New insights on human essential genes based on integrated multi-omics analysis

Hebing Chen[1,2], Zhuo Zhang[1,2], Shuai Jiang [1,2], Ruijiang Li[1], Wanying Li[1], Hao Li[1,*] and Xiaochen Bo[1,*]

[1]Beijing Institute of Radiation Medicine, Beijing 100850, China.

[2] Co-first author

*Correspondence: lihao527_thu@foxmail.com; boxc@bmi.ac.cn

Abstract

Essential genes are those whose functions govern critical processes that sustain life in the organism. Comprehensive understanding of human essential genes could enable breakthroughs in biology and medicine. Recently, there has been a rapid proliferation of technologies for identifying and investigating the functions of human essential genes. Here, according to gene essentiality, we present a global analysis for comprehensively and systematically elucidating the genetic and regulatory characteristics of human essential genes. We explain why these genes are essential from the genomic, epigenomic, and proteomic perspectives, and we discuss their evolutionary and embryonic developmental properties. Importantly, we find that essential human genes can be used as markers to guide cancer treatment. We have developed an interactive web server, the Human Essential Genes Interactive Analysis Platform (HEGIAP) (http://sysomics.com/HEGIAP/), which integrates abundant analytical tools to give a global, multidimensional interpretation of gene essentiality.

Essential genes are indispensable for organism survival and maintaining basic functions of cells or tissues(Koonin 2000; Koonin 2003; Gerdes et al. 2006). Systematic identification of essential genes in different organisms(Luo et al. 2014) has provided critical insights into the molecular bases of many biological processes(Giaever et al. 2002). Such information may be useful for applications in areas such as synthetic biology(Lartigue et al. 2007) and drug target identification(Galperin and Koonin 1999; Hu et al. 2007). The identification of human essential genes is a particularly attractive area of research because of the potential for medical applications(Liao and Zhang 2008; Chen et al. 2012a). Utilizing gene-editing technologies based on CRISPR-Cas9 and retroviral gene-trap screens, three independent genome-wide studies(Blomen et al. 2015; Hart et al. 2015; Wang et al. 2015) identified sets of essential genes that are indispensable for human cell viability. The results agreed very well among the three studies, confirming the robustness of the evaluating approaches. All of the studies(Blomen et al. 2015; Hart et al. 2015; Wang et al. 2015) found that ~10% of the ~20,000 genes in human cells are essential for cell survival, highlighting the fact that eukaryotic genomes have intrinsic mechanisms for buffering against genetic and environmental insults(Hartman et al. 2001).

In a recent review, Norman Pavelka and colleagues(Rancati et al. 2017) stated that gene essentiality is not a fixed property but instead strongly depends on the environmental and genetic contexts and can be altered during short- or long-term evolution. However, this paradigm leaves some confusion, as we do not know how essential genes are interconnected within the cell, why these genes are essential, or what their underlying mechanisms may be. Furthermore, we do not know whether these genes are associated with disease (e.g., cancer) or have the potential to be

exploited as targets for therapeutic strategies. With the recent and rapid development of next-generation sequencing and other experimental technologies, we now have access to myriad data on genomic sequences, epigenetic modifications, structures, and disease-related information. These data will enable researchers to examine human essential genes from multiple perspectives.

In light of this background, we performed a comprehensive study of the set of human essential genes, including their genomic, epigenetic, proteomic, evolutionary, and embryonic patterning characteristics. Genetic and regulatory characteristics were studied to understand what makes these genes essential for human cell viability. We analyzed the evolutionary status of human essential genes and their profiles during embryonic development. Our findings suggest that human essential genes could be used as tumor markers in cancer. Essential genes have important implications for drug discovery, which may inform the next generation of cancer therapeutics (Fig. 1). Finally, we have developed a new web server, the Human Essential Genes Interactive Analysis Platform (HEGIAP), to facilitate the global research community in the comprehensive exploration of human essential genes.

