Am J Cancer Res 2014;4(4):394-410
www.ajcr.us /ISSN:2156-6976/ajcr0000975

Original Article

Genome-wide prediction of cancer driver genes based on SNP and cancer SNV data

Quanze He[1*], Quanyuan He[2*], Xiaohui Liu[3,4], Youheng Wei[1], Suqin Shen[1], Xiaohui Hu[1], Qiao Li[5], Xiangwen Peng[5], Lin Wang[6], Long Yu[1]

[1]The State Key Laboratory of Genetic Engineering, Institute of Biomedical Science, Fudan University, 220 Handan Rd, Shanghai 200433, China; [2]Verna and Marrs Mclean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA; [3]Department of Chemistry, Fudan University, Shanghai 200032, China; [4]Institute of Biomedical Sciences, Fudan University, Shanghai 200032, China; [5]The State Key Laboratory of Genetic Engineering, Department of Genetics, Fudan University, 220 Handan Rd, Shanghai 200433, China; [6]Key Laboratory of Crop Genetics and Physiology of Jiangsu Province, College of Bioscience and Biotechnology, Yangzhou University, Yangzhou 225009, China. [*]Equal contributors.

Received June 2, 2014; Accepted June 17, 2014; Epub July 16, 2014; Published July 30, 2014

Abstract: Identifying cancer driver genes and exploring their functions are essential and the most urgent need in basic cancer research. Developing efficient methods to differentiate between driver and passenger somatic mutations revealed from large-scale cancer genome sequencing data is critical to cancer driver gene discovery. Here, we compared distinct features of SNP with SNV data in detail and found that the weighted ratio of SNV to SNP (termed as WVPR) is an excellent indicator for cancer driver genes. The power of WVPR was validated by accurate predictions of known drivers. We ranked most of human genes by WVPR and did functional analyses on the list. The results demonstrate that driver genes are usually highly enriched in chromatin organization related genes/pathways. And some protein complexes, such as histone acetyltransferase, histone methyltransferase, telomerase, centrosome, sin3 and U12-type spliceosomal complexes, are hot spots of driver mutations. Furthermore, this study identified many new potential driver genes (e.g. NTRK3 and ZIC4) and pathways including oxidative phosphorylation pathway, which were not deemed by previous methods. Taken together, our study not only developed a method to identify cancer driver genes/pathways but also provided new insights into molecular mechanisms of cancer development.

Keywords: Bioinformatics, SNV, SNP, mutation frequency, cancer driver gene

Introduction

Cancer is characterized by accumulated somatic mutations during tumorigenesis, of which only a small subset contributes to the tumor progression [1]. Distinguishing these "driver" mutations from the preponderance of "passenger" mutations is still a challenge because of the genetic heterogeneity of cancer [2]. In recent years, several methods were developed for predicting driver genes by taking advantage of the wealth of data produced by high-throughput cancer genome sequencing studies [3-5]. Till now more than 125 driver genes have been identified [1]. These genes have relatively high frequency mutations, which are usually shared by different tumors. However, more and more studies suggested that a great number of mutations with low frequency are shared by various cancers and remain discovered [6-9].

To identify driver genes, most of current methods test whether the mutation rate of each gene is significantly higher than the background (passenger) mutation rate using binomial or likelihood test. They used a common approach to define background non-silent mutation rate ρN, which is a product of ρS * R, where the ρS is the result of dividing the number of observed silent mutations by the number of base pairs and R is the average ratio of the number of potential non-silent mutation sites to the number of potential silent mutation sites. However this model is not as simple as it looks like. The most challenging part of this strategy is how to define pS and R for genes in distinct contexts.

Many elaborate models designed to deal with the problem using additional parameters such as mutation type, gene length and nucleotide context to optimize the model [3, 5]. Although working well for the genes with high frequency mutations, they have three unavoidable shortcomings for identifying low frequency ones. First, previous approaches ignore the fact that even the same type of mutations may have different impact on proteins' function when they occur at different sites (such as active site and no essential site); Second, in most cases, the observation number of silent mutations is too small to be used for estimating silent mutation ratio (ρN) accurately for each gene. For example, only 108 silent mutations were identified in the data from Ding et al. who sequenced 623 genes in 188 tumor samples. In average, even one sample has less than one silent mutation; Third, many non-silence mutations are actually "silence" and haven't significant impact on protein function, which results in overestimation of the value of R and losing of their sensitivity. As it is hard for current methods to overcome these shortcomings, developing new strategies to identify low frequency mutations is an urgent requirement in the field.

A single-nucleotide polymorphism (SNP) is an inheritable single nucleotide variation between members of species or paired chromosomes. Evolutionally, the SNP profile in genome is the fixating result of natural selection in evolution. Because unfavorable mutations will typically be eliminated while favorable changes are quickly fixed in a population, and only neutral (or nearly neutral) mutations, which has little effect on an organism's fitness, can be accumulated across the genome [10, 11]. They were considered as common variants in general population (minor allele frequency (MAF) of somatic mutation is > 1%) [10] and were derived from a lot of whole genome sequencing experiments. Therefore, the density of SNP can be used to estimate the frequency of neutral mutation for each gene [12, 13]. Single-nucleotide variants (SNV) are somatic point mutations found in cancer tissues. Majority of them are non-silent mutations locating at exons and lead to alterations of protein's structure/function. SNVs are enriched in cancer driver genes and cellular pathways essential for tumorigenesis. Here, we propose a new method that compares cancer SNV data to single-nucleotide polymorphism (common SNP) data to identify novel cancer driver genes. We validated the method by precise predictions of known drivers. Functional analyses on driver genes uncovered new protein complexes and pathways that are enriched with driver mutations, which provides new insights into molecular mechanisms of cancer development.

Materials and methods

Data collection

In this research, we collected four types of data: Chip-Seq data: Gene expression (RNA-Seq); mutation data (common SNPs [14] and cancer SNVs [15]) which were download from GEO [16], UCSC, SRA (NCBI Sequence Read Archive http://www.ncbi.nlm.nih.gov/Traces/sra), NIH website (http://dir.nhlbi.nih.gov) and COSMIC [15] databases. The detailed information about datasource can be found in Table S4. The reference genome [17] (NCBI37/hg19) of human was downloaded from UCSC FTP site.

Data pre-processing

Firstly, all ChiP-Seq and RNA-Seq data from SRA were converted to fastq files using SRA toolbox. Genomic read alignment and assemble were done by Bowtie [18] using default parameters with human reference genome of NCBI37/hg19. Secondly, the conversion of gene locations from NCBI36/hg18 to NCBI37/hg19 for sequence assembling was done by liftOver [19]. Finally, all aligned ChIP-Seq, RNA-Seq data files (bed, bam files) were converted into wig files using software MACS [20] with default setting. The cancer SNVs that co-localize with common SNPs were filtered out for consequential analyses.

Calculate weighted SNV/SNP ratios (WVPRs) for genes

Firstly, we used the longest isoform of a gene to define six gene related regions including exon, intron, promoter, tail, acceptor and donor. For each gene, promoter region includes upstream 1000bp and downstream 200bp of TSS; Donors include around 36bp of split site 5' [21]; Acceptors include upstream 36bp and downstream 24bp of split site 3' [22]; Tail includes upstream 200bp and downstream 1000bp of TTS. The location information of exons and introns were extracted from reference genomes

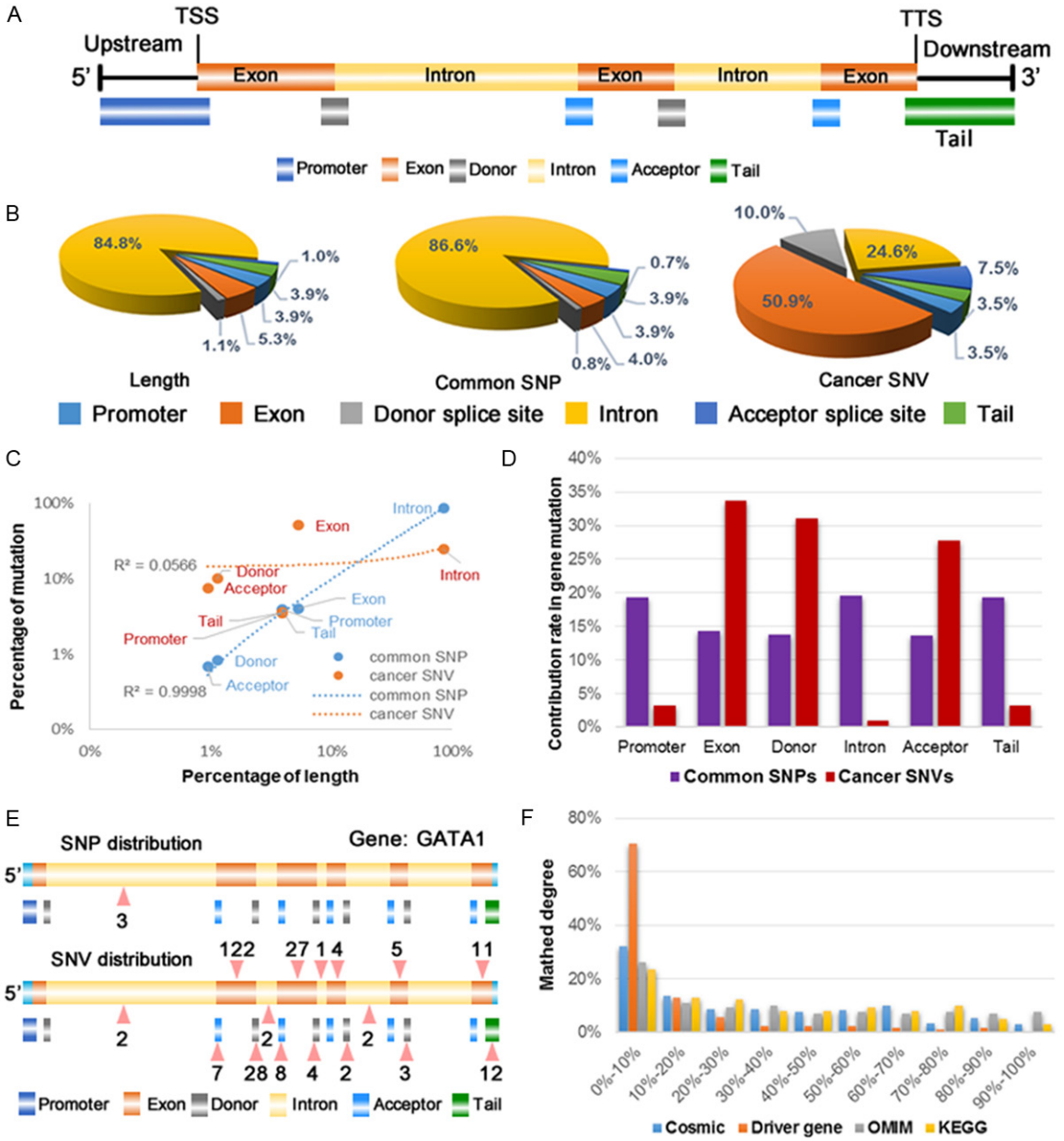


Figure 1. Predict cancer driver genes using SNP and SNV data. A. A carton to illustrate the definition of six regions in a gene. B. The distribution of length of six regions; the proportions of common SNPs and cancer SNVs within six different regions. C. The correlation between relative lengths and the numbers of SNVs and SNPs of six regions. D. The percentages of SNPs and SNVs in six regions. E. The distribution of SNPs and SNVs in GATA1. F. The distribution of cancer related genes annotated by databases in our ranked gene list. All genes were sorted and categorized into ten groups based on their WVPR value. The 0%-10% group contains genes with top10% highest WVPRs. There are 455 genes in Cosmic database [15], 180 genes in OMIM database and 168 gene in KEGG database (ver 2011-7-13) were annotated as cancer related genes. Driver gene list includes 125 genes and is adopted from the reference 1.

[17] (NCBI37/hg19) of human. All SNVs and SNPs were mapped to these regions. For each region, the weight (W) is the percentages of total number of SNPs or SNVs within this region to total numbers of SNPs and SNVs across genome. For each gene, we calculated the mutation densities for each region, which is the ratio of the number of mutations within the region to the length of the regions in the gene. Finally, a line model was used to speculate mutation risk in normal and cancer (MRN and MRC) cells. The weighted SNV/SNP ratio

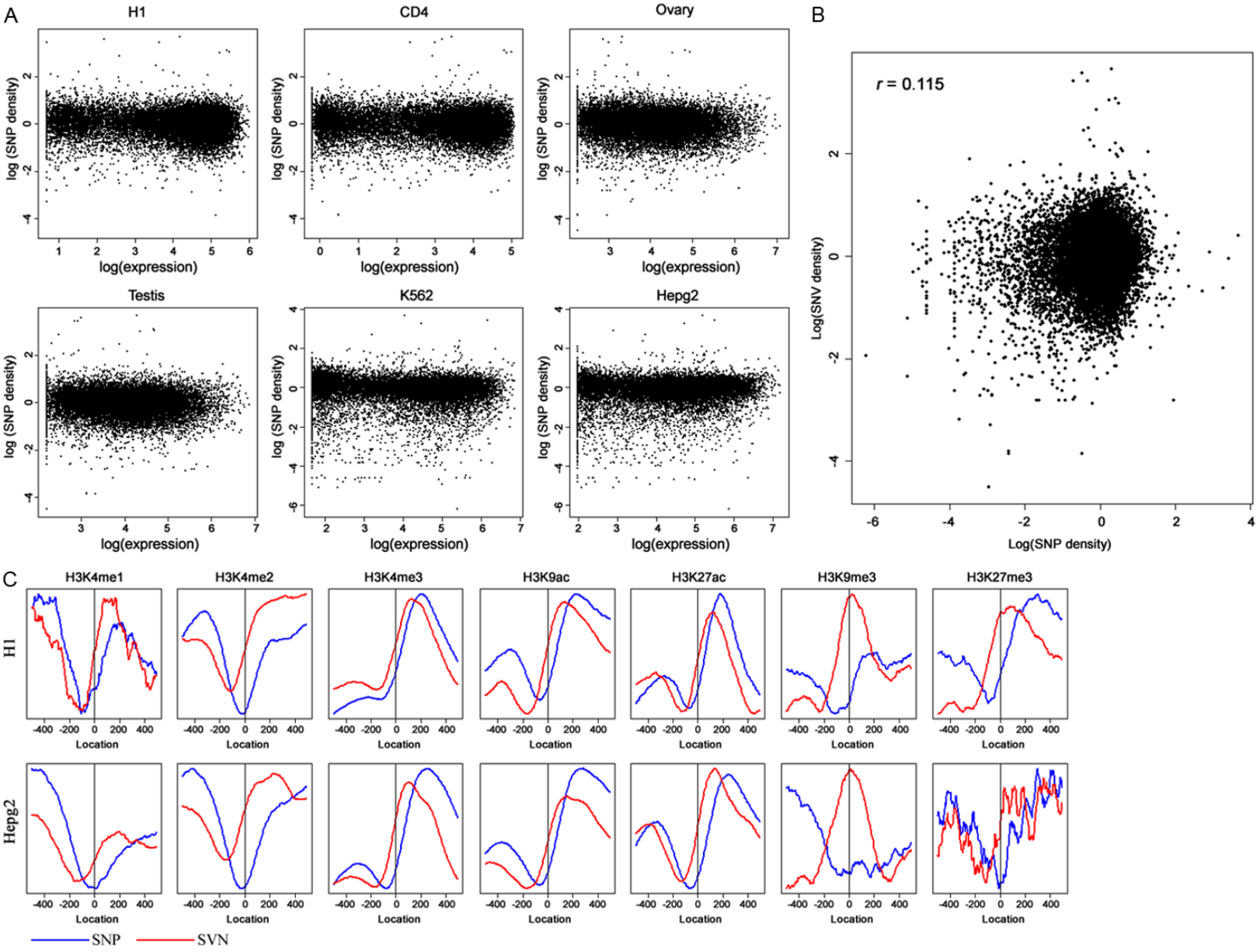

A
H1
CD4
Ovary
Testis
K562
Hepg2
log (SNP density)
log(expression)
B
r = 0.115
Log(SNV density)
Log(SNP density)
C
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K27ac
H3K9me3
H3K27me3
H1
Hepg2
Location
SNP
SVN

Figure 2. The correlation of SNP, SNV with gene expression and epigenetics marks. A. The correlation between SNV/SNP density and gene expression in six cells. B. The correlation between SNV density and SNP density in 17,498 genes. C. The profile of seven epigenetics markers around SNV and SNP sites in H1-ESC and Hepg2 cells.

