline-regulated genes for LUSC. **Table S4.** Final germline-regulated genes for SCLC. **Table S5.** Final germline-regulated genes for GWAS Catalog SNPs. **Table S6.** Pathway enrichment results for LUAD. **Table S7.** Pathway enrichment results for LUSC. **Table S8.** Pathway enrichment results for SCLC. **Table S9.** Independent locus level analysis for genes uniquely identified in LUAD. **Table S10.** Independent locus-level analysis for genes uniquely identified in LUSC. **Table S11.** Independent locus-level analysis for genes uniquely identified in SCLC. **Table S12.** Independent locus-level analysis for LUAD overlap with LUSC. **Table S13.** Independent locus-level analysis for LUAD overlap with SCLC. **Table S14.** Independent locus-level analysis for LUSC overlap with SCLC. **Table S15.** Independent locus-level analysis for all overlaps. **Figure S1.** Determination of significance for GTEx multi-tissue eQTLs. **Figure S2.** Comparison of SNPs from the GWAS for lung cancer. **Figure S3.** Pipelines used to obtain overlap in the LD expanded and LD trimmed SNPs per lung cancer subtype. **Figure S4.** Comparison of germline-regulated genes to original report and the GWAS Catalog. **Figure S5.** PCA of germline genotype data from TCGA in six cancer types. (DOCX 628 kb)

Abbreviations

eQTL: Expression quantitative trait loci; FANTOM: Functional Annotation of the Mammalian Genome; GTEx: Genotype-Tissue Expression project; GWAS: Genome-wide association study; IM-PET: Integrated Methods for Predicting Enhancer Targets; KEGG: Kyoto Encyclopedia of Genes and Genomes; LD: Linkage disequilibrium; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; NCBI: National Center for Biotechnology Information; NSCLC: Non-small cell lung cancer; PCA: Principal component analysis; SCLC: Small cell lung cancer; SNP: Single-nucleotide polymorphism; TCGA: The Cancer Genome Atlas

Acknowledgements

We thank the two reviewers for their constructive comments that helped improve the quality of the manuscript. We thank Drs. Nancy Cox, David Samuels, Melinda Aldrich, Tony Capra, and Jirong Long for their valuable suggestions on this project, and members of the Zhao lab for helpful comments on this work.
We also thank Dr. Barbara O'Brien for comments on an earlier version of the manuscript and Dr. Pora Kim for help with the website development.

Funding

This work was partially supported by U.S. National Institutes of Health grants (R01LM012806 and R01LM011177) to ZZ. TDO was supported by a National Institutes of General Medical Sciences training grant (T32GM080178). The GWAS data used in this study was originally generated with support from the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, to MTL. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

We used summary results from an initial GWAS that was conducted for LUAD, LUSC, and SCLC [20]. The computer code files that were used for this study can be obtained at the website https://bioinfo.uth.edu/LungCancerSubtypes/.
The code we used for this study is also stored on GitHub at https://github.com/timothyob/LungCancerSubtypes.

Authors' contributions

TDO, PJ, and ZZ contributed to the conception and design of the study. TDO and PJ performed the formal analysis of the data. The GWAS data was generated by MTL and NEC. TDO, PJ, MTL, and ZZ wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN, USA. [2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [3]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St. Suite 820, Houston, TX 77030, USA. [4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [5]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA.

Received: 12 November 2017 Accepted: 13 February 2018
Published online: 27 February 2018

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015; 65:5–29.
2. Alberg AJ, Wallace K, Silvestri GA, Brock MV. Invited commentary: the etiology of lung cancer in men compared with women. Am J Epidemiol. 2013;177:613–6.
3. Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. Int J Cancer. 2002;99:260–6.
4. Hemminki K, Lonnstedt I, Vaittinen P, Lichtenstein P. Estimation of genetic and environmental components in colorectal and lung cancer and melanoma. Genet Epidemiol. 2001;20:107–16.
5. Pao W, Girard N. New driver mutations in non-small-cell lung cancer. Lancet Oncol. 2011;12:175–80.
6. Houston KA, Henley SJ, Li J, White MC, Richards TB. Patterns in lung cancer incidence rates and trends by histologic type in the United States, 2004–2009. Lung Cancer. 2014;86:22–8.
7. Lemjabbar-Alaoui H, Hassan OU, Yang YW, Buchanan P. Lung cancer: biology and treatment options. Biochim Biophys Acta. 2015;1856:189–210.
8. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455:1069–75.
9. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012;150:1107–20.
10. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.

11. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, Maher CA, Fulton R, Fulton L, Wallis J, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell. 2012;150:1121–34.
12. Hammerman PS, Sos ML, Ramos AH, Xu C, Dutt A, Zhou W, Brace LE, Woods BA, Lin W, Zhang J, et al. Mutations in the DDR2 kinase gene identify a novel therapeutic target in squamous cell lung cancer. Cancer Discov. 2011;1:78–89.
13. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.
14. Weiss J, Sos ML, Seidel D, Peifer M, Zander T, Heuckmann JM, Ullrich RT, Menon R, Maier S, Soltermann A, et al. Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. Sci Transl Med. 2010;2:62ra93.
15. Peifer M, Fernandez-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, Plenker D, Leenders F, Sun R, Zander T, et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet. 2012;44:1104–10.
16. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, Bergbower EA, Guan Y, Shin J, Guillory J, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet. 2012;44:1111–6.
17. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158:929–44.
18. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, Shukla SA, Guo G, Brooks AN, Murray BA, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet. 2016;48:607–16.
19. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008;452:633–7.
20. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet. 2009;85:679–91.
21. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, Zong X, Laplana M, Wei Y, Han Y, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. 2014;46:736–41.
22. Yoon KA, Park JH, Han J, Park S, Lee GK, Han JY, Zo JI, Kim J, Lee JE, Takahashi A, et al. A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population. Hum Mol Genet. 2010;19:4948–54.
23. Broderick P, Wang Y, Vijayakrishnan J, Matakidou A, Spitz MR, Eisen T, Amos CI, Houlston RS. Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. Cancer Res. 2009;69:6633–41.
24. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. 2008;40:616–22.
25. Liu P, Vikis HG, Wang D, Lu Y, Wang Y, Schwartz AG, Pinney SM, Yang P, de Andrade M, Petersen GM, et al. Familial aggregation of common sequence variants on 15q24-25.1 in lung cancer. J Natl Cancer Inst. 2008;100:1326–30.
26. Timofeeva MN, Hung RJ, Rafnar T, Christiani DC, Field JK, Bickeboller H, Risch A, McKay JD, Wang Y, Dai J, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14,900 cases and 29,485 controls. Hum Mol Genet. 2012;21:4980–95.
27. Zanetti KA, Wang Z, Aldrich M, Amos CI, Blot WJ, Bowman ED, Burdette L, Cai Q, Caporaso N, Chung CC, et al. Genome-wide association study confirms lung cancer susceptibility loci on chromosomes 5p15 and 15q25 in an African–American population. Lung Cancer. 2016;98:33–42.
28. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, Qureshi M, Dong Q, Gu X, Chen WV, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat Genet. 2008;40:1407–9.
29. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, et al. Lung cancer susceptibility locus at 5p15.33. Nat Genet. 2008;40:1404–6.
30. Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, Yang L, Dai J, Hu L, Tan W, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. Nat Genet. 2011;43:792–6.
31. Shi J, Chatterjee N, Rotunno M, Wang Y, Pesatori AC, Consonni D, Li P, Wheeler W, Broderick P, Henrion M, et al. Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma. Cancer Discovery. 2012;2:131.
32. Zuber V, Marconett CN, Shi J, Hua X, Wheeler W, Yang C, Song L, Dale AM, Laplana M, Risch A, et al. Pleiotropic analysis of lung cancer and blood triglycerides. J National Cancer Inst. 2016;108:djw167.
33. Jiang J, Jia P, Shen B, Zhao Z. Top associated SNPs in prostate cancer are significantly enriched in cis-expression quantitative trait loci and at transcription factor binding sites. Oncotarget. 2014;5:6168–77.
34. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ, Trait-associated SNP. are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6:e1000888.
35. Chelala C, Khan A, Lemoine NR. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics. 2009;25:655–61.
36. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27:718–9.
37. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
39. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353–8.
40. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet. 2013;9:e1003486.
41. Gen Li, Shabalin AA, Rusyn I, Fred A. Wright, AB. Nobel; An empirical Bayes approach for multiple tissue eQTL analysis. Biostatistics. https://doi.org/10.1093/biostatistics/kxx048.
42. Hao K, Bosse Y, Nickle DC, Pare PD, Postma DS, Laviolette M, Sandford A, Hackett TL, Daley D, Hogg JC, et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet. 2012;8:e1003029.
43. He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci. 2014;111:E2191–9.
44. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
45. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. Ensembl 2014. Nucleic Acids Res. 2014;42:D749–55.
46. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21:3439–40.
47. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. 2005;33:W741–8.
48. Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.
49. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: visualization of intersecting sets. IEEE Trans Vis Comput Graph. 2014;20:1983–92.
50. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, Caporaso NE, Johansson M, Xiao X, Li Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017;49:1126–32.
51. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.
52. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648–60.
53. Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau O, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum Mol Genet. 2007;16:36–49.
54. Wen L, Jiang K, Yuan W, Cui W, Li MD. Contribution of variants in CHRNA5/A3/B4 gene cluster on chromosome 15 to tobacco smoking: from genetic association to mechanism. Mol Neurobiol. 2016;53:472–84.
55. Wang T, Chen T, Thakur A, Liang Y, Gao L, Zhang S, Tian Y, Jin T, Liu JJ, Chen M. Association of PSMA4 polymorphisms with lung cancer susceptibility and response to cisplatin-based chemotherapy in a Chinese Han population. Clin Transl Oncol. 2015;17:564–9.
56. Lee MP, Reeves C, Schmitt A, Su K, Connors TD, Hu RJ, Brandenburg S, Lee MJ, Miller G, Feinberg AP. Somatic mutation of TSSC5, a novel imprinted gene from human chromosome 11p15.5. Cancer Res. 1998;58:4155–9.