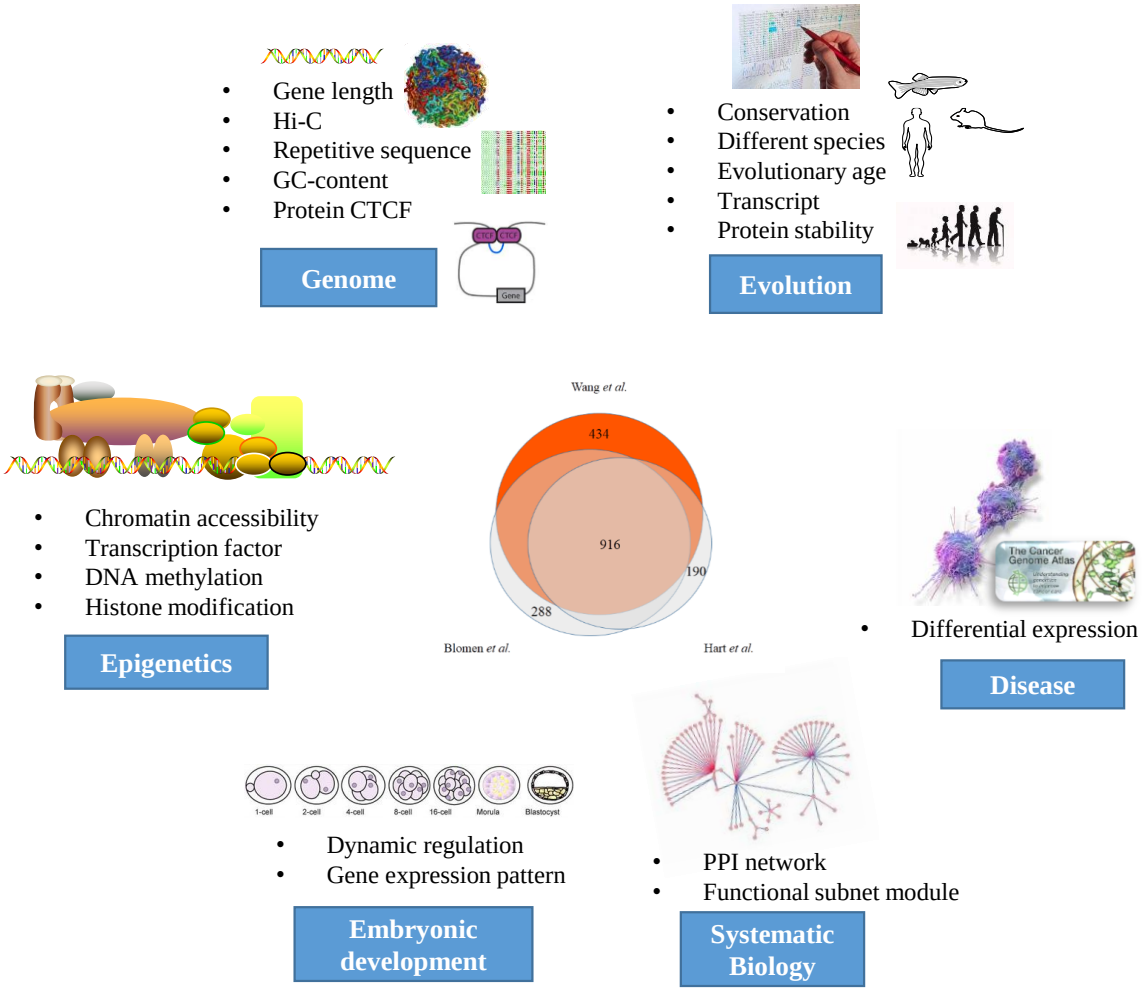


Figure 1 Comprehensive overview of analysis of human essential genes

Results

Multilevel essentiality of human essential genes.

Essential genes define the central biological functions that are required for cell growth, proliferation, and survival. To characterize the nature of human essential genes, we compared essential gene lists generated by different experimental methods from three independent genome-

wide studies(Blomen et al. 2015; Hart et al. 2015; Wang et al. 2015). More than 60% of essential genes were cataloged by each two lists (Venn diagram in Fig. 1). We focused on the essential gene list from Wang et al(Wang et al. 2015). These authors used a CRISPR score (CS) to assess the differential essentiality of human protein-coding genes, where low CS indicates high essentiality. The 18,166 protein-coding genes were divided into groups CS0 to CS9, with ascending CS values. The first group (CS0) had the lowest CS values and was defined as the set of human essential genes. This detailed classification scheme provided an accurate representation of the various features for subsequent analyses.