(WVPR) is the ratio of MRC to MRN and was used to estimate gene mutation risk for each gene. The formula is as following:

$$\begin{aligned}\mathrm{MRN} = {} & W_{pa} \times M_{pa} + W_{\mathrm{pd}} \times M_{\mathrm{pd}} + W_{\mathrm{pe}} \times M_{\mathrm{pe}} + \\ & W_{\mathrm{pi}} \times M_{\mathrm{pi}} + W_{\mathrm{pp}} \times M_{\mathrm{pp}} + W_{\mathrm{pt}} \times M_{\mathrm{pt}}\end{aligned}$$

$$\begin{aligned}\mathrm{MRC} = {} & W_{va} \times M_{va} + W_{\mathrm{vd}} \times M_{\mathrm{vd}} + W_{\mathrm{ve}} \times M_{\mathrm{ve}} + \\ & W_{\mathrm{vi}} \times M_{\mathrm{vi}} + W_{\mathrm{vp}} \times M_{\mathrm{vp}} + W_{\mathrm{vt}} \times M_{\mathrm{vt}}\end{aligned}$$

$$\mathrm{WVPR} = \frac{\mathrm{MRC}}{\mathrm{MRN}}$$

Here, the "W" and "M" represent the weight parameters and mutation densities; "p" indicates common SNP; "v" indicates cancer SNV; a, d, e, I, p and t represent six different ranges of gene including acceptor, donor, exon, intron, promoter and tail respectively.

Fisher test for KEGG pathways

Fisher's exact test was performed to identify KEGG pathways enriched in cancer driver genes. A two-way contingency table was created based on the numbers of common SNPs and cancer SNVs in/out of a certain pathway to calculate the *p* values using R.

Results

Distinct characteristics of SNPs and SNVs

In this research, we collected 13,608,948 common SNPs and 2,342,135 unique cancer SNVs in which 5,140,763 and 1,735,291 mutation sites were found from 1000bp upstream to 1000bp downstream of 17,498 genes. All of these genes expressed at least in three out of six cells (H1-ESC, CD4, K562 Testis, Ovary and Hepg2) (see 'Materials and methods' section). All of SNPs and SNVs are classified into six groups based on the elements they locate on (promoter (upstream 1000bp and downstream 200bp of TSS), exon, intron, donor (around 36bp of split site 5') [21], acceptor (upstream 36bp and downstream 24bp of split site 3') [22] and tail (upstream 200bp and downstream 1000bp of TTS)) (**Figure 1A**). As shown in **Figure 1B**, although SNPs occur in non-coding regions with little higher frequently than in coding regions, in general, the number of SNPs is highly correlated with the length of elements (**Figure 1C**) suggesting their nature of neutral mutations. SNVs are highly enriched in exons and two splicing sites (donor and acceptor), which is consistent with previous reports [23-25]. It is notable that both majorities of SNPs (65,697/84,938) and SNVs (690,827/886,381) in exons are non-silent mutations and have almost consistent percentage (77.34% and 77.93%) in all SNP and SNV suggesting that many non-silent mutations are actually neutral. More important, no significant correlation between densities of SNV and SNP (R = 0.115) has been found (**Figure 2A**). And the size of genes is not correlated with SNV/SNP ratio (Figure S1). As that whether highly expressed genes in cancer cell have elevated mutation rates is a controversial question [26, 27], we checked the correlation of mutation density and gene expression. We calculated the correlations between SNP density and gene expression in four normal cell lines/tissues (H1, CD4, Ovary, Testis) and correlations between SNV density and gene expression in two cancer cell lines (K562, Hepg2). No significant correlations were found in all tests supporting the null hypothesis that the gene expression doesn't take an essential role in regulating SNV and SNP distribution. (**Figure 2B**).

Distinct chromatin structure at SNV and SNP mutation sites

One possible mechanism to affect generation of DNA sequence variations is the alteration of chromatin structure [28]. However what are epigenetic statues that correlate with SNPs and SNVs occurrence are still unknown. To address the question, we calculated the accumulated profiles of active transcriptional epigenetic markers (such as H3K4me1, H3K4me2, H3K4me3, H3K9ac and H3K27ac) and transcriptional repressive markers (such as H3K9me3 and H3K27me3) at SNPs and SNVs site in H1 (normal human ES Cell) and Hepg2 (a liver cancer cell) cells. We found that the profiles of the same marker from two cells are usually similar. However, the profiles of H3K4me2, H3K9ac and H3K9me3 are significant different

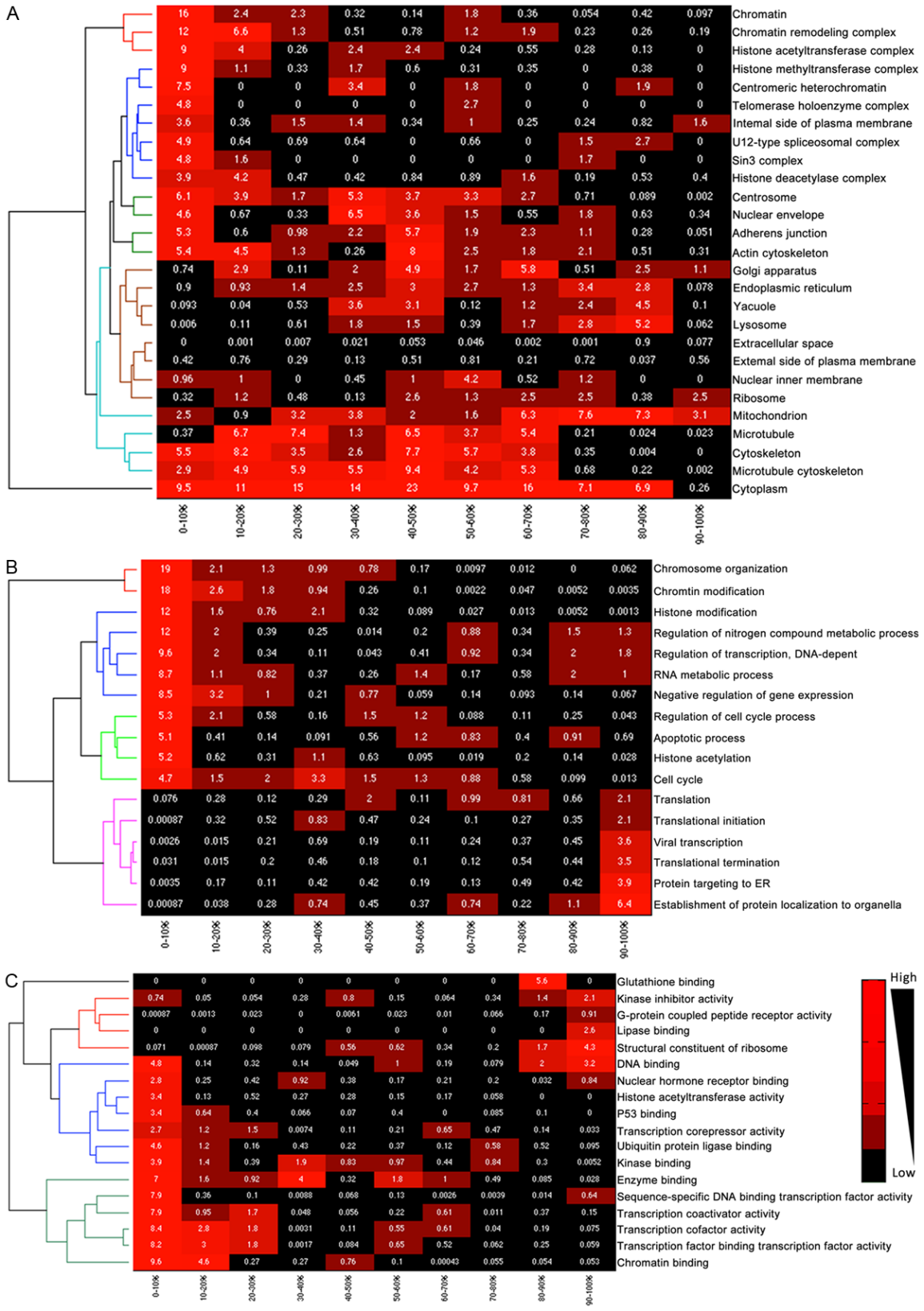
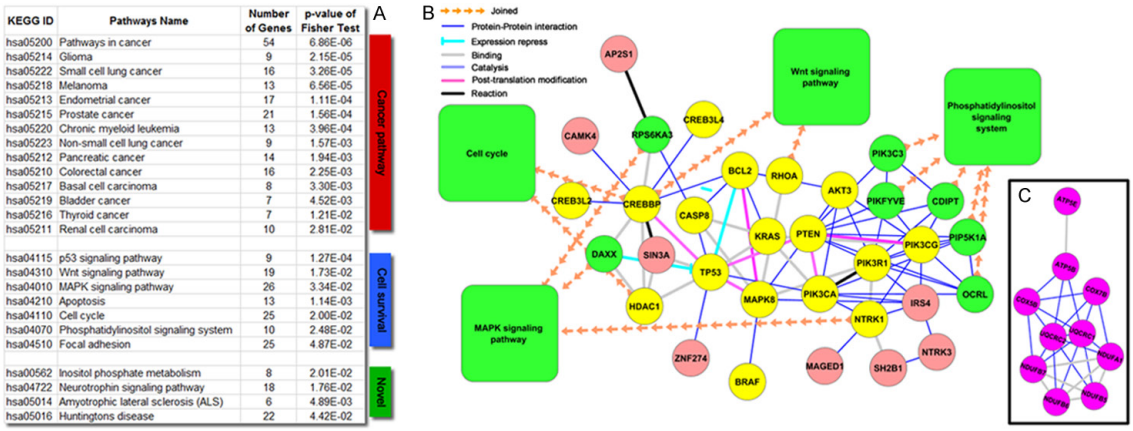


Figure 3. Discovering functional preference of driver genes by GO analysis. All of genes were categorized into ten groups according to their WVPR values. Hypergeometric test was used to test the enrichment of genes with certain GO items in these groups. The number in each grid is the -log10 (P), where P is the *p* value of hypergeometirc test. A.

The enriched GO items in cell component namespace. B. The enriched GO items in molecular function namespace. C. The enriched GO items in biology processes namespace.



KEGG ID	Pathways Name	Number of Genes	p-value of Fisher Test	
hsa05200	Pathways in cancer	54	6.86E-06	Cancer pathway
hsa05214	Glioma	9	2.15E-05	
hsa05222	Small cell lung cancer	16	3.26E-05	
hsa05218	Melanoma	13	6.56E-05	
hsa05213	Endometrial cancer	17	1.11E-04	
hsa05215	Prostate cancer	21	1.56E-04	
hsa05220	Chronic myeloid leukemia	13	3.96E-04	
hsa05223	Non-small cell lung cancer	9	1.57E-03	
hsa05212	Pancreatic cancer	14	1.94E-03	
hsa05210	Colorectal cancer	16	2.25E-03	
hsa05217	Basal cell carcinoma	8	3.30E-03	
hsa05219	Bladder cancer	7	4.52E-03	
hsa05216	Thyroid cancer	7	1.21E-02	
hsa05211	Renal cell carcinoma	10	2.81E-02	
hsa04115	p53 signaling pathway	9	1.27E-04	Cell survival
hsa04310	Wnt signaling pathway	19	1.73E-02	
hsa04010	MAPK signaling pathway	26	3.34E-02	
hsa04210	Apoptosis	13	1.14E-03	
hsa04110	Cell cycle	25	2.00E-02	
hsa04070	Phosphatidylinositol signaling system	10	2.48E-02	
hsa04510	Focal adhesion	25	4.87E-02	
hsa00562	Inositol phosphate metabolism	8	2.01E-02	Novel
hsa04722	Neurotrophin signaling pathway	18	1.76E-02	
hsa05014	Amyotrophic lateral sclerosis (ALS)	6	4.89E-03	
hsa05016	Huntingtons disease	22	4.42E-02	

Figure 4. Pathways and networks enriched with cancer driver genes. A. Table of KEGG pathways enriched with genes with top 10% WVPR, which are classified into three groups: known cancer pathways, cell survival pathways and novel pathways. B. The core network of cancer driver genes. All genes are represented as circles. They are linked by lines with different colors representing interactions and regulation among them. C. A novel pathway of cancer, the members are location on mitochondrial inner membrane and involving oxidative phosphorylation (Pathway 2).

between SNPs and SNVs sites in both cells. The SNP sites have lower level of two transcription activation epigenetics markers (H3K4me2, H3K9ac) than around regions and localize at the bottom of valleys, And SNV sites however usually localized at the boundary between regions with high and low level of two epigenetic markers. Intriguingly, the difference of H3K9me3 profiles around SNP and SNV sites are totally different. Usually, SNP sites have low H3K9me3 marker but SNV sites are rich for the modification, which is consistent to recent study [29]. Similar observation also found for H3K27me3 profile in J1 cell. Taken together, SNPs are enriched at the regions free of epigenetics markers and SNVs are usually found at chromatin structure transition regions which are repressed by repressive epigenetic markers (**Figure 2C**).

Ranking genes by the weighted SNV/SNP ratio (WVPR)

It is reasonable to speculate that driver genes usually have higher SNV density and lower SNP density, as individuals who have mutations in these genes usually get more chances to be eliminated by cancer. This hypothesis also was supported by the analyses of known driver genes such as GATA1 (**Figure 1E**). Thus the simplest way to identify driver genes is ranking genes with ratio of density of SNV to SNP. However, we found although working well for most of driver genes, this method is not sensitive to some driver genes that have high SNP density such as TP53, PTEN, NF2. To improve its performance, we used a line model to speculate the weighted SNV and SNP ratio by multiplying relative frequencies of each group with the densities of them in each gene. Details can be found in the materials and method section. The formula is as following:

All of genes were then sorted by the ratio of weighted SNV to weighted SNP (WVPR) and classified into ten groups for further analysis.