57. Ziolkowska-Suchanek I, Mosor M, Gabryel P, Grabicki M, Zurawek M, Fichna M, Strauss E, Batura-Gabryel H, Dyszkiewicz W, Nowak J. Susceptibility loci in lung cancer and COPD: association of IREB2 and FAM13A with pulmonary diseases. Sci Rep. 2015;5:13502.
58. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.
59. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, et al. CAGE: cap analysis of gene expression. Nat Methods. 2006;3:211–22.
60. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.
61. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39:311–8.
62. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci. 2010;107:21931–6.
63. Levine AJ, Puzio-Kuter AM. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. Science. 2010;330:1340–4.
64. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–6.
65. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012;8:e1002822.
66. Taly A, Corringer PJ, Guedin D, Lestage P, Changeux JP. Nicotinic receptors: allosteric transitions and therapeutic targets in the nervous system. Nat Rev Drug Discov. 2009;8:733–50.
67. Improgo MR, Scofield MD, Tapper AR, Gardner PD. From smoking to lung cancer: the CHRNA5/A3/B4 connection. Oncogene. 2010;29:4874–84.
68. Huh TL, Kim YO, Oh IU, Song BJ, Inazawa J. Assignment of the human mitochondrial NAD+-specific isocitrate dehydrogenase α subunit (IDH3A) gene to 15q25.1 → q25.2 by in situ hybridization. Genomics. 1996;32:295–6.
69. Zeng L, Morinibu A, Kobayashi M, Zhu Y, Wang X, Goto Y, Yeom CJ, Zhao T, Hirota K, Shinomiya K, et al. Aberrant IDH3α expression promotes malignant tumor growth by inducing HIF-1-mediated metabolic reprogramming and angiogenesis. Oncogene. 2015;34:4758–66.
70. Zhang D, Wang Y, Shi Z, Liu J, Sun P, Hou X, Zhang J, Zhao S, Zhou BP, Mi J. Metabolic reprogramming of cancer-associated fibroblasts by IDH3α downregulation. Cell Rep. 2015;10:1335–48.
71. Liu Y, Liu P, Wen W, James MA, Wang Y, Bailey-Wilson JE, Amos CI, Pinney SM, Yang P, de Andrade M, et al. Haplotype and cell proliferation analyses of candidate lung cancer susceptibility genes on chromosome 15q24-25.1. Cancer Res. 2009;69:7844–50.
72. Jin M, Ye D, Li Y, Jing F, Jiang X, Gu S, Mao Y, Li Q, Chen K. Association of a novel genetic variant in RP11-650L12.2 with risk of colorectal cancer in Han Chinese population. Gene. 2017;624:21–5.
73. Popovic D, Akutsu M, Novak I, Harper JW, Behrends C, Dikic I. Rab GTPase-activating proteins in autophagy: regulation of endocytic and autophagy pathways by direct binding to human ATG8 modifiers. Mol Cell Biol. 2012; 32:1733–44.
74. Dang CV. Links between metabolism and cancer. Genes Dev. 2012;26:877–90.
75. Kim P, Cheng F, Zhao J, Zhao Z. ccmGDB: a database for cancer cell metabolism genes. Nucleic Acids Res. 2016;44:D959–68.
76. Mani A, Gelmann EP. The ubiquitin-proteasome pathway and its role in cancer. J Clin Oncol. 2005;23:4776–89.
77. Zheng J. Energy metabolism of cancer: glycolysis versus oxidative phosphorylation. Oncol Lett. 2012;4:1151–7.
78. Kalluri R, Weinberg RA. The basics of epithelial–mesenchymal transition. J Clin Invest. 2009;119:1420–8.
79. Thiery JP. Epithelial–mesenchymal transitions in tumour progression. Nat Rev Cancer. 2002;2:442–54.
80. Wang Q, Yu H, Zhao Z, Jia P. EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. Bioinformatics. 2015;31:2591–4.
81. Jia P, Wang L, Meltzer HY, Zhao Z. Pathway-based analysis of GWAS datasets: effective but caution required. Int J Neuropsychopharmacol. 2011;14:567–72.