Protein essentiality: expression, stability, and network hubs. We analyzed the expression levels of human genes using data for 2,916 individuals from the GTEx program(Consortium 2013). Essential genes were expressed at extremely high levels, whereas nonessential genes were expressed at lower levels that decreased with decreasing level of essentiality (Fig. 2A). Proteins encoded by essential genes were the most stable of all proteins (Supplementary Fig. 1A). Our results are consistent with those of a recent study(Leuenberger et al. 2017), which found that highly expressed proteins are stable because they are designed to tolerate translational errors that would lead to accumulation of toxic misfolded species. This rationale explains that proteins encoded by human essential genes play an important role in evolution because of their more destructive features.

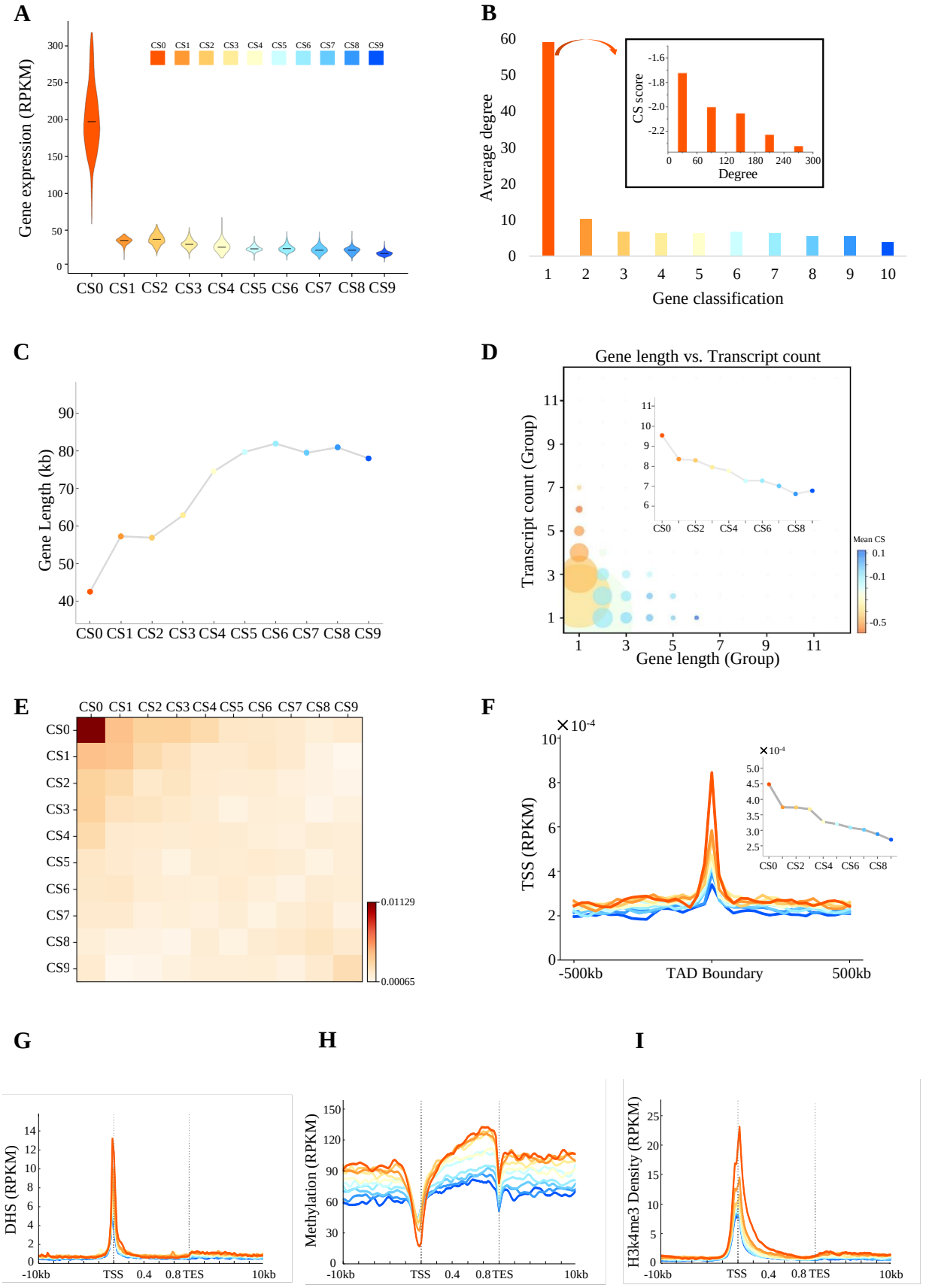

A
Gene expression (RPKM)
CS0 CS1 CS2 CS3 CS4 CS5 CS6 CS7 CS8 CS9
B
Average degree
Gene classification
CS score
Degree
C
Gene Length (kb)
D
Gene length vs. Transcript count
Transcript count (Group)
Gene length (Group)
Mean CS
E
F
TSS (RPKM)
TAD Boundary
-500kb
500kb
G
DHS (RPKM)
H
Methylation (RPKM)
I
H3k4me3 Density (RPKM)
-10kb
TSS
0.4
0.8 TES
10kb

Figure 2 General properties of human essential genes

(A) Violin plot of gene expression data of 2916 individuals from GTEx program. Mean expression level was calculated for each group of genes (CS0-CS9).

(B) Degree of connectivity of each gene group. Inset: relationship between degree of connectivity and CS value of group CS0 (set of human essential genes).

(C) Relationship between gene essentiality and gene length.

(D) Relationship between gene length and transcript count.

(E) Heatmap showing level of overlap among 10 types of gene TSSs.

(F) Profile of 10 types of gene TSSs near IMR90 TAD boundary. Inset: mean value of TAD boundary and near-boundary regions (≤100 kb).

Chromatin accessibility **(G)**, methylation level (MeDIP-seq, **H**), and H3k4me3 density **(I)** in 10-kb regions up- or downstream of gene body.

In many model organisms, essential genes tend to encode abundant proteins that engage extensively in protein-protein interactions (PPIs)(Jeong et al. 2001; Davierwala et al. 2005; Ishihama et al. 2008; Costanzo et al. 2010; Hart et al. 2014; Costanzo et al. 2016). We constructed PPI networks for each of the 10 groups of human genes (Supplementary Fig. 1B). Essential human genes tended to act as central hubs in the PPI networks (Fig. 2B). Gene Ontology (GO) analysis to compare biological functions of gene groups CS0 to CS9 showed that the set of essential genes (CS0) was enriched in many fundamental cellular processes, such as rRNA processing, translational initiation, mRNA splicing, and DNA replication. Nonessential human genes (groups CS1–CS9) did not show similar functional enrichment, although functional enrichment was different in each group (Supplementary Fig. 2).