Method validation

To validate the method, we divided the sorted gene list into ten groups and tested whether known driver genes are enriched in groups with high WVPR. We extracted potential driver genes based on annotations in Cosmic, OMIM and KEGG database. And a driver gene list presented by Bert Vogelstein et al. was also included. As shown in **Figure 1E**, there is a clear trend of enrichment for each dataset on highly ranked groups. Especially 70% of known driver genes are ranked in top 10% gene group. And most of the well-known oncogenes and tumor suppressor (for example: TP53, PTEN, VHL, NF2, GATA1)

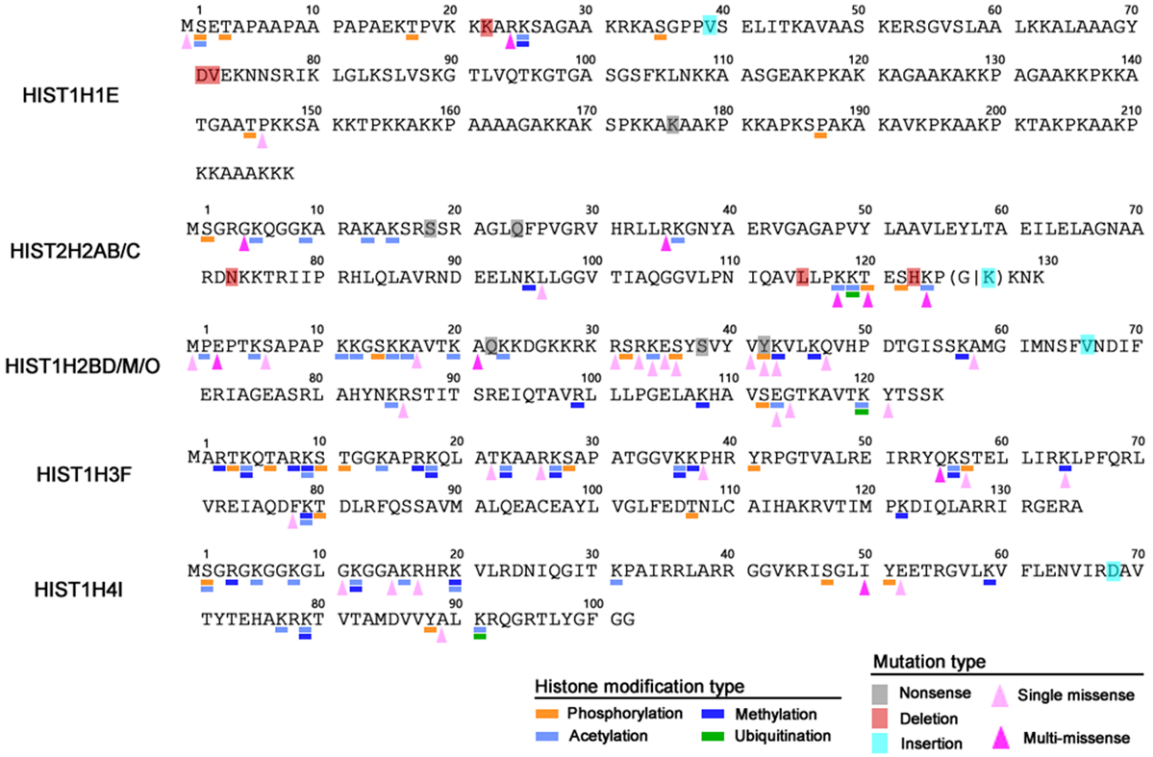


Figure 5. The distribution of driver mutations in histone family members with high WVPRs. The histone modification sites are marked by colored rectangles and mutations are represented by triangles and colored backgrounds as the legends in the figure.

have top 10 highest WVPR scores. Some high-risk genes reported by recent GWAS studies are also included in top 10% such as STAT4 and TNFAIP3. They were firstly linked to human diseases in GWAS researches on hepatitis B virus-related hepatocellular carcinoma [30] and systemic lupus erythematosus [31] respectively. We categorized the top 10% genes into 18 gene families and some of them were not reported by previously studies such as ANKRD family (involving cell cycle, immune response, cell structure and cell's signaling); ZNF family (as key role in gene transcription especially C2H2 zinc finger proteins); Histone family from H1 to H4 (contracture nucleosome); PCDHC family (involving cell adhere) (Table S2). The WVPR distribution in top 10% was shown in Figure S2. These results support the high accuracy of our prediction method and indicate there are more driver genes remained to be discovered.

GO analysis

To understand the functional preference of driver genes, we identify the enriched GO terms for all gene groups by Kolmogorov-Smirnov test. Using the value of -log p where p is the p-value of the test, we construct three matrixes for three GO name spaces and did hierarchical clustering to classify enriched GO items. As **Figure 3A** shown, the chromosome organization and its related biological processes (such as histone modification) are exclusive enriched in top groups suggesting their significant role in cancer development, which is consistent with previous reports [32-34]. Other processes of transcription regulation, cell cycle regulation and apoptosis are also highly enriched in high ranked groups. It is interesting that genes involved in translation and transportation to organelles have less SNVs than others in cancer cells and are highly enriched in the group with lowest WVPR, which suggests that although having lower possibility to be driver genes, these genes are important for viability of cancer cells. (**Figure 3A**).

Intriguingly, in cellular component matrix, we found that chromatin remodeling complexes,

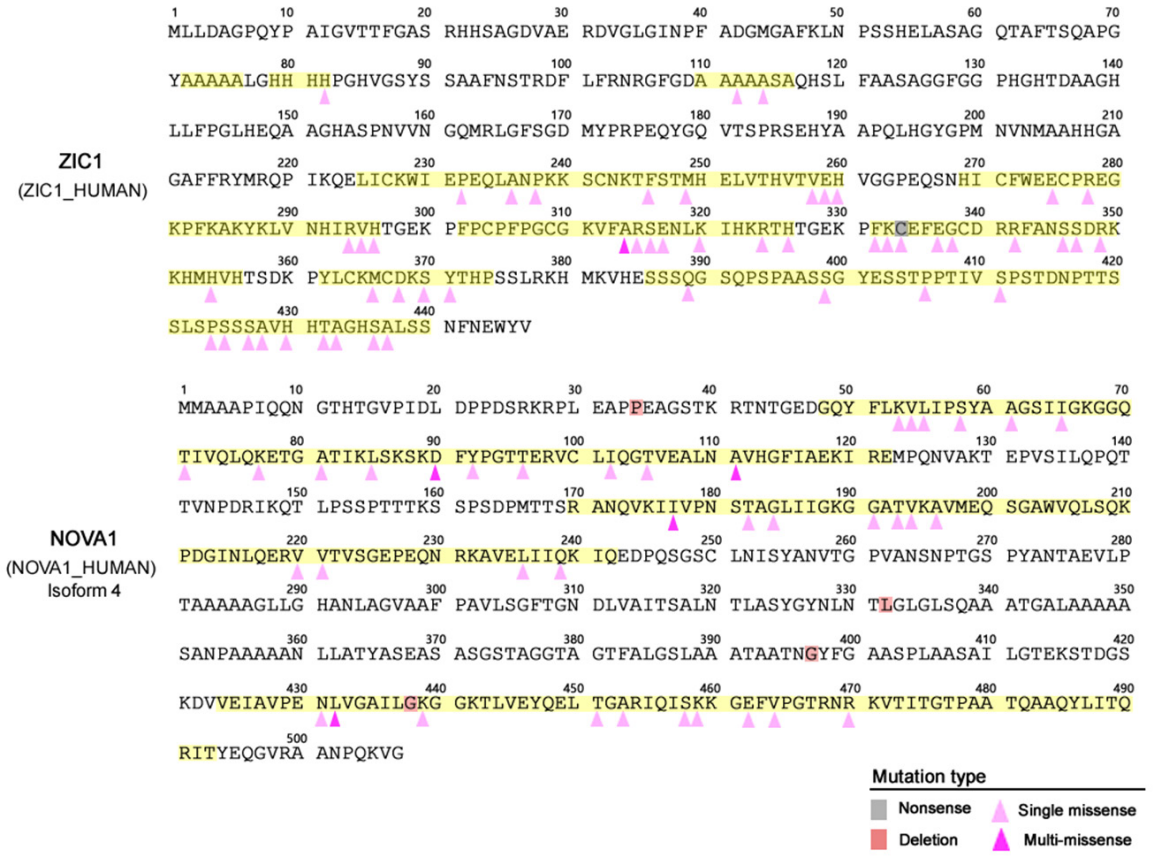


Figure 6. The distribution of missense mutations in ZIC1 and NOVA1. Mutations are represented as colored triangles and the domain regions are highlighted with yellow background.

especially histone acetytransferse complex and histone methytransferase complex are hottest spots of cancer driver mutations. For example, MEN1, OGT, RUVBL1, TAF1L have high WVPR in which MEN1 location on 27 in top 10% highest WVPR. These results suggest that histone modification alterations are one of most fundamental driving mechanisms for tumor genesis. Additionally, genes forming centromere and telomerase complex are also highly enriched in top ranked groups. It makes sense because centromere and telomerase mutations have long been linked to cancer development [35, 36]. Furthermore, some complexes, such as U12-type spliceosomal complex and Sin3 complex, which were not deemed by previous studies, were firstly found as hotspots of driver mutations and remain for further study. Finally, no significant enrichment of driver genes was found in other cellular components such as Goligi apparatus, lysosome, ribosome, cytoskeleton and nuclear inner membrane. (**Figure 3B**).

In molecular function namespace, chromatin-binding genes and transcription cofactors are highly enriched in top ranked group (**Figure 3C**). The *p*-value of enrichment analysis in GO item for ten groups have been shown in Tables S5, S6, S7. It is consistent with previous observations [37, 38] and reinforces the conclusion that alterations of cis and trans transcription regulation is the major driver force of cancer development.

Discovering cancer driver genes enriched pathways and networks

To discover the pathways involved in cancer development, we searched KEGG pathway database with top 10% high WVPR genes. 25 pathways are rich in these genes (149 genes in total) significantly by Fisher statistic test ($P < 0.01$). These pathways can be categorized into three groups including 14 cancer related pathways, 7 cell survival pathways and 4 novel pathways (**Figure 4A**). In cancer related pathways

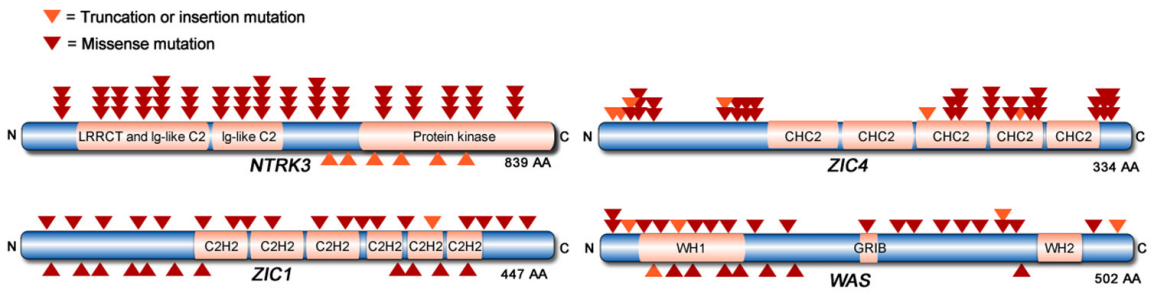


Figure 7. The different patterns of mutations in two cancer driver genes (NTRK3 and ZIC4) and two tumor suppressors (ZIC1 and WAS). For NTRK3 and ZIC4, mutations information was obtained from the COSMIC database and recurrent mutations including truncation or insertion in different samples have been shown. For ZIC1 and WAS, the first 30 mutation sites are plotted from COSMIC database.

group, 62 high WVPR genes were found in the common cancer pathway (hsa05200) with the most significant *p*-value (6.97e-06) suggesting good accuracy of our method. The cell survival groups include p53 signaling pathway, cell cycle, apoptosis, Wnt, MAPK and phosphatidylinositol and ubiquitin mediated proteolysis signaling system. All of them have long been thought related to cancer development [5, 39-44]. Our data suggests what components are the driver parts of the cancer pathways. For novel pathways, which wasn't linked to cancer before, 50 unique genes are involved in four potential signal pathways including inositol phosphate metabolim, neurotrophin signaling pathway, amyotrophic lateral sclerosis and Huntinton's disease. Interestingly, three of them related to neuron diseases. Whether some types of cancer sharing similar bio molecular mechanisms with these diseases remain further exploration.

The ranked list of cancer driver genes also presents a good opportunity to construct a core network of cancer development. We use the high confident PPI (protein-protein interaction) data (*p*-value > 0.7) extracted from STRING database [45] to assemble the networks de novo. Finally 41 genes and two networks have been discovered which were named Network 1 and 2 (**Figure 4B**, **4C**). Network 1 contains 32 genes, more than half of them (HDAC1, TP53, AKT3, CREB3L4, CASP8, PTEN, MAPK8, PIK3CA, BCL2, RHOA, CREB3L2, CREBBP, KRAS, PIK3R1, NTRK1, PIK3CG, BRAF) have reported in known cancer pathway (yellow node), which forms a core of the network. For example, HDAC1 as a deacetylase is responsible for deacetylation lysine residues on the N-terminal part of histones (H2-H4), which are not only involved in chronic myeloid leukemia but also discovered effectively in cell cycle process [46, 47]. GSK3B is an oncogene in basal cell carcinoma, endometrial cancer, prostate cancer, and colorectal cancer and is involved in Wnt signaling pathway and two novel pathways in our result (Neurotrophin signaling pathway, Insulin signaling pathway) [48-54]. Other genes are usually involved in cell survival pathways (such as cell cycle and Wnt, MAPK and phosphatidylinositol signaling pathway) and may serve as interface of the core to link to other pathways. For example, DAXX is a transcription repressor and histone 3.3 specific chaperon and involved in MAPK signaling way. Recent reports suggested that mutations of DAXX result in dysfunction of telomeres and pancreatic neuroendocrine tumors [55].

Network 2 is constructed by 10 genes, which are involved in oxidative phosphorylation. It is notable that three out of five complexes in the network contain high mutation risk genes: Four genes (NDUFA1, NDUFB6, NDUFB5, NDUFB7) belong to mitochondrial respiratory chain complex I; two genes (UQCRC1 and UQCRC2) located on complex III; two genes (COX7B and COX5B) located on complex VI and ATP5B and ATP5E are subunits of F-type ATPases in complex V. As most cancer cells exhibit increased glycolysis for generation of ATP as a main source of their energy supply [56], this surprising result reveals that accumulated mutations and defect of oxidative phosphorylation pathway may be an initial step in cancer development. It also partially answers the question that why cancer cells prefer glycolysis but not oxidative phosphorylation even if oxygen is available.

Novel candidates of cancer driver genes

One of major goals of the study is founding new cancer driver genes. We discovered several gene families that are enriched in top 100 genes (Table S2). Here we focused on histone family and transcription related families.

Although the histone epigenetics modifications have long been linked to cancer development, until recently the missense mutations of histones were given more and more attentions [57, 58]. Our data suggests that histone family is a hot spot of cancer related mutations and more ten histone genes have high WVPRs. More importantly, we found that most of unsilent mutations locate on/around (± 1) epigenetics modification sites in H2A, H2B, H3 and H4 (Table S3, **Figure 5**). For example, in HIST2H2AB and HIST2H2AC, 45 missense mutations and 6 other mutations (including three deletion mutations at N73, L115 and H123, one insertion at K126 and two nonsense mutations at Q24 and S18) accumulate on 10 sites. In which 7 of 10 mutation sites locate on or aside histone modification sites respectively (Tables S8, S9, S10, S11, S12). Although, till now, few studies investigated the biological effect of mutations around these modification sites, it is reasonable to speculate that these mutations may affect the structure and epigenetic statue of chromatin because most of them are highly conserved in evolution. How do these mutations influence histone function and cancer development is an interesting topic for further study.

Aberrant transcription/translation regulation is a key step of cancer development. Some transcription/translcation factors (e.g. ZIC1, ZIC4, ZNF26, ZNF513, ZNF536, ZMYM3, HOXA1), which were not deemed by previous methods, were highlighted in our list. For example, ZIC1 is sequence-specific transcription factor which involving developmental regulatory and regulation cell cycle and cell migration in gastric cancer [59]. Mutation analysis showed that ZIC1 gene is rich in mutations including 31 silence mutations, 103 missense mutations, two nonsense and one unknown mutations. Intriguingly, 47/50 mutations accumulated on its five C2H2 domains, which only take 30% of protein in length and are responsible for DNA binding (**Figure 6**). NOVA1 is another example. 90% unsilent mutations of NOVA1 have occurs on its conserved three KH domains, which function as RNA binding domains. Till now, no studies linked it to cancer and it was thought playing a role in regulating RNA splicing or metabolism in a specific subset of developing neurons [60, 61].

The pattern of mutations

Recent studies suggested that, aside from mutation frequency, the pattern of mutations is also an important feature to identify Mut-driver genes. Oncogenes are usually recurrently mutated at the same positions, whereas tumor suppressor genes may mutate evenly through the gene body [1]. Based on the hypothesis, we tried to classify the top ranked genes into oncogenes and tumor suppressor by calculating the average mutation number per site for top 200 genes (Table S1). 98 driver cancer genes were identified as oncogenes based on the rules: more than 10% mutation sites are recurrently mutations. The other 102 genes may be tumor suppressor genes (Addition File). Two novel candidates of oncogenes including NTRK3 (SNVs are 54 times more than SNPs and 12% mutation sites have been repeated identified in different sample) and ZIC4 (SNVs are 14 times more than SNPs and 14% mutation site has been repeated identified in different sample) have high ratio of recurrently mutations were shown in **Figure 6**. Most of mutations of NTRK3 were discovered in lung and colon tumors. And the fusion protein ETV6-NTRK3 has been considered as a biomarker in breast carcinoma [62, 63]. Although the last studies suggested that NTRK3 is a potential tumor suppressor gene [64], the high mutation frequencies and recurrently mutation rate suggest that it looks like a oncogene. ZIC4 gene encodes a member of ZIC family of C2H2 type zinc finger protein. Although its function is unknown, member of this family were linked to several human diseases such as visceral heterotaxy, and paraneoplastic neurologic disorders [65, 66]. Another interesting observation about ZIC4 gene is that most of its cancer SNVs cluster at the exons encoding the N-terminal, C-terminal and two CHC2 domains of the proteins suggesting a potential function of ZIC4 in cancer development. Further experiments are needed to figure out the role of NTRK3 and ZIC4 in the process of cancer development. We also showed two genes (ZIC1 and WAS) as examples

of tumor suppressors that usually have even distribution of SNVs (**Figure 7**).

Discussion

Our analyses suggest that although both of SNPs and cancer SNVs are single-nucleotide polymorphisms, their underline driver mechanisms might be dramatically different. It is supported the result that SNPs are enriched at the regions free of epigenetics markers and SNVs usually was found at chromatin structure transition regions which accumulate some repressive epigenetic markers. Here we propose a model: SNP sites are usually more sensitive to mutagen attack in germ cell than SNVs sites because they are not protected by epigenetic marker binding proteins. Comparing with gene body, which usually occupied/protected by transcriptional and epigenetic factors, intergenic regions have higher chances to get SNP. The chromatin statue of transition regions is changing dynamically in cell and need by protected by certain repressive epigenetic mechanisms such as H3K9me3 or PRC2 complex, which recognizes H3K27me3 marker. In cancer cells, the defect of these repressive epigenetic mechanisms may result in high frequency of mutations within these regions.

Tumorigenesis is an evolutionary process of accumulation of somatic mutations (driver mutation), which promotes a selective growth advantage for cancer cells. Numerous statistical methods to identify driver genes have been proposed. Most of them estimated background mutation ratio using the ratio of frequency of no silent mutations to silent mutations with the hypothesis that most of non-silent mutations are unfavorable and will affect gene function as well as fitness of species. However the validity of the hypothesis is controversy. Because the ratio of non-silent mutations in SNP and SNV is comparable suggesting that in most cases the non-silent mutations are natural, then it means previous methods usually can't estimate background accurately.

One of significant advantages of our new method is using SNP to estimate the background mutation ratio for each gene. The SNP data cross human genome presents a natural map of neutral mutations. The genomic distribution of SNPs is not homogenous. For each gene, the final distribution of SNPs is the product of natural selection and affected by many factors such as mutation context, gene structure, location, size, nucleotide composition and basal mutation ratio, which vary among genes. As all of these factors already been taken into account, a complex model for correction is not necessary, which enables the method very simple. As the same SNP/SNV at different gene features might have distinct possibilities to be a neutral mutation/driver mutation, we used a line model to estimate the weight parameters for calculating the background as well as cancer mutation ratio, which dramatically increased the sensitivity of the method. In addition, it is notable that the new method also has some limitations. For example, as SNPs are the affixation result of neutral mutations in germ cell, using it to estimate the ratio of somatic neutral mutations in cancer cells may be risky, especially when the statue of gene (such as transcription statue, distribution of epigenetic markers) are dramatically different between germ cell and cancer cell. As a result, some driver genes for specific tumor type may not be identified by the new method efficiently.

The analysis of cancer genomics data is of key importance for understanding oncogenesis. Although vast amounts of cancer genome sequencing data are now available, deciphering this information to draw meaningful conclusions is still challenging. In this study we presented a large ranked list of cancer driver genes and highlighted a lot of new candidates for further analysis. Some complexes such as sin3 and U12-type spliceosomal complexes (Figure S3) and pathways such as oxidative phosphorylation pathway (Figure S4), which were not deemed by previous methods, now were linked to cancer development. Our study not only develops a method to identify cancer driver genes/pathways but also provides new insights into molecular mechanisms of cancer development. The WVPR of 17498 genes have been shown in additional file and the top 200 gene list also provided. Further experiments are needed to validate these findings in the future.

Acknowledgements

The study is supported by the National Key Sci-Tech Special Project of China (2013ZX10002010 and 2008ZX10002-020 to L.Y.), the National

Natural Science Foundation of China for Creative Research Groups (30024001 to L.Y.), the Project of the Shanghai Municipal Science and Technology Commission (to L.Y.), the National Natural Science Foundation of China (31071193 to L.Y. and 31100895 to D.-K.J.), Director Foundation of the State Key Laboratory of Genetic Engineering (to L.Y.).

Disclosure of conflict of interest

The conflict of interest is none among authors.

Address correspondence to: Quanze He or Long Yu, The State Key Laboratory of Genetic Engineering, Institute of Biomedical Science, Fudan University, 220 Handan Rd, Shanghai 200433, China. Tel: +86 021 65643713; Fax: +86 021 65643250; E-mail: hqzlul@gmail.com (QH); longyu@fudan.edu.cn (LY)

References

[1] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr and Kinzler KW. Cancer genome landscapes. Science 2013; 339: 1546-1558.

[2] Zhang J, Liu J, Sun J, Chen C, Foltz G and Lin B. Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. Brief Bioinform 2014; 15: 244-55.

[3] Youn A and Simon R. Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics 2011; 27: 175-181.

[4] Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA and Stratton MR. Patterns of somatic mutation in human cancer genomes. Nature 2007; 446: 153-158.

[5] Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M and Wilson RK. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 2008; 455: 1069-1075.

[6] Vinagre J, Almeida A, Populo H, Batista R, Lyra J, Pinto V, Coelho R, Celestino R, Prazeres H, Lima L, Melo M, da Rocha AG, Preto A, Castro P, Castro L, Pardal F, Lopes JM, Santos LL, Reis RM, Cameselle-Teijeiro J, Sobrinho-Simoes M, Lima J, Maximo V and Soares P. Frequency of TERT promoter mutations in human cancers. Nat Commun 2013; 4: 2185.

[7] Salvesen HB, Kumar R, Stefansson I, Angelini S, MacDonald N, Smeds J, Jacobs IJ, Hemminki K, Das S and Akslen LA. Low frequency of BRAF and CDKN2A mutations in endometrial cancer. Int J Cancer 2005; 115: 930-934.

[8] Kwiatkowska E, Skasko E, Niwinska A, Wojciechowska-Lacka A, Rachtan J, Molong L, Nowakowska D, Konopka B, Janiec-Jankowska A, Paszko Z and Steffen J. Low frequency of the CHEK2*1100delC mutation among breast cancer probands from three regions of Poland. Neoplasma 2006; 53: 305-308.

[9] Jiang L, Huang J, Morehouse C, Zhu W, Korolevich S, Sui D, Ge X, Lehmann K, Liu Z, Kiefer C, Czapiga M, Su X, Brohawn P, Gu Y, Higgs BW and Yao Y. Low frequency KRAS mutations in colorectal cancer patients and the presence of multiple mutations in oncogenic drivers in non-small cell lung cancer patients. Cancer Genet 2013; 206: 330-9.

[10] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME and Visscher PM. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 2010; 42: 565-569.

[11] Visscher PM, Yang J and Goddard ME. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). Twin Res Hum Genet 2010; 13: 517-524.

[12] Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjorib-

anks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP and Cox DR. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 2001; 294: 1719-1723.

[13] Nachman MW. Single nucleotide polymorphisms and recombination rate in humans. Trends Genet 2001; 17: 481-485.

[14] ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrímsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447: 799-816.

[15] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR and Futreal PA. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 2011; 39: D945-950.

[16] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S and Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 2013; 41: D991-995.

[17] Pruitt KD, Tatusova T and Maglott DR. NCBI Reference Sequence (RefSeq): a curated nonredundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005; 33: D501-504.

[18] Langmead B, Trapnell C, Pop M and Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009; 10: R25.

[19] Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR,

Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D and Kent WJ. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res 2013; 41: D64-69.

[20] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W and Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008; 9: R137.

[21] Joseph DR, Hall SH, Conti M and French FS. The gene structure of rat androgen-binding protein: identification of potential regulatory deoxyribonucleic acid elements of a follicle-stimulating hormone-regulated protein. Mol Endocrinol 1988; 2: 3-13.

[22] Fu QH, Zhou RF, Liu LG, Wang WB, Wu WM, Ding QL, Hu YQ, Wang XF, Wang ZY and Wang HL. Identification of three F5 gene mutations associated with inherited coagulation factor V deficiency in two Chinese pedigrees. Haemophilia 2004; 10: 264-270.

[23] Diez O and Gutierrez-Enriquez S. BRCA2 splice site mutations in an Italian breast/ovarian cancer family. Ann Oncol 2009; 20: 1285; author reply 1285-1286.

[24] Bianchi F, Rosati S, Belvederesi L, Loretelli C, Catalani R, Mandolesi A, Bracci R, Bearzi I, Porfiri E and Cellerino R. MSH2 splice site mutation and endometrial cancer. Int J Gynecol Cancer 2006; 16: 1419-1423.

[25] Walsh T, Casadei S, Coats KH, Swisher E, Stray SM, Higgins J, Roach KC, Mandell J, Lee MK, Ciernikova S, Foretova L, Soucek P and King MC. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. JAMA 2006; 295: 1379-1388.

[26] Park C, Qian W and Zhang J. Genomic evidence for elevated mutation rates in highly expressed genes. EMBO Rep 2012; 13: 1123-1129.

[27] Masica DL and Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer Res 2011; 71: 4550-4561.

[28] Tolstorukov MY, Volfovsky N, Stephens RM and Park PJ. Impact of chromatin structure on sequence variability in the human genome. Nat Struct Mol Biol 2011; 18: 510-515.

[29] Schuster-Bockler B and Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 2012; 488: 504-507.

[30] Jiang DK, Sun J, Cao G, Liu Y, Lin D, Gao YZ, Ren WH, Long XD, Zhang H, Ma XP, Wang Z, Jiang W, Chen TY, Gao Y, Sun LD, Long JR, Huang HX, Wang D, Yu H, Zhang P, Tang LS, Peng B, Cai H, Liu TT, Zhou P, Liu F, Lin X, Tao S, Wan B, Sai-Yin HX, Qin LX, Yin J, Liu L, Wu C, Pei Y, Zhou YF, Zhai Y, Lu PX, Tan A, Zuo XB, Fan J, Chang J, Gu X, Wang NJ, Li Y, Liu YK, Zhai K, Zhang H, Hu Z, Liu J, Yi Q, Xiang Y, Shi R, Ding Q, Zheng W, Shu XO, Mo Z, Shugart YY, Zhang XJ, Zhou G, Shen H, Zheng SL, Xu J and Yu L. Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma. Nat Genet 2013; 45: 72-75.

[31] Han JW, Zheng HF, Cui Y, Sun LD, Ye DQ, Hu Z, Xu JH, Cai ZM, Huang W, Zhao GP, Xie HF, Fang H, Lu QJ, Xu JH, Li XP, Pan YF, Deng DQ, Zeng FQ, Ye ZZ, Zhang XY, Wang QW, Hao F, Ma L, Zuo XB, Zhou FS, Du WH, Cheng YL, Yang JQ, Shen SK, Li J, Sheng YJ, Zuo XX, Zhu WF, Gao F, Zhang PL, Guo Q, Li B, Gao M, Xiao FL, Quan C, Zhang C, Zhang Z, Zhu KJ, Li Y, Hu DY, Lu WS, Huang JL, Liu SX, Li H, Ren YQ, Wang ZX, Yang CJ, Wang PG, Zhou WM, Lv YM, Zhang AP, Zhang SQ, Lin D, Li Y, Low HQ, Shen M, Zhai ZF, Wang Y, Zhang FY, Yang S, Liu JJ and Zhang XJ. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat Genet 2009; 41: 1234-1237.

[32] Cross NC. Histone modification defects in developmental disorders and cancer. Oncotarget 2012; 3: 3-4.

[33] Biancotto C, Frige G and Minucci S. Histone modification therapy of cancer. Adv Genet 2010; 70: 341-386.

[34] Chervona Y and Costa M. Histone modifications and cancer: biomarkers of prognosis? Am J Cancer Res 2012; 2: 589-597.

[35] Blackburn EH. Telomerase and Cancer: Kirk A. Landon--AACR prize for basic cancer research lecture. Mol Cancer Res 2005; 3: 477-482.

[36] Shay JW, Zou Y, Hiyama E and Wright WE. Telomerase and cancer. Hum Mol Genet 2001; 10: 677-685.

[37] Shaikhibrahim Z and Wernert N. ETS transcription factors and prostate cancer: the role of the family prototype ETS-1 (review). Int J Oncol 2012; 40: 1748-1754.

[38] Shimizu R, Engel JD and Yamamoto M. GATA1-related leukaemias. Nat Rev Cancer 2008; 8: 279-287.

[39] Stegh AH. Targeting the p53 signaling pathway in cancer therapy - the promises, challenges and perils. Expert Opin Ther Targets 2012; 16: 67-83.

[40] Collins K, Jacks T and Pavletich NP. The cell cycle and cancer. Proc Natl Acad Sci U S A 1997; 94: 2776-2778.

[41] Lowe SW and Lin AW. Apoptosis in cancer. Carcinogenesis 2000; 21: 485-495.

[42] Lascorz J, Forsti A, Chen B, Buch S, Steinke V, Rahner N, Holinski-Feder E, Morak M, Schackert HK, Gorgens H, Schulmann K, Goecke T, Kloor M, Engel C, Buttner R, Kunkel N, Weires M, Hoffmeister M, Pardini B, Naccarati A, Vodickova L, Novotny J, Schreiber S, Krawczak M, Broring CD, Volzke H, Schafmayer C, Vodicka P, Chang-Claude J, Brenner H, Burwinkel B, Propping P, Hampe J and Hemminki K. Genome-wide association study for colorectal cancer identifies risk polymorphisms in German familial cases and implicates MAPK signalling pathways in disease susceptibility. Carcinogenesis 2010; 31: 1612-1619.

[43] Hernandez-Aya LF and Gonzalez-Angulo AM. Targeting the phosphatidylinositol 3-kinase signaling pathway in breast cancer. Oncologist 2011; 16: 404-414.

[44] Ciechanover A, Orian A and Schwartz AL. Ubiquitin-mediated proteolysis: biological regulation via destruction. Bioessays 2000; 22: 442-451.

[45] Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C and Jensen LJ. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 2013; 41: D808-815.

[46] Ammanamanchi S, Freeman JW and Brattain MG. Acetylated sp3 is a transcriptional activator. J Biol Chem 2003; 278: 35775-35780.

[47] Wilting RH, Yanover E, Heideman MR, Jacobs H, Horner J, van der Torre J, DePinho RA and Dannenberg JH. Overlapping functions of Hdac1 and Hdac2 in cell cycle regulation and haematopoiesis. EMBO J 2010; 29: 2586-2597.

[48] Boyle WJ, Smeal T, Defize LH, Angel P, Woodgett JR, Karin M and Hunter T. Activation of protein kinase C decreases phosphorylation of c-Jun at sites that negatively regulate its DNA-binding activity. Cell 1991; 64: 573-584.

[49] Welsh GI and Proud CG. Glycogen synthase kinase-3 is rapidly inactivated in response to insulin and phosphorylates eukaryotic initiation factor eIF-2B. Biochem J 1993; 294: 625-629.

[50] Beals CR, Sheridan CM, Turck CW, Gardner P and Crabtree GR. Nuclear export of NF-ATc enhanced by glycogen synthase kinase-3. Science 1997; 275: 1930-1934.

[51] Li Y, Bharti A, Chen D, Gong J and Kufe D. Interaction of glycogen synthase kinase 3beta with the DF3/MUC1 carcinoma-associated antigen and beta-catenin. Mol Cell Biol 1998; 18: 7216-7224.

[52] Yook JI, Li XY, Ota I, Fearon ER and Weiss SJ. Wnt-dependent regulation of the E-cadherin repressor snail. J Biol Chem 2005; 280: 11740-11748.

[53] Hashimoto YK, Satoh T, Okamoto M and Takemori H. Importance of autophosphorylation at Ser186 in the A-loop of salt inducible kinase 1 for its sustained kinase activity. J Cell Biochem 2008; 104: 1724-1739.

[54] Heyd F and Lynch KW. Phosphorylation-dependent regulation of PSF by GSK3 controls CD45 alternative splicing. Mol Cell 2010; 40: 126-137.

[55] Jiao Y, Shi C, Edil BH, de Wilde RF, Klimstra DS, Maitra A, Schulick RD, Tang LH, Wolfgang CL, Choti MA, Velculescu VE, Diaz LA Jr, Vogelstein B, Kinzler KW, Hruban RH and Papadopoulos N. DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. Science 2011; 331: 1199-1203.

[56] Gatenby RA and Gillies RJ. Why do cancers have high aerobic glycolysis? Nat Rev Cancer 2004; 4: 891-899.