We calculated the degree of connectivity for human essential genes and compared these results with the CS values (block diagram of Fig. 2B). The CS values tended to decrease as the

degree of connectivity increased, suggesting that the gene essentiality gradient is associated with centrality in the PPI network. We also investigated the proportion of essential genes with different degrees of connectivity in the PPI network. The number of genes decreased with increasing degree of connectivity in the human PPI network, but the proportion of essential genes increased (Supplementary Fig. 3A-B).

In summary, we found that proteins encoded by essential human genes are highly expressed, very stable, perform important biological functions, and serve as core hubs in the PPI network. In other words, essential genes in human cells show essentiality at the protein level.

Structural essentiality: clustering, stability, and chromosomal scaffold. We examined the general structural characteristics of genes in the 10 groups. Compared to nonessential genes (groups CS1–CS9), essential genes (group CS0) were shorter and had higher transcript counts, which is a typical property of long genes (Fig. 2C, Fig. 2D inset). Transcripts were denser for essential genes.

We examined the distribution of GC content to understand the internal gene structure. The GC content increased with decreasing CS value, with the highest GC content being obtained for the set of essential genes (Supplementary Fig. 4A). Variations in the GC ratio within the genome result in variations in staining intensity in chromosomes(Comings 1978; Holmquist et al. 1982; Bernardi 1989; Furey and Haussler 2003). High GC content means greater thermostability, suggesting that the DNA of essential genes may be more stable than that of nonessential genes.

Repetitive sequences in the genome have multiple functions, including a major

architectonic role in higher-order physical structuring(Shapiro and von Sternberg 2005). Therefore, we investigated the structural characteristics of essential genes within repetitive elements. DNA sequences for interspersed repeats and low-complexity DNA sequences were identified and masked by RepeatMasker(Smit et al. 2015). We found that repetitive elements were more enriched in essential than in nonessential genes (Supplementary Fig. 4B). Previous studies(Ivessa et al. 2000; Ivessa et al. 2002; Argueso et al. 2008; Shore and Bianchi 2009; Branzei and Foiani 2010) reported that the repetitive component of the genome can detect and repair errors and damage to the genome, even restructuring the genome when necessary (as part of the normal life cycle or in response to a critical selective challenge). Thus, our findings of gene colocalization with repetitive elements and high GC content suggest that essential genes help maintain genomic stability.