[57] Rheinbay E, Louis DN, Bernstein BE and Suva ML. A tell-tail sign of chromatin: histone mutations drive pediatric glioblastoma. Cancer Cell 2012; 21: 329-331.

[58] Moorefield B. Helicase disc breaks. Nat Struct Mol Biol 2013; 20: 1242.

[59] Zhong J, Chen S, Xue M, Du Q, Cai J, Jin H, Si J and Wang L. ZIC1 modulates cell-cycle distributions and cell migration through regulation of sonic hedgehog, PI(3)K and MAPK signaling pathways in gastric cancer. BMC Cancer 2012; 12: 290.

[60] Buckanovich RJ, Posner JB and Darnell RB. Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system. Neuron 1993; 11: 657-672.

[61] Jensen KB, Dredge BK, Stefani G, Zhong R, Buckanovich RJ, Okano HJ, Yang YY and Darnell RB. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron 2000; 25: 359-371.

[62] Tognon C, Knezevich SR, Huntsman D, Roskelley CD, Melnyk N, Mathers JA, Becker L, Carneiro F, MacPherson N, Horsman D, Poremba C and Sorensen PH. Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. Cancer Cell 2002; 2: 367-376.

[63] Li Z, Tognon CE, Godinho FJ, Yasaitis L, Hock H, Herschkowitz JI, Lannon CL, Cho E, Kim SJ, Bronson RT, Perou CM, Sorensen PH and Orkin SH. ETV6-NTRK3 fusion oncogene initiates breast cancer from committed mammary progenitors via activation of AP1 complex. Cancer Cell 2007; 12: 542-558.

[64] Luo Y, Kaz AM, Kanngurn S, Welsch P, Morris SM, Wang J, Lutterbaugh JD, Markowitz SD and Grady WM. NTRK3 is a potential tumor

suppressor gene commonly inactivated by epigenetic mechanisms in colorectal cancer. PLoS Genet 2013; 9: e1003552.

[65] Cowan J, Tariq M and Ware SM. Genetic and functional analyses of ZIC3 variants in congenital heart disease. Hum Mutat 2014; 35: 66-75.

[66] Bataller L, Wade DF, Graus F, Stacey HD, Rosenfeld MR and Dalmau J. Antibodies to Zic4 in paraneoplastic neurologic disorders and small-cell lung cancer. Neurology 2004; 62: 778-782.

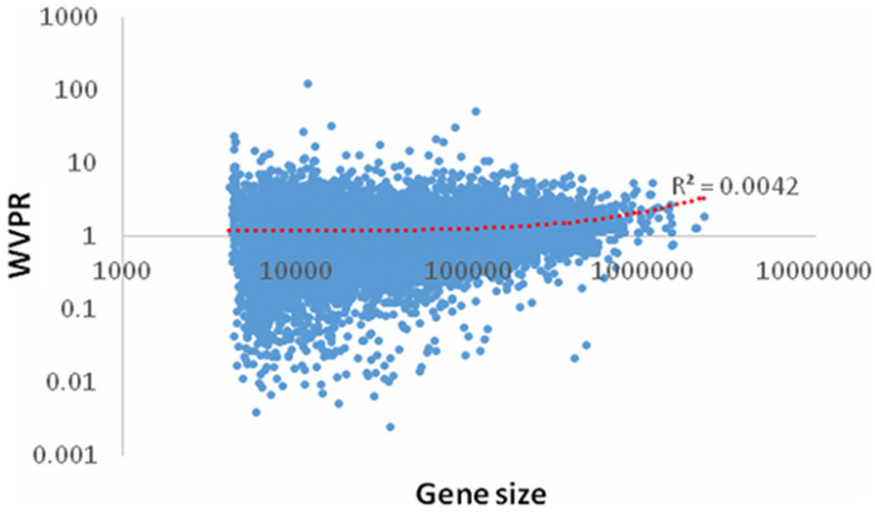


Figure S1. The dot plot of gene size against WVPR for 17,498 genes.

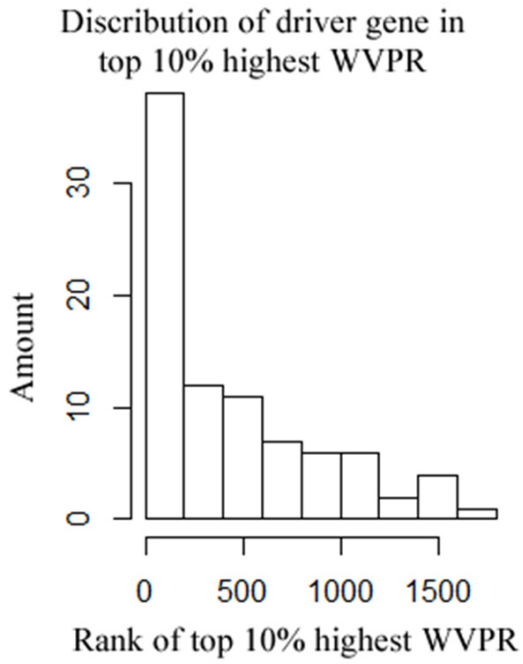
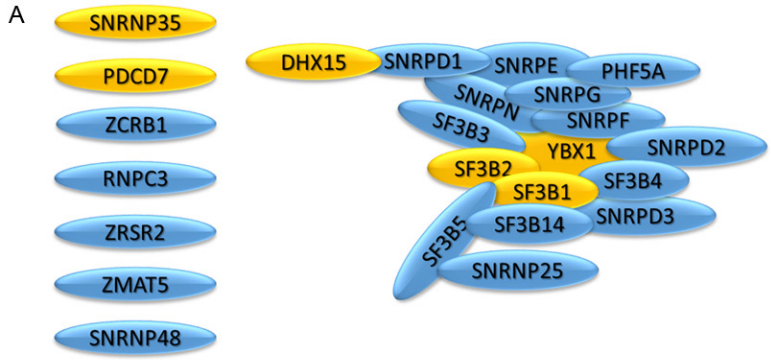


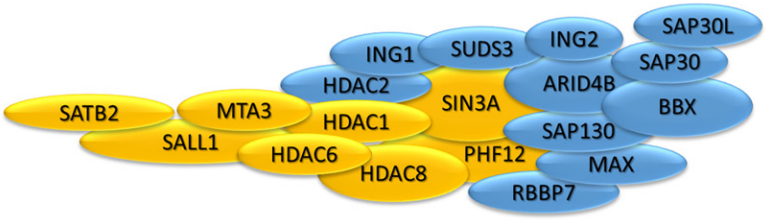
Figure S2. The distribution of known cancer driver genes in top 10% genes with highest WVPR.

A



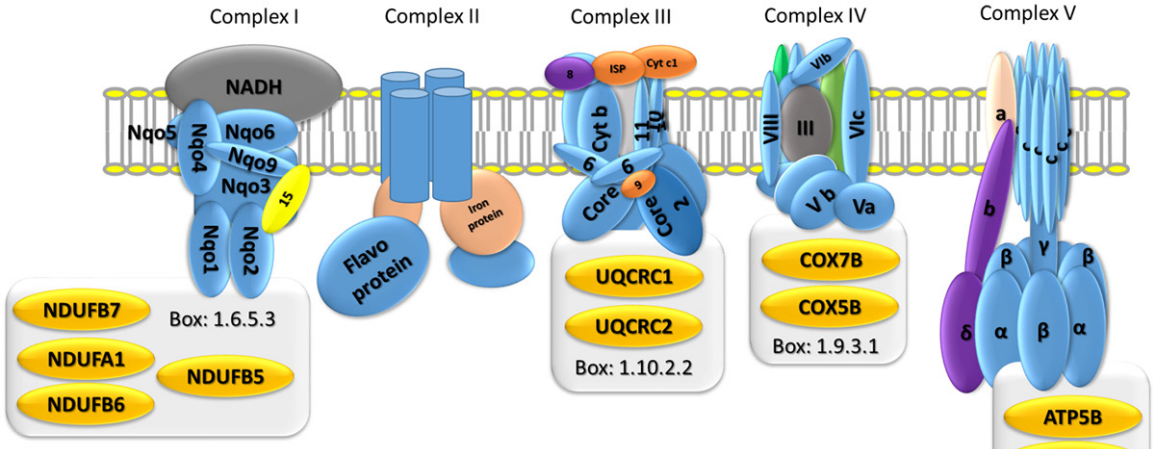
Gene Name	WVPR	Rank
SF3B1	4.403	296
YBX1	3.76	462
PDCD7	2.858	944
SNRNP35	2.850	949
DHX15	2.648	1152
SF3B2	2.321	1599

B



Gene Name	WVPR	Rank
SIN3A	4.498	279
HDAC6	4.498	278
SALL1	2.857	946
HDAC1	2.685	1109
HDAC8	2.476	1368
SATB2	2.497	1340
PHF12	2.264	1702
MTA3	2.245	1746

Figure S3. High WVPR genes are enriched in U12 and SIN3 protein complex. The protein-protein interaction (PPI) data was adopted from STRING database (Ver 9.05). Genes with high WVPR are marked by orange and others are colored as blue; some genes which were annotated as components of complex by GO but lack PPI information are listed aside. The values of WVPR and rank numbers of high WVPR genes are shown at the right side table. A. U12-type protein complex; B. SIN3 protein complex.



Gene Name	WVPR	Rank	Gene Name	WVPR	Rank	Gene Name	WVPR	Rank
NDUFA1	2.912	898	UQCRC1	2.590	1224	ATP5B	2.821	977
NDUFB6	2.539	1289	UQCRC2	2.422	1441	ATP5E	3.345	624
NDUFB7	2.526	1303	COX7B	4.665	250			
NDUFB5	3.340	627	COX5B	2.427	1442			

Figure S4. High WVPR genes are enriched in oxidative phosphorylation pathway. The architecture of the pathway follows the oxidative phosphorylation pathway of KEGG database (http://www.genome.jp/kegg/pathway/map/map00190.html). The high WVPR genes are marked by orange. The value of WVPR and rank number of these high WVPR genes are shown in the table at bottom.

Table S1. Mapping result of gene list with known oncogene list

Database/List name	Amount	Match	Miss match
COSMIC	455	451	4
OMIM	180	130	50
KEGG cancer pathway	162	163	1
Driver gene of cancer	125	125	0

Table S2. Gene families with high WVPR

Family name	Gene List
ANKRD	ANKRD17, BCOR, ASB12, ANKHD1, KRIT1, ASB11, ANKHD1-EIF4EBP3, ASB14, BCORL1, FPGT-TNNI3K, CDKN2C, NOTCH2, ANK2, NOTCH1, ANKIB1, SNCAIP
bHLH	TCF23, MYC, MAX, MITF, HAND2, MSC, HIF1A, TFEC, HES1, MYCN, TCF4
HOXL	HOXA1, HOXC6, HOXD4, HOXD11, HOXC8, HOXA2, HOXC5, HOXB1, HOXD1, HOXA13, MEOX2
HIST	HIST2H2AB, HIST2H2AC, HIST1H2BM, HIST1H2BD, HIST1H2BO, HIST1H1E, HIST1H3F, HIST1H4I, HIST1H2AG, HIST1H2BH, HIST1H2BL, HIST1H2AM, HIST1H2AK, HIST2H3D, HIST1H3H, HIST1H2BK, HIST3H2BB, HIST1H4L, HIST1H3G, HIST2H2BF, HIST1H2BB, HIST1H4D, HIST1H2BN, HIST1H2BC, HIST1H2AE, HIST1H2BI, HIST1H2AJ, HIST1H2AI, HIST1H3I, HIST2H2BE, HIST1H3E, HIST1H2BG, HIST1H2BJ, HIST1H3B, HIST1H3C, HIST1H2AH, HIST1H4G, HIST1H1D, HIST1H1C, HIST1H4E, HIST1H1T
OR2	OR2A1, OR2A42, OR2L2, OR2A25, OR2A4, OR2A14, OR2AK2, OR2B6, OR2K2, OR2A7, OR2AG2, OR2J3
ISET	IL1RAPL1, NTRK3, CNTN1, LRRC4C, PDGFRA, LRFN5, NTM, LRRN3, FGFR2, LRRC4, MDGA2, NEGR1, HMCN1, UNC5D, FLT1, PXDNL, TTN, OPCML, FSTL5, ADAMTSL1, NCAM2, CNTN5, L1CAM, MYOT, ROBO2, KALRN
PCDHC	PCDHA13, PCDHGA10, PCDHGA5, PCDHGB7, PCDHA6, PCDHGA1, PCDHGA2, PCDHGB3, PCDHB1, PCDHGA3, PCDHGB1, PCDHGA11, PCDHGA12, PCDHA1, PCDHGB4, PCDHB4, PCDHGA6, PCDHGB6, PCDHB7, PCDHA5, PCDHA7, PCDHGB2, PCDHB5, PCDHA9, PCDHB2, PCDHA3, PCDHGC4, PCDHB12, PCDHGA7, PCDHB15, PCDHB8, PCDHGA8, PCDHGC3, PCDHGA4, PCDHAC1, PCDHGB5
PHF	PHF6, PHF7, PHF16, KDM5C, PHF8, WHSC1, PYGO2, ASH1L, ING3, PYGO1, ING4, NSD1, SP140, PHF3, PHF12, BAZ1B
PLEKH	ARHGEF9, AKT3, GAB3, ARHGAP15, ARHGEF6, PHLDA2, PLEKHH3, ARHGEF2, RTKN, AGAP3, PRKD1, ARHGAP21, AGAP2
PRD	RHOXF2, RHOXF2B, ALX1, OTX2, ARX, OTX1, MIXL1, RHOXF1, ESX1, SEBOX, DUXA, PAX5, HESX1, SHOX2, PHOX2A
RBM	HTATSF1, NONO, ELAVL4, RBM14, SYNCRIP, ENOX2, PSPC1, ELAVL2, PPRC1, MYEF2, RALYL, RBMX, SNRNP35, RBM7, EWSR1, RBM14-RBM4, RBM45, IGF2BP2, HNRNPA1, CELF4, HNRNPH2, RBM5, ESRP2, RBPMS,
RNF	MKRN3, ZFP36, PJA1, CBL, SCAF11, RNF128, TRAIP, RNF145, MARCH11, 41337, ZFP36L1, RNF113B, LONRF3, MARCH1, MSL2, BRCA1, RPL10L, RPLP0, RPL5
SAMD	EPHA3, SCML2, DDHD2, PPFIA2, ASZ1, EPHA6, CNKSR2, EPHA5, SFMBT2, SAMHD1, EPHB1, SCML1, EPHA7, EPHB2, ARAP2
SH2D	PIK3R1, SUPT6H, SH2D1A, LCK, JAK2, ABL2, RASA1, HCK, SH2B1, STAT4, ZAP70, GRAP2
SLC	SLC25A14, SLC24A5, SLC35A2, SLC16A2, SLC38A5, SLC4A10, SLC10A7, SLC45A2, SLC4A3, SLC8A1, SLC38A6, SLC26A7, SLC17A8, SLC25A20, SLC10A3, SLC50A1, SLC25A31, SLC39A5, SLC9A6, SLC6A13, UCP1, SLC22A7, SLC9A5, SLC25A13, SLC44A5, SLC5A2
VSET	TREML1, CD79A, CD2, HEPACAM2, SIGLEC7, TIMD4, CADM3, SIGLEC8, KDR, CD86, PILRA, PVRL4, PVRL1
WDR	FBXW7, DCAF12L1, WDR6, WDR45, DCAF4L2, WDR54, DCAF8L2, BRWD3, DCAF12L2, WDR78, DCAF8L1, WDR49, DCAF6, NBEAL1, WDR13, PPP2R2B, RBBP7, WDR20, AMBRA1, RFWD2, FBXW4, EIF3I, TRAF7, CDC20
ZNF	ZIC1, SNAI2, ZNF513, CTCF, ZIC4, ZNF536, PRDM9, PRDM1, OSR2, WT1, ZNF267, ZNF688, ZNF747, ZNF296, ZIK1, SALL4, ZNF827, BCL11A, EGR2, ZNF254, ZNF689, FEZF2, BCL11B, ZNF548, ZNF676, ZNF790, ZFPM2, PEG3, ZNF449, ZBTB16, BNC1, ZNF521, ZNF274, ZNF451, TRPS1, EGR3, ZNF786, SALL1, ZIM2, ZNF423, ZNF514, ZFP37, KLF9, IKZF3, ZNF326, ZNF560, ZNF569, ZNF732, ZNF711, KLF8, GLI1, ZNF14, ZBTB7B, ZNF462, ZNF583, ZNF574

Table S3. Statistics of mutation sites on/aside histone modification sites in histones

Gene Name	Coding silent	Missense	Nonsense	Others	Matched
HIST1H1E	14	43	1	5	3
HIST2H2AC	36	28	3	2	14
HIST2H2AB	26	17	0	3	
HIST1H2BD	8	33	1	1	
HIST1H2BM	21	13	1	0	21
HIST1H2BO	5	23	1	0	
HIST1H3F	3	13	0	0	7
HIST1H4I	7	16	0	1	7

The naming of “Matched” is mutation counting of matched in modification site or around of them.