The level of overlap among the 10 types of gene transcription start sites (TSSs) of the one-dimensional (1D) genome (Fig. 2E, Supplementary Fig. 5A) indicated that the essential genes were prone to clustering. Therefore, we investigated the three-dimensional (3D) structural organization of the genes by collecting high-throughput/high-resolution chromosome conformation capture (Hi-C) data from Ren(Jin et al. 2013) and generated contact maps with a bin size of 10 kb. The TSSs of essential genes were closer than the TSSs of nonessential genes to the boundary of topologically associated domains (TADs)(Dixon et al. 2012) (Fig. 2F, Supplementary Fig. 5B-E). TADs are highly conserved between cell types and species(Dixon et al. 2012; Nora et al. 2012; Sexton et al. 2012; Nagano et al. 2013; Dixon et al. 2015), and proximity to the TAD boundary likely contributes to the stabilization of gene expression.

Next, we calculated the mean intra-TAD local (<100 kb) Hi-C contacts for bins containing 10 types of gene TSSs (Supplementary Fig. 6A-B) and determined the level of overlap among the 10 types of gene TSSs and chromatin loop anchors (Supplementary Fig. 6C). Essential genes had greater amounts of local contacts and greater clustering in chromatin loop anchors. Chromatin may be associated with proteins having affinity for each other, resulting in chromatin loops(Krijger and de Laat 2016). Architectural loops strongly depend on the protein CTCF(Heath et al. 2008), which can form chromatin loops between dispersed binding sites(Splinter et al. 2006; Handoko et al. 2011), and the cohesin protein complex. We found that CTCF was more likely to appear near human essential genes (Supplementary Fig. 6D).

In summary, human essential genes are significantly different from other genes in terms of their sequence and structural characteristics. Essential genes are highly stable due to their short gene length, high GC content, high transcript count, and presence in highly repetitive elements. Essential genes cluster in the genome and are located close to the chromosomal scaffold, reflecting their key role in genomic structural stability. Thus, these genes show more structural essentiality than other genes.

Epigenetic essentiality: clustering of active epigenetic modifications. The recent adaptation of high-throughput genomic assays has greatly facilitated the global study of epigenetics(Laird 2010; Zhu et al. 2013). Therefore, we examined the effect of epigenetic regulation in essential genes. To investigate the level of chromatin accessibility near essential genes, we compared DNase I hypersensitive site (DHS) profiles among gene groups in the H1-hESC cell line (Fig.

2G). The TSSs of essential genes corresponded to the strongest DHS profiles, and strength decreased with decreasing gene essentiality. This result suggests that decreased expression arose from decreased chromatin accessibility due to remodeling of the chromatin structure and alteration of the DNA packaging density(Kouzarides 2007). The density of transcription factor binding sites near gene TSSs also decreased with decreasing gene essentiality (Supplementary Fig. 7).

DNA methylation is a major epigenetic modification in multicellular organisms(Bird 2002). We used two sequencing-based methods, methylated DNA immunoprecipitation sequencing(Down et al. 2008; Jacinto et al. 2008) (MeDIP-seq) and methylation-sensitive restriction enzyme sequencing (MRE-seq), to perform genome-wide DNA methylation profiling of essential and nonessential genes. Both methods revealed a minimum methylation level at the TSSs of essential genes. Comparing among groups CS0 to CS9, methylation level decreased with increasing level of essentiality (Fig. 2H and Supplementary Fig. 8A).