Table S4. Data collection

Cell line	ID	Chip-seq Type	Data Source	File type
H1	GSM915328 [1]	mRNA	GEO	BED
K562	wgEncodeEH000124 [2]	mRNA	UCSC	BAM
CD4	GSM669617 [3]	mRNA	GEO	BED
Hepg2	wgEncodeEH000127 [4]	mRNA	UCSC	BAM
Testis	SRR531456 [5]	mRNA	GEO	SRA
Ovary	SRR531458 [5]	mRNA	GEO	SRA
Mutation data				
common SNPs [2]	Download from UCSC	Cancer SNVs	Download from COSMIC [6]	
Epigenetic marks				
H1-ESC				
ID	Chip-seq Type	ID	Chip-seq Type	
GSM466739 [1]	H3K4me1	GSM605323 [1]	H3K9ac	
GSM602260 [1]	H3K4me2	GSM466732 [1]	H3K27ac	
GSM469971 [1]	H3K4me3	GSM466734 [1]	H3K27me3	
GSM433174 [1]	H3K9me3			
Hepg2				
ID	Chip-seq Type	ID	Chip-seq Type	
GSM646355 [7]	H3K27ac	GSM646357 [7]	H3K27me3	
GSM646361 [7]	H3K4me1	GSM646362 [7]	H3K4me2	
GSM646364 [7]	H3K4me3	GSM646366 [7]	H3K9ac	
wgEncodeEH003087 [4]	H3K9me3			

Table S5. Enriched GO items in cell component namespace of GO in ten gene groups

GO ID	Item	0%-10%	10%-20%	20%-30%	30%-40%	40%-50%	50%-60%	60%-70%	70%-80%	80%-90%	90%-100%
GO:0005635	Nuclear envelope	9.61E-03	5.12E-01	7.16E-01	1.48E-03	2.77E-02	2.28E-01	5.79E-01	1.67E-01	5.31E-01	7.10E-01
GO:0005637	Nuclear inner membrane	3.83E-01	3.53E-01	1.00E+00	6.36E-01	3.59E-01	1.48E-02	5.94E-01	3.16E-01	1.00E+00	1.00E+00
GO:0000785	Chromatin	1.80E-07	9.39E-02	9.90E-02	7.23E-01	8.70E-01	1.71E-01	6.97E-01	9.47E-01	6.56E-01	9.08E-01
GO:0005739	Mitochondrion	8.14E-02	4.08E-01	3.93E-02	2.33E-02	1.40E-01	1.96E-01	1.79E-03	5.23E-04	6.83E-04	4.35E-02
GO:0005783	Endoplasmic reticulum	4.08E-01	3.96E-01	2.52E-01	8.60E-02	4.96E-02	6.89E-02	2.82E-01	3.39E-02	6.06E-02	9.25E-01
GO:0005794	Golgi apparatus	4.78E-01	5.39E-02	8.93E-01	1.41E-01	7.58E-03	1.75E-01	3.08E-03	6.03E-01	7.82E-02	3.34E-01
GO:0005764	Lysosome	9.94E-01	8.95E-01	5.45E-01	1.65E-01	2.32E-01	6.77E-01	1.79E-01	6.23E-02	5.77E-03	9.40E-01
GO:0005773	Vacuole	9.11E-01	9.61E-01	5.87E-01	2.72E-02	4.43E-02	8.89E-01	2.88E-01	8.98E-02	1.08E-02	9.03E-01
GO:0005840	Ribosome	7.25E-01	2.89E-01	6.19E-01	8.79E-01	7.63E-02	2.69E-01	7.92E-02	8.19E-02	6.87E-01	7.93E-02
GO:0005737	Cytoplasm	7.59E-05	1.43E-05	3.70E-07	9.73E-07	1.14E-10	6.44E-05	1.16E-07	8.35E-04	1.01E-03	7.73E-01
GO:0005615	Extracellular space	1.00E+00	9.99E-01	9.93E-01	9.79E-01	9.48E-01	9.55E-01	9.98E-01	9.99E-01	4.06E-01	9.26E-01
GO:0005912	Adherens junction	5.22E-03	5.51E-01	3.75E-01	1.08E-01	3.23E-03	1.44E-01	1.04E-01	3.44E-01	7.58E-01	9.50E-01
GO:0005813	Centrosome	2.27E-03	2.05E-02	1.77E-01	5.22E-03	2.40E-02	3.86E-02	7.03E-02	4.92E-01	9.15E-01	9.98E-01
GO:0005856	Cytoskeleton	3.98E-03	2.71E-04	2.90E-02	7.47E-02	4.53E-04	3.42E-03	2.18E-02	7.06E-01	9.96E-01	1.00E+00
GO:0005874	Microtubule	6.93E-01	1.28E-03	5.97E-04	2.78E-01	1.57E-03	2.59E-02	4.49E-03	8.08E-01	9.76E-01	9.77E-01
GO:0009897	External side of plasma membrane	6.57E-01	4.70E-01	7.45E-01	8.74E-01	6.02E-01	4.44E-01	8.09E-01	4.89E-01	9.64E-01	5.73E-01
GO:0009898	Internal side of plasma membrane	2.85E-02	6.99E-01	2.18E-01	2.59E-01	7.10E-01	3.69E-01	7.82E-01	7.86E-01	4.41E-01	1.97E-01
GO:0015629	Actin cytoskeleton	4.33E-03	1.06E-02	2.65E-01	7.70E-01	3.29E-04	8.51E-02	1.62E-01	1.20E-01	6.03E-01	7.34E-01
GO:0015630	Microtubule cytoskeleton	5.33E-02	7.61E-03	2.85E-03	4.09E-03	8.32E-05	1.51E-02	4.90E-03	5.08E-01	8.03E-01	9.98E-01
GO:0035097	Histone methyltransferase complex	1.22E-04	3.37E-01	7.22E-01	1.85E-01	5.50E-01	7.37E-01	7.04E-01	1.00E+00	6.86E-01	1.00E+00
GO:0000123	Histone acetyltransferase complex	1.25E-04	1.81E-02	7.72E-01	9.21E-02	9.46E-02	7.88E-01	5.75E-01	7.57E-01	8.78E-01	1.00E+00
GO:0000118	Histone deacetylase complex	2.04E-02	1.56E-02	6.24E-01	6.57E-01	4.33E-01	4.11E-01	1.93E-01	8.27E-01	5.86E-01	6.71E-01
GO:0016580	Sin3 complex	8.15E-03	1.98E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.79E-01	1.00E+00	1.00E+00
GO:0016585	Chromatin remodeling complex	3.99E-06	1.37E-03	2.70E-01	5.98E-01	4.60E-01	2.94E-01	1.48E-01	7.96E-01	7.70E-01	8.26E-01
GO:0005689	U12-type spliceosomal complex	7.53E-03	5.25E-01	5.02E-01	5.28E-01	1.00E+00	5.15E-01	1.00E+00	2.18E-01	6.78E-02	1.00E+00
GO:0005697	Telomerase holoenzyme complex	8.21E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.70E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00
GO:0005721	Centromeric heterochromatin	5.50E-04	1.00E+00	1.00E+00	3.46E-02	1.00E+00	1.65E-01	1.00E+00	1.00E+00	1.46E-01	1.00E+00

Table S6. Enriched GO items in molecular function namespace of GO in ten gene groups

GO ID	Item	0%-10%	10%-20%	20%-30%	30%-40%	40%-50%	50%-60%	60%-70%	70%-80%	80%-90%	90%-100%
GO:0003735	Structural constituent of ribosome	8.49E-01	9.98E-01	7.98E-01	8.34E-01	2.77E-01	2.42E-01	4.55E-01	6.24E-01	2.21E-02	4.61E-05
GO:0019210	Kinase inhibitor activity	1.82E-01	8.92E-01	8.83E-01	5.24E-01	1.59E-01	7.01E-01	8.62E-01	4.53E-01	3.99E-02	8.52E-03
GO:0008528	G-protein coupled peptide receptor activity	9.98E-01	9.97E-01	9.49E-01	1.00E+00	9.86E-01	9.48E-01	9.77E-01	8.59E-01	6.82E-01	1.24E-01
GO:0035257	Nuclear hormone receptor binding	1.58E-03	5.63E-01	3.84E-01	1.19E-01	4.18E-01	6.71E-01	6.19E-01	6.35E-01	9.30E-01	1.46E-01

GO:0003682	Chromatin binding	2.30E-10	2.59E-05	5.38E-01	5.31E-01	1.75E-01	7.92E-01	9.99E-01	8.82E-01	8.84E-01	8.86E-01
GO:0003712	Transcription cofactor activity	3.81E-09	1.57E-03	1.64E-02	9.93E-01	7.73E-01	2.79E-01	2.43E-01	9.11E-01	6.49E-01	8.42E-01
GO:0000989	Transcription factor binding transcription factor activity	6.05E-09	1.06E-03	1.75E-02	9.96E-01	8.24E-01	2.24E-01	3.04E-01	8.67E-01	5.63E-01	8.73E-01
GO:0003713	Transcription coactivator activity	1.17E-08	1.12E-01	2.22E-02	8.96E-01	8.80E-01	6.02E-01	2.44E-01	9.76E-01	4.31E-01	7.02E-01
GO:0003700	Sequence-specific DNA binding transcription factor activity	1.18E-08	4.41E-01	7.94E-01	9.80E-01	8.56E-01	7.36E-01	9.94E-01	9.91E-01	9.69E-01	2.27E-01
GO:0003677	DNA binding	1.67E-05	7.29E-01	4.78E-01	7.28E-01	8.94E-01	9.47E-02	6.51E-01	8.33E-01	1.09E-02	6.11E-04
GO:0031625	Ubiquitin protein ligase binding	2.31E-05	5.63E-02	6.92E-01	3.73E-01	6.01E-01	4.25E-01	7.50E-01	2.66E-01	3.05E-01	8.04E-01
GO:0002039	P53 binding	4.09E-04	2.30E-01	3.99E-01	8.59E-01	8.52E-01	3.94E-01	1.00E+00	8.23E-01	7.94E-01	1.00E+00
GO:0003714	Transcription corepressor activity	2.22E-03	6.40E-02	2.90E-02	9.83E-01	7.68E-01	6.17E-01	2.22E-01	3.37E-01	7.18E-01	9.27E-01
GO:0019900	Kinase binding	1.38E-04	3.66E-02	4.03E-01	1.33E-02	1.47E-01	1.06E-01	3.63E-01	1.45E-01	5.06E-01	9.88E-01
GO:0004402	Histone acetyltransferase activity	3.61E-04	7.37E-01	3.01E-01	5.41E-01	5.25E-01	7.14E-01	6.82E-01	8.75E-01	1.00E+00	1.00E+00
GO:0035473	Lipase binding	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.52E-03
GO:0019899	Enzyme binding	9.31E-08	2.47E-02	1.20E-01	1.08E-04	4.78E-01	1.44E-02	9.07E-02	3.25E-01	8.23E-01	9.37E-01
GO:0043295	Glutathione binding	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.56E-06	1.00E+00

Table S7. Enriched GO items in biology process namespace of GO in ten gene groups

GO ID	Item	0%-10%	10%-20%	20%-30%	30%-40%	40%-50%	50%-60%	60%-70%	70%-80%	80%-90%	90%-100%
GO:0072594	Establishment of protein localization to organelle	9.98E-01	9.17E-01	5.23E-01	1.84E-01	3.55E-01	4.28E-01	1.81E-01	6.01E-01	7.82E-02	4.26E-07
GO:0045047	Protein targeting to ER	9.92E-01	6.80E-01	7.79E-01	3.82E-01	3.81E-01	6.53E-01	7.48E-01	3.21E-01	3.82E-01	1.32E-04
GO:0019083	Viral transcription	9.94E-01	9.65E-01	6.16E-01	2.06E-01	6.47E-01	7.78E-01	5.80E-01	4.26E-01	3.53E-01	2.55E-04
GO:0006415	Translational termination	9.32E-01	9.67E-01	6.28E-01	3.43E-01	6.59E-01	7.87E-01	7.51E-01	2.89E-01	3.65E-01	2.88E-04
GO:0006413	Translational initiation	9.98E-01	4.76E-01	3.04E-01	1.47E-01	3.40E-01	5.80E-01	7.86E-01	5.38E-01	4.48E-01	7.38E-03
GO:0006412	Translation	8.40E-01	5.24E-01	7.66E-01	5.12E-01	1.05E-02	7.85E-01	1.02E-01	1.55E-01	2.18E-01	7.86E-03
GO:0006355	Regulation of transcription, DNA-dependent	2.44E-10	1.08E-02	4.62E-01	7.70E-01	9.05E-01	3.93E-01	1.21E-01	4.52E-01	9.37E-03	1.53E-02
GO:0051276	Chromosome organization	2.10E-19	8.01E-03	4.97E-02	1.03E-01	1.67E-01	6.75E-01	9.78E-01	9.73E-01	1.00E+00	8.66E-01
GO:0016568	Chromatin modification	1.70E-18	2.77E-03	1.72E-02	1.15E-01	5.50E-01	7.92E-01	9.95E-01	8.97E-01	9.88E-01	9.92E-01
GO:0016570	Histone modification	4.26E-13	2.68E-02	1.75E-01	8.32E-03	4.75E-01	8.14E-01	9.40E-01	9.71E-01	9.88E-01	9.97E-01
GO:0051171	Regulation of nitrogen compound metabolic process	2.33E-12	9.43E-03	4.03E-01	5.63E-01	9.69E-01	6.26E-01	1.31E-01	4.59E-01	3.31E-02	5.37E-02
GO:0016070	RNA metabolic process	2.08E-09	7.65E-02	1.51E-01	4.31E-01	5.50E-01	4.23E-02	6.77E-01	2.66E-01	1.01E-02	8.92E-02
GO:0010629	Negative regulation of gene expression	2.96E-09	5.67E-04	8.94E-02	6.18E-01	1.71E-01	8.72E-01	7.32E-01	8.08E-01	7.27E-01	8.58E-01
GO:0010564	Regulation of cell cycle process	5.59E-06	8.80E-03	2.61E-01	6.90E-01	2.93E-02	6.79E-02	8.17E-01	7.81E-01	5.65E-01	9.06E-01
GO:0016573	Histone acetylation	6.56E-06	2.40E-01	4.89E-01	7.49E-02	2.36E-01	8.04E-01	9.58E-01	6.28E-01	7.21E-01	9.37E-01
GO:0006915	Apoptotic process	7.92E-06	3.90E-01	7.29E-01	8.11E-01	2.78E-01	6.13E-02	1.47E-01	3.94E-01	1.23E-01	2.06E-01
GO:0007049	Cell cycle	1.89E-05	3.50E-02	1.10E-02	4.51E-04	3.10E-02	5.35E-02	1.31E-01	2.64E-01	7.97E-01	9.70E-01

Table S8. HIST2H2AB and HIST2H2AC missense mutations

Gene Name	Reference ID	Sample Name	Cancer Type	Mutation	Mutation type	Location (hg19)
HIST2H2AC	**ENST00000331380**	**280**	**haematopoietic**	**p.G5D**	**Substitution - Missense**	**1:149858538-149858538**

HIST2H2AC	**ENST00000331380**	**280-01-4TD**	**haematopoietic**	**p.G5D**	**Substitution-Missense**	**1:149858538-149858538**
HIST2H2AC	ENST00000331380	TCGA-55-7283-01	lung	p.S19*	Substitution-Nonsense	1:149858580-149858580
HIST2H2AC	ENST00000331380	TCGA-AD-6895-01	large_intestine	p.S20P	Substitution-Missense	1:149858582-149858582
HIST2H2AC	ENST00000331380	TCGA-09-2044-01	ovary	p.Q25*	Substitution-Nonsense	1:149858597-149858597
HIST2H2AC	ENST00000331380	TCGA-AF-2687-01	large_intestine	p.Q25*	Substitution - Nonsense	1:149858597-149858597
HIST2H2AC	ENST00000331380	TCGA-AF-4110-01	large_intestine	p.Q25R	Substitution-Missense	1:149858598-149858598
HIST2H2AC	ENST00000331380	TCGA-CM-6172-01	large_intestine	p.Q25R	Substitution-Missense	1:149858598-149858598
HIST2H2AC	ENST00000331380	TCGA-F5-6812-01	large_intestine	p.Q25R	Substitution-Missense	1:149858598-149858598
HIST2H2AC	**ENST00000331380**	**pfg008T**	**stomach**	**p.R36C**	**Substitution-Missense**	**1:149858630-149858630**
HIST2H2AC	ENST00000331380	TCGA-DS-A10D-01	cervix	p.M52L	Substitution-Missense	1:149858678-149858678
HIST2H2AC	ENST00000331380	LUAD-RT-S01818	lung	p.E57Q	Substitution-Missense	1:149858693-149858693
HIST2H2AC	ENST00000331380	TCGA-73-4677-01	lung	p.R72L	Substitution-Missense	1:149858739-149858739
HIST2H2AC	ENST00000331380	TCGA-CM-5861-01	large_intestine	p.T77A	Substitution-Missense	1:149858753-149858753
HIST2H2AC	ENST00000331380	TCGA-20-1682-01	ovary	p.R89P	Substitution-Missense	1:149858790-149858790
HIST2H2AC	ENST00000331380	TCGA-55-7907-01	lung	p.D91Y	Substitution-Missense	1:149858795-149858795
HIST2H2AC	ENST00000331380	TCGA-60-2698-01	lung	p.D91H	Substitution-Missense	1:149858795-149858795
HIST2H2AC	ENST00000331380	TCGA-A0-A03T-01	breast	p.K96M	Substitution-Missense	1:149858811-149858811
HIST2H2AC	**ENST00000331380**	**91T**	**skin**	**p.L97P**	**Substitution-Missense**	**1:149858814-149858814**
HIST2H2AC	ENST00000331380	MD-034	central	p.T102S	Substitution-Missense	1:149858829-149858829
HIST2H2AC	ENST00000331380	TCGA-73-4666-01	lung	p.Q105E	Substitution-Missense	1:149858837-149858837
HIST2H2AC	ENST00000331380	TCGA-49-4487-01	lung	p.Q105H	Substitution-Missense	1:149858839-149858839
HIST2H2AC	ENST00000331380	ESO-0125	oesophagus	p.P110L	Substitution-Missense	1:149858853-149858853
HIST2H2AC	ENST00000331380	LC_S35	lung	p.N111S	Substitution-Missense	1:149858856-149858856
HIST2H2AC	ENST00000331380	TCGA-BG-A18B-01	endometrium	p.L116fs* > 14	Deletion-Frameshift	1:149858869-149858870
HIST2H2AC	**ENST00000331380**	**LC_S25**	**lung**	**p.K119E**	**Substitution-Missense**	**1:149858879-149858879**
HIST2H2AC	**ENST00000331380**	**LC_S35**	**lung**	**p.K119E**	**Substitution-Missense**	**1:149858879-149858879**
HIST2H2AC	**ENST00000331380**	**SWE-33**	**prostate**	**p.K119N**	**Substitution-Missense**	**1:149858881-149858881**
HIST2H2AC	**ENST00000331380**	**TCGA-50-5072-01**	**lung**	**p.T121I**	**Substitution-Missense**	**1:149858886-149858886**
HIST2H2AC	**ENST00000331380**	**TCGA-66-2763-01**	**lung**	**p.T121N**	**Substitution-Missense**	**1:149858886-149858886**
HIST2H2AC	**ENST00000331380**	**8013222**	**pancreas**	**p.K125E**	**Substitution-Missense**	**1:149858897-149858897**
HIST2H2AC	**ENST00000331380**	**ICGC_0002**	**pancreas**	**p.K125E**	**Substitution-Missense**	**1:149858897-149858897**
HIST2H2AC	ENST00000331380	MN-1025	meninges	p.K127fs* > 6	Insertion-Frameshift	1:149858902-149858903
HIST2H2AB	**ENST00000331128**	**RK126_C01**	**liver**	**p.K125R**	**Substitution-Missense**	**1:149859093-149859093**
HIST2H2AB	**ENST00000331128**	**TCGA-EI-6507-01**	**large_intestine**	**p.H124fs* > 7**	**Deletion-Frameshift**	**1:149859099-149859100**
HIST2H2AB	**ENST00000331128**	**TCGA-BG-A0MI-01**	**endometrium**	**p.H124fs* > 7**	**Deletion-Frameshift**	**1:149859103-149859104**
HIST2H2AB	ENST00000331128	LC_S25	lung	p.L116P	Substitution-Missense	1:149859120-149859120

HIST2H2AB	ENST00000331128	587268	large_intestine	p.V115I	Substitution-Missense	1:149859124-149859124
HIST2H2AB	ENST00000331128	TCGA-DI-A0WH-01	endometrium	p.G107S	Substitution-Missense	1:149859148-149859148
HIST2H2AB	ENST00000331128	LUAD-CHTN-MAD06-00668	lung	p.Q105H	Substitution-Missense	1:149859152-149859152
HIST2H2AB	ENST00000331128	TCGA-G4-6317-01	large_intestine	p.V101A	Substitution-Missense	1:149859165-149859165
HIST2H2AB	ENST00000331128	TCGA-CM-5860-01	large_intestine	p.V101I	Substitution-Missense	1:149859166-149859166
HIST2H2AB	ENST00000331128	TCGA-DR-A0ZM-01	cervix	p.R82G	Substitution-Missense	1:149859223-149859223
HIST2H2AB	ENST00000331128	TCGA-CM-4746-01	large_intestine	p.N74delN	Deletion-In frame	1:149859243-149859245
HIST2H2AB	ENST00000331128	LUAD-NYU408	lung	p.N69I	Substitution-Missense	1:149859261-149859261
HIST2H2AB	ENST00000331128	TCGA-50-6595-01	lung	p.T60S	Substitution-Missense	1:149859288-149859288
HIST2H2AB	ENST00000331128	TCGA-B4-5832-01	kidney	p.E57Q	Substitution-Missense	1:149859298-149859298
HIST2H2AB	ENST00000331128	TCGA-DS-A10C-01	cervix	p.L52M	Substitution-Missense	1:149859313-149859313
HIST2H2AB	ENST00000331128	TCGA-DS-A10D-01	cervix	p.L52M	Substitution-Missense	1:149859313-149859313
HIST2H2AB	ENST00000331128	TCGA-39-5035-01	lung	p.A48G	Substitution-Missense	1:149859324-149859324
HIST2H2AB	ENST00000331128	HN_63080	upper_aerodigestive_tract	p.G45R	Substitution-Missense	1:149859334-149859334
HIST2H2AB	ENST00000331128	TCGA-AA-3715-01	large_intestine	p.R36H	Substitution-Missense	1:149859360-149859360
HIST2H2AB	ENST00000331128	TCGA-66-2778-01	lung	p.R33G	Substitution-Missense	1:149859370-149859370

The marked by bold mutation site are located on around of known modification site in HIST2H2AB and HIST2H2AC.

Table S9. HIST1H1E missense mutations

Gene Name	Reference ID	Sample Name	Cancer Type	Primary site	Mutation	Mutation type	Location (hg19)
HIST1H1E	**ENST00000304218**	**TCGA-BT-A20R-01**	**urinary_tract**	**bladder**	**p.M1L**	**Substitution-Missense**	**6:26156619-26156619**
HIST1H1E	ENST00000304218	PD4192a	NS	NS	p.P14R	Substitution-Missense	6:26156659-26156659
HIST1H1E	ENST00000304218	TCGA-A6-5665-01	large_intestine	colon	p.K23delK	Deletion-In frame	6:26156678-26156680
HIST1H1E	ENST00000304218	TCGA-AZ-6598-01	large_intestine	caecum	p.K23delK	Deletion-In frame	6:26156678-26156680
HIST1H1E	ENST00000304218	TCGA-CM-4746-01	large_intestine	colon	p.K23delK	Deletion-In frame	6:26156678-26156680
HIST1H1E	**ENST00000304218**	**TCGA-35-5375-01**	**lung**	**NS**	**p.R25P**	**Substitution-Missense**	**6:26156692-26156692**
HIST1H1E	**ENST00000304218**	**TCGA-55-5899-01**	**lung**	**NS**	**p.R25P**	**Substitution-Missense**	**6:26156692-26156692**
HIST1H1E	ENST00000304218	HN_62672	aerodigestive_tract	pharynx	p.G29D	Substitution-Missense	6:26156704-26156704
HIST1H1E	ENST00000304218	tumor_4163639	haematopoietic	NS	p.E42D	Substitution-Missense	6:26156744-26156744
HIST1H1E	ENST00000304218	TCGA-B5-A0JZ-01	endometrium	NS	p.K46E	Substitution-Missense	6:26156754-26156754
HIST1H1E	ENST00000304218	DLBCL705	haematopoietic	NS	p.A47V	Substitution-Missense	6:26156758-26156758
HIST1H1E	ENST00000304218	WA35	prostate	NS	p.S51F	Substitution-Missense	6:26156770-26156770
HIST1H1E	ENST00000304218	TCGA-39-5035-01	lung	NS	p.R54C	Substitution-Missense	6:26156778-26156778
HIST1H1E	ENST00000304218	TCGA-24-1471-01	ovary	NS	p.S58P	Substitution-Missense	6:26156790-26156790

HIST1H1E	ENST00000304218	TCGA-CM-6169-01	large_intestine	caecum	p.S58P	Substitution-Missense	6:26156790-26156790
HIST1H1E	ENST00000304218	DLBCL922	haematopoietic	NS	p.S58F	Substitution-Missense	6:26156791-26156791
HIST1H1E	ENST00000304218	tumor_4163639	haematopoietic	NS	p.S58F	Substitution-Missense	6:26156791-26156791
HIST1H1E	ENST00000304218	TCGA-09-1674-01	ovary	NS	p.A61D	Substitution-Missense	6:26156800-26156800
HIST1H1E	ENST00000304218	TCGA-BH-A0BP-01	breast	NS	p.A61V	Substitution-Missense	6:26156800-26156800
HIST1H1E	ENST00000304218	91	haematopoietic	NS	p.A65P	Substitution-Missense	6:26156811-26156811
HIST1H1E	ENST00000304218	091-01-6TD	haematopoietic	NS	p.A65P	Substitution-Missense	6:26156811-26156811
HIST1H1E	ENST00000304218	TCGA-GV-A3JV-01	urinary_tract	bladder	p.Y71C	Substitution-Missense	6:26156830-26156830
HIST1H1E	ENST00000304218	CLL052	haematopoietic	NS	p.D72_V73delDV	Deletion-In frame	6:26156832-26156837
HIST1H1E	ENST00000304218	9534	salivary_gland	NS	p.K75N	Substitution-Missense	6:26156843-26156843
HIST1H1E	ENST00000304218	CLL107	haematopoietic	NS	p.R79H	Substitution-Missense	6:26156854-26156854
HIST1H1E	ENST00000304218	HCC06T	liver	NS	p.R79L	Substitution-Missense	6:26156854-26156854
HIST1H1E	ENST00000304218	ESO-0013	oesophagus	NS	p.V88L	Substitution-Missense	6:26156880-26156880
HIST1H1E	ENST00000304218	587338	large_intestine	colon	p.G91C	Substitution-Missense	6:26156889-26156889
HIST1H1E	ENST00000304218	TCGA-BS-A0TG-01	endometrium	NS	p.V94L	Substitution-Missense	6:26156898-26156898
HIST1H1E	ENST00000304218	TCGA-AX-A0IW-01	endometrium	NS	p.T96S	Substitution-Missense	6:26156904-26156904
HIST1H1E	ENST00000304218	DLBCL899	haematopoietic	NS	p.N108H	Substitution-Missense	6:26156940-26156940
HIST1H1E	ENST00000304218	DLBCL899	haematopoietic	NS	p.K109N	Substitution-Missense	6:26156945-26156945
HIST1H1E	ENST00000304218	TCGA-BL-A3JM-01	urinary_tract	bladder	p.K109N	Substitution-Missense	6:26156945-26156945
HIST1H1E	ENST00000304218	585260	lung	NS	p.A145S	Substitution-Missense	6:26157051-26157051
HIST1H1E	**ENST00000304218**	**ME024T**	**NS**	**NS**	**p.P147L**	**Substitution-Missense**	**6:26157058-26157058**
HIST1H1E	ENST00000304218	ME048T	skin	extremity	p.K153N	Substitution-Missense	6:26157077-26157077
HIST1H1E	ENST00000304218	TCGA-64-5775-01	lung	NS	p.K156R	Substitution-Missense	6:26157085-26157085
HIST1H1E	ENST00000304218	TCGA-AP-A0LM-01	endometrium	NS	p.A158T	Substitution-Missense	6:26157090-26157090
HIST1H1E	ENST00000304218	90983	haematopoietic	NS	p.A164P	Substitution-Missense	6:26157108-26157108
HIST1H1E	ENST00000304218	LPJ023	haematopoietic	NS	p.A164V	Substitution-Missense	6:26157109-26157109
HIST1H1E	ENST00000304218	tumor_4194218	haematopoietic	NS	p.A164G	Substitution-Missense	6:26157109-26157109
HIST1H1E	ENST00000304218	CLL129	haematopoietic	NS	p.A167V	Substitution-Missense	6:26157118-26157118
HIST1H1E	ENST00000304218	TCGA-22-5492-01	lung	NS	p.K177*	Substitution-Nonsense	6:26157147-26157147
HIST1H1E	ENST00000304218	TCGA-CM-5861-01	large_intestine	caecum	p.K190R	Substitution-Missense	6:26157187-26157187
HIST1H1E	ENST00000304218	CLL106	haematopoietic	NS	p.P196S	Substitution-Missense	6:26157204-26157204
HIST1H1E	ENST00000304218	TCGA-CA-6717-01	large_intestine	colon	p.P196A	Substitution-Missense	6:26157204-26157204
HIST1H1E	ENST00000304218	CLL058	haematopoietic	NS	p.K202E	Substitution-Missense	6:26157222-26157222
HIST1H1E	ENST00000304218	TCGA-D1-A15X-01	endometrium	NS	p.A204T	Substitution-Missense	6:26157228-26157228
HIST1H1E	ENST00000304218	587284	large_intestine	colon	p.*220Q	Nonstop extension	6:26157276-26157276

The marked by bold mutation site are located on around of known modification site in HIST1H1E.