Histone modifications are commonly associated with active transcription of nearby genes(Guenther et al. 2007) and act in diverse biological processes, such as gene regulation, DNA repair, chromosome condensation (mitosis), and spermatogenesis (meiosis)(Song et al. 2011). We focused on two histone modifications: trimethylation of H3 lysine 4 (H3K4me3) and trimethylation of H3 lysine 27 (H3K27me3). The H3K4me3 modification, which is associated with transcriptional activation, occurs at the promoter region of active genes(Krogan et al. 2003; Ng et al. 2003; Bernstein et al. 2005) and in nucleosomes adjacent to TSSs, where it relaxes the chromatin structure(Calo and Wysocka 2013). In contrast, the H3K27me3 modification is a clear

marker of gene repression(Cao et al. 2002). We observed opposite results for these two modifications: the H3K4me3 modification level decreased as CS value increased (Fig. 2I), whereas the H3K27me3 modification level was lowest for the set of essential genes with the lowest CS values (Supplementary Fig. 8B).

We calculated the abundance of noncoding RNAs (ncRNAs) for the 10 gene groups (Supplementary Fig. 8C). Although understanding of ncRNA function is still quite limited, research suggests that ncRNAs have diverse regulatory activities, and that regulatory RNA networks represent a crucial phase between genomics and proteomics. We found that ncRNAs were present at high levels near the TSSs of essential genes, suggesting that ncRNAs may promote gene expression by remodeling chromatin activity.

Overall, our results indicate a very strong correlation between epigenetics and gene essentiality, highlighting the critical role of epigenetics in cell survival. Essential genes were located near highly accessible chromatin regions that can bind many transcription factors and histones, thus enabling nearby genes to have direct effects on the surrounding microenvironment. Absence of an essential gene in this context will greatly affect the surrounding epigenetic modifications and microenvironment. In other words, gene essentiality is reflected in the epigenetics and cellular environment.

Characteristics of human essential genes in evolution and embryonic development.

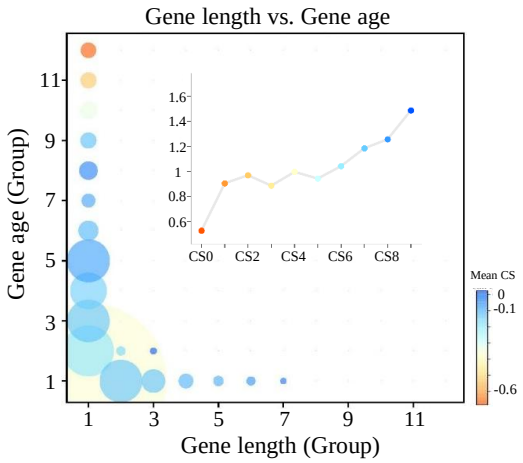
Evolutionary status of human essential genes. We investigated but found no significant

correlation between gene essentiality and gene conservation (Supplementary Fig. 9A-C). Therefore, we next examined the evolutionary status of human essential genes. Previous studies found that younger genes tend to be short, and that longer genes tend to have high transcript counts(Alba and Castresana 2005; Wolf et al. 2009; Chen et al. 2012b). Although the results of our genome-wide analysis confirmed these relationships, we observed a clear inverse correlation when we focused solely on essential genes, which tended to be shorter but older than most of the nonessential genes (Fig. 3A inset). This inverse correlation was due to the uneven distribution of essential genes (Fig. 3A). Although young genes were shorter, they also were mainly nonessential genes, whereas most of the essential genes were shorter and older genes. We found a high level of essentiality in the youngest genes, suggesting the existence of a special minority of very young and short essential genes. Short genes usually had smaller transcript counts, whereas essential genes were the minority of short genes that had the highest transcript counts. One explanation for this finding is that the higher transcript diversity of essential genes led to greater diversity of encoded proteins, thereby contributing to gene essentiality in cell function.

Most of the human essential genes are homologous genes inherited from ancient species, although their essentiality often cannot be traced back to these ancestors. We found that human essential genes showed only slight overlaps in sequence homology with orthologs in mouse, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Danio rerio*, and *Schizosaccharomyces pombe* (Supplementary Fig. S10). Mouse essential genes were slightly more homologous with human essential genes than were essential genes of other species. In this analysis, Human essential genes were derived from two different gene sets: our set of 1,878 essential genes from

group CS0, and another set of 8,253 essential genes from a database of multiple experimental data(Luo et al. 2014). The weak overlap of essential genes among species suggests that genes constantly gain and lose essentiality throughout evolution, and that essential genes of different species should be analyzed separately. Moreover, the evolvability of gene essentiality may be dependent on environmental and genetic contexts(Rancati et al. 2017).