Table S10. HIST1H4I missense mutations

Gene Name	Reference ID	Sample Name	Cancer Type	Primary site	Mutation	Mutation type	Location (hg19)
HIST1H4I	**ENST00000354348**	**TCGA-37-4133-01**	**lung**	**NS**	**p.G12E**	**Substitution-Missense**	**6:27107122-27107122**
HIST1H4I	ENST00000354348	DLBCL701	haematopoietic	NS	p.D69fs* > 36	Insertion-Frameshift	6:27107292-27107293
HIST1H4I	**ENST00000354348**	587256	**large_intestine**	**colon**	**p.A16V**	**Substitution-Missense**	**6:27107134-27107134**
HIST1H4I	**ENST00000354348**	**TCGA-18-3406-01**	**lung**	**NS**	**p.R18L**	**Substitution-Missense**	**6:27107140-27107140**
HIST1H4I	ENST00000354348	TCGA-69-7979-01	lung	right_upper_lobe	p.R41L	Substitution-Missense	6:27107209-27107209
HIST1H4I	ENST00000354348	587336	large_intestine	colon	p.K45Q	Substitution-Missense	6:27107220-27107220
HIST1H4I	**ENST00000354348**	**TCGA-GV-A3JZ-01**	**urinary_tract**	**bladder**	**p.I51M**	**Substitution-Missense**	**6:27107240-27107240**
HIST1H4I	ENST00000354348	Toledo	haematopoietic	**NS**	**p.I51M**	**Substitution-Missense**	**6:27107240-27107240**
HIST1H4I	**ENST00000354348**	**TCGA-BT-A20J-01**	**urinary_tract**	**bladder**	**p.E53Q**	**Substitution-Missense**	**6:27107244-27107244**
HIST1H4I	ENST00000354348	TCGA-DM-A1DB-01	large_intestine	colon	p.R56C	Substitution-Missense	6:27107253-27107253
HIST1H4I	ENST00000354348	TCGA-D1-A0Z0-01	endometrium	NS	p.V58L	Substitution-Missense	6:27107259-27107259
HIST1H4I	ENST00000354348	TCGA-44-6777-01	lung	NS	p.E64K	Substitution-Missense	6:27107277-27107277
HIST1H4I	ENST00000354348	TCGA-AZ-6600-01	large_intestine	colon	p.H76Q	Substitution-Missense	6:27107315-27107315
HIST1H4I	ENST00000354348	TCGA-AX-A060-01	endometrium	NS	p.V87M	Substitution-Missense	6:27107346-27107346
HIST1H4I	**ENST00000354348**	**TCGA-AH-6643-01**	**large_intestine**	**rectum**	**p.A90T**	**Substitution-Missense**	**6:27107355-27107355**
HIST1H4I	ENST00000354348	TCGA-73-4668-01	lung	NS	p.G100C	Substitution-Missense	6:27107385-27107385
HIST1H4I	ENST00000354348	TCGA-G4-6586-01	large_intestine	colon	p.G102D	Substitution-Missense	6:27107392-27107392

The marked by bold mutation site are located on around of known modification site in HIST1H4I.

Table S11. HIST1H3F missense mutations

Gene Name	Reference ID	Sample Name	Cancer Type	Primary site	Mutation	Mutation type	Location (hg19)
HIST1H3F	**ENST00000446824**	**TCGA-55-7570-01**	**lung**	**lobe**	**p.T23P**	**Substitution-Missense**	**6:26250767-26250767**
HIST1H3F	**ENST00000446824**	**pfg006T**	**stomach**	**NS**	**p.R27C**	**Substitution-Missense**	**6:26250755-26250755**
HIST1H3F	**ENST00000446824**	**TCGA-50-5944-01**	**lung**	**NS**	**p.P39T**	**Substitution-Missense**	**6:26250719-26250719**
HIST1H3F	ENST00000446824	TCGA-BL-A0C8-01	urinary_tract	bladder	p.A48T	Substitution-Missense	6:26250692-26250692
HIST1H3F	**ENST00000446824**	**TCGA-55-8089-01**	**lung**	**lobe**	**p.Q56E**	**Substitution-Missense**	**6:26250668-26250668**
HIST1H3F	**ENST00000446824**	**LUAD-S01357**	**lung**	**NS**	**p.Q56H**	**Substitution-Missense**	**6:26250666-26250666**
HIST1H3F	**ENST00000446824**	**TCGA-E9-A247-01**	**breast**	**NS**	**p.S58L**	**Substitution-Missense**	**6:26250661-26250661**
HIST1H3F	**ENST00000446824**	**585260**	**lung**	**NS**	**p.K65N**	**Substitution-Missense**	**6:26250639-26250639**
HIST1H3F	ENST00000446824	TCGA-AR-A0TX-01	breast	NS	p.V72A	Substitution-Missense	6:26250619-26250619
HIST1H3F	ENST00000446824	TCGA-CK-5916-01	intestine	caecum	p.V72A	Substitution-Missense	6:26250619-26250619
HIST1H3F	**ENST00000446824**	**TCGA-60-2698-01**	**lung**	**NS**	**p.F79L**	**Substitution-Missense**	**6:26250597-26250597**

HIST1H3F	ENST00000446824	TCGA-18-5595-01	lung	NS	p.E98K	Substitution-Missense	6:26250542-26250542
HIST1H3F	ENST00000446824	TCGA-05-4410-01	lung	NS	p.Y100F	Substitution-Missense	6:26250535-26250535

The marked by bold mutation site are located on around of known modification site in HIST1H3F.

Table S12. HIST1H2BD/M/O missense mutations

Gene Name	Reference ID	Sample Name	Cancer Type	Mutation	Mutation type	Location (hg19)
HIST1H2BD	**ENST00000377777**	**TCGA-DK-A1AC-01**	**urinary_tract**	**p.E3K**	**Substitution-Missense**	**6:26158404-26158404**
HIST1H2BD	**ENST00000377777**	**TCGA-20-1684-01**	**ovary**	**p.E3Q**	**Substitution-Missense**	**6:26158404-26158404**
HIST1H2BD	ENST00000377777	ESO14T	oesophagus	p.P4R	Substitution-Missense	6:26158408-26158408
HIST1H2BD	**ENST00000377777**	**TCGA-A5-A0GB-01**	**endometrium**	**p.S7Y**	**Substitution-Missense**	**6:26158417-26158417**
HIST1H2BD	ENST00000377777	PD3988a	breast	p.P9L	Substitution-Missense	6:26158423-26158423
HIST1H2BD	ENST00000377777	LUAD-E00934	lung	p.A10D	Substitution-Missense	6:26158426-26158426
HIST1H2BD	ENST00000377777	TCGA-D1-A0ZO-01	endometrium	p.A10V	Substitution-Missense	6:26158426-26158426
HIST1H2BD	**ENST00000377777**	**TCGA-13-0899-01**	**ovary**	**p.A22G**	**Substitution-Missense**	**6:26158462-26158462**
HIST1H2BD	ENST00000377777	TCGA-AP-A054-01	endometrium	p.K31E	Substitution-Missense	6:26158488-26158488
HIST1H2BD	**ENST00000377777**	**TCGA-AP-A0LT-01**	**endometrium**	**p.K35R**	**Substitution-Missense**	**6:26158501-26158501**
HIST1H2BD	ENST00000377777	DU-145	prostate	p.S39*	Substitution-Nonsense	6:26158513-26158513
HIST1H2BD	ENST00000377777	24	pancreas	p.S39L	Substitution-Missense	6:26158513-26158513
HIST1H2BD	ENST00000377777	TCGA-AA-A01R-01	large_intestine	p.V40M	Substitution-Missense	6:26158515-26158515
HIST1H2BD	ENST00000377777	TCGA-D1-A17D-01	endometrium	p.Y41C	Substitution-Missense	6:26158519-26158519
HIST1H2BD	**ENST00000377777**	**TCGA-CM-5861-01**	**large_intestine**	**p.K44N**	**Substitution-Missense**	**6:26158529-26158529**
HIST1H2BD	**ENST00000377777**	**DLBCL-PatientB**	**haematopoietic**	**p.Q48E**	**Substitution-Missense**	**6:26158539-26158539**
HIST1H2BD	ENST00000377777	TCGA-78-7542-01	lung	p.V49F	Substitution-Missense	6:26158542-26158542
HIST1H2BD	ENST00000377777	LAU149	skin	p.G54S	Substitution-Missense	6:26158557-26158557
HIST1H2BD	ENST00000377777	TCGA-GD-A3OQ-01	urinary_tract	p.I55M	Substitution-Missense	6:26158562-26158562
HIST1H2BD	ENST00000377777	TCGA-AA-3811-01	large_intestine	p.I62N	Substitution-Missense	6:26158582-26158582
HIST1H2BD	ENST00000377777	TARGET-30-PARKNP	autonomic_ganglia	p.V67fs* > 62	Insertion-Frameshift	6:26158587-26158588
HIST1H2BD	ENST00000377777	PD4120a	breast	p.E77K	Substitution-Missense	6:26158626-26158626
HIST1H2BD	ENST00000377777	TCGA-DK-A1AC-01	urinary_tract	p.E77K	Substitution-Missense	6:26158626-26158626
HIST1H2BD	ENST00000377777	TCGA-DR-A0ZM-01	cervix	p.E77K	Substitution-Missense	6:26158626-26158626
HIST1H2BD	ENST00000377777	TCGA-66-2773-01	lung	p.E77Q	Substitution-Missense	6:26158626-26158626
HIST1H2BD	ENST00000377777	RK079_C01	liver	p.R80C	Substitution-Missense	6:26158635-26158635
HIST1H2BD	**ENST00000377777**	**LC_C34**	**lung**	**p.R87G**	**Substitution-Missense**	**6:26158656-26158656**
HIST1H2BD	ENST00000377777	HCC72T	liver	p.S88W	Substitution - Missense	6:26158660-26158660

HIST1H2BD	ENST00000377777	TCGA-B5-A11U-01	endometrium	p.E94D	Substitution - Missense	6:26158679-26158679
HIST1H2BD	ENST00000377777	TCGA-D1-A0ZV-01	endometrium	p.I95V	Substitution - Missense	6:26158680-26158680
HIST1H2BD	ENST00000377777	T80	endometrium	p.L103F	Substitution - Missense	6:26158704-26158704
HIST1H2BD	**ENST00000377777**	**TCGA-CM-4746-01**	**large_intestine**	**p.Y122H**	**Substitution - Missense**	**6:26158761-26158761**
HIST1H2BD	ENST00000377777	TCGA-BG-A0M4-01	endometrium	p.T123A	Substitution - Missense	6:26158764-26158764
HIST1H2BD	ENST00000377777	DLBCL-PatientK	haematopoietic	p.S124N	Substitution - Missense	6:26158768-26158768
HIST1H2BD	ENST00000377777	TCGA-22-1002-01	lung	p.K126N	Substitution - Missense	6:26158775-26158775
HIST1H2BM	**ENST00000359465**	**DLBCL-PatientM**	**haematopoietic**	**p.E3Q**	**Substitution - Missense**	**6:27782828-27782828**
HIST1H2BM	ENST00000359465	PD4003a	breast	p.Q23*	Substitution - Nonsense	6:27782888-27782888
HIST1H2BM	**ENST00000359465**	**LUAD-CHTN-MAD06-00668**	**lung**	**p.R34H**	**Substitution - Missense**	**6:27782922-27782922**
HIST1H2BM	**ENST00000359465**	**TCGA-B5-A0JR-01**	**endometrium**	**p.Y43C**	**Substitution - Missense**	**6:27782949-27782949**
HIST1H2BM	ENST00000359465	ME020T	NS	p.H50Y	Substitution - Missense	6:27782969-27782969
HIST1H2BM	ENST00000359465	TCGA-91-6848-01	lung	p.I62V	Substitution - Missense	6:27783005-27783005
HIST1H2BM	ENST00000359465	TCGA-AG-A002-01	large_intestine	p.N64H	Substitution - Missense	6:27783011-27783011
HIST1H2BM	ENST00000359465	TCGA-18-3419-01	lung	p.S65C	Substitution - Missense	6:27783015-27783015
HIST1H2BM	ENST00000359465	TCGA-D1-A163-01	endometrium	p.V67I	Substitution - Missense	6:27783020-27783020
HIST1H2BM	ENST00000359465	TCGA-EI-6507-01	large_intestine	p.R80C	Substitution - Missense	6:27783059-27783059
HIST1H2BM	ENST00000359465	TCGA-D1-A0ZS-01	endometrium	p.A82V	Substitution - Missense	6:27783066-27783066
HIST1H2BM	**ENST00000359465**	**TCGA-AA-3516-01**	**large_intestine**	**p.R87C**	**Substitution - Missense**	**6:27783080-27783080**
HIST1H2BM	ENST00000359465	TCGA-05-4396-01	lung	p.R93T	Substitution - Missense	6:27783099-27783099
HIST1H2BM	ENST00000359465	TCGA-61-1722-01	ovary	p.S124I	Substitution - Missense	6:27783192-27783192
HIST1H2BO	**ENST00000303806**	**TCGA-22-5471-01**	**lung**	**p.M1I**	**Substitution - Missense**	**6:27861243-27861243**
HIST1H2BO	**ENST00000303806**	**tumor_4163639**	**haematopoietic**	**p.A18V**	**Substitution - Missense**	**6:27861293-27861293**
HIST1H2BO	ENST00000303806	DLBCL-PatientC	haematopoietic	p.V19L	Substitution - Missense	6:27861295-27861295
HIST1H2BO	**ENST00000303806**	**TCGA-B0-5113-01**	**kidney**	**p.A22V**	**Substitution - Missense**	**6:27861305-27861305**
HIST1H2BO	**ENST00000303806**	**TCGA-24-1469-01**	**ovary**	**p.E36K**	**Substitution - Missense**	**6:27861346-27861346**
HIST1H2BO	**ENST00000303806**	**LUAD-S00488**	**lung**	**p.S37R**	**Substitution - Missense**	**6:27861351-27861351**
HIST1H2BO	ENST00000303806	110	haematopoietic	p.I40M	Substitution - Missense	6:27861360-27861360
HIST1H2BO	ENST00000303806	110-0218-04TD	haematopoietic	p.I40M	Substitution - Missense	6:27861360-27861360
HIST1H2BO	ENST00000303806	01-19969	haematopoietic	p.Y41D	Substitution - Missense	6:27861361-27861361
HIST1H2BO	**ENST00000303806**	**TCGA-36-2534-01**	**ovary**	**p.Y43***	**Substitution - Nonsense**	**6:27861369-27861369**
HIST1H2BO	ENST00000303806	TCGA-22-5471-01	lung	p.G54C	Substitution - Missense	6:27861400-27861400
HIST1H2BO	ENST00000303806	ESO-0280	oesophagus	p.G54R	Substitution - Missense	6:27861400-27861400
HIST1H2BO	ENST00000303806	TCGA-D1-A17D-01	endometrium	p.A59V	Substitution - Missense	6:27861416-27861416

HIST1H2BO	ENST00000303806	587376	large_intestine	p.M63I	Substitution - Missense	6:27861429-27861429
HIST1H2BO	ENST00000303806	TCGA-AA-3672-01	large_intestine	p.F66V	Substitution - Missense	6:27861436-27861436
HIST1H2BO	ENST00000303806	LUAD-YINHD	lung	p.I70M	Substitution - Missense	6:27861450-27861450
HIST1H2BO	ENST00000303806	Patient1_Tu	haematopoietic	p.R73C	Substitution - Missense	6:27861457-27861457
HIST1H2BO	ENST00000303806	TCGA-66-2759-01	lung	p.E77K	Substitution - Missense	6:27861469-27861469
HIST1H2BO	ENST00000303806	cSCCP4	skin	p.E77K	Substitution - Missense	6:27861469-27861469
HIST1H2BO	ENST00000303806	TCGA-91-7771-01	lung	p.T91N	Substitution - Missense	6:27861512-27861512
HIST1H2BO	ENST00000303806	LUAD-D02185	lung	p.S92F	Substitution - Missense	6:27861515-27861515
HIST1H2BO	**ENST00000303806**	**TCGA-AP-A0LF-01**	**endometrium**	**p.E114Q**	**Substitution - Missense**	**6:27861580-27861580**
HIST1H2BO	**ENST00000303806**	**ESO-147**	**oesophagus**	**p.G115S**	**Substitution - Missense**	**6:27861583-27861583**
HIST1H2BO	ENST00000303806	07-32561	haematopoietic	p.S124T	Substitution - Missense	6:27861611-27861611

Notice: The mutation sites marked by bold are located on/around (+/-) of known epigenetic modification sites in HIST1H2BD, HIST1H2BM and HIST1H2BO.

References

[1] Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, Downes M, Yu R, Stewart R, Ren B, Thomson JA, Evans RM and Ecker JR. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 2011; 471: 68-73.

[2] ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrímsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447: 799-816.

[3] Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS and Thomson JA. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 2010; 28: 1045-1048.

[4] Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, Suh BB, Hinrichs AS, Clawson H, Zweig AS, Kirkup V, Fujita PA, Rhead B, Smith KE, Pohl A, Kuhn RM, Karolchik D, Haussler D and Kent WJ. ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res 2011; 39: D871-875.

[5] Gkountela S, Li Z, Vincent JJ, Zhang KX, Chen A, Pellegrini M and Clark AT. The ontogeny of cKIT+ human primordial germ cells proves to be a resource for human germ line reprogramming, imprint erasure and in vitro differentiation. Nat Cell Biol 2013; 15: 113-122.

[6] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR and Futreal PA. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 2011; 39: D945-950.

[7] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M and Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011; 473: 43-49.