A


Gene length vs. Gene age
Gene age (Group)
Gene length (Group)
1
3
5
7
9
11
1.6
1.4
1.2
1
0.8
0.6
CS0
CS2
CS4
CS6
CS8
Mean CS
0
-0.1
-0.6

B

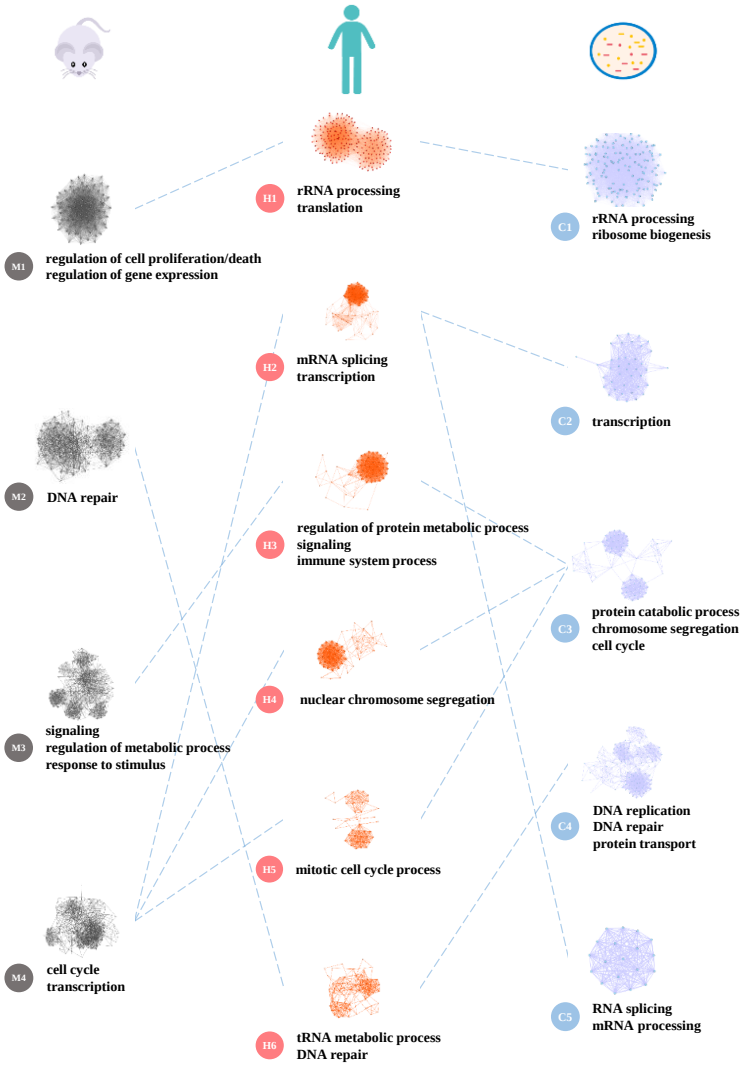

H1
rRNA processing
translation
C1
rRNA processing
ribosome biogenesis
M1
regulation of cell proliferation/death
regulation of gene expression
H2
mRNA splicing
transcription
C2
transcription
M2
DNA repair
H3
regulation of protein metabolic process
signaling
immune system process
C3
protein catabolic process
chromosome segregation
cell cycle
M3
signaling
regulation of metabolic process
response to stimulus
H4
nuclear chromosome segregation
C4
DNA replication
DNA repair
protein transport
H5
mitotic cell cycle process
M4
cell cycle
transcription
C5
RNA splicing
mRNA processing
H6
tRNA metabolic process
DNA repair

Figure 3 Evolutionary status of human essential genes

(A) Scatter-plot of gene length and evolutionary age. Size of the circle indicates number of genes; color represents average CS value.

(B) Essential gene associated (EG-associated) specific functional modules. The network of specific functional interactions among the 1,878 human essential genes was clustered by using a graph theory clustering algorithm to elucidate gene modules. Six clusters that containing ≥40 genes (H1-H6) were tested for functional enrichment by using genes annotated with GO biological process terms. Representative processes and pathways enriched within each cluster are presented alongside the cluster label. Enriched functions provide a landscape of the potential effects of cellular functions for essential genes. Similar functional processes were shared by essential genes in mouse (four subnet modules) and *S. cerevisiae* (five subnet modules).

Although not conserved across species, all essential genes play critical roles in cell survival and may be involved in similar functional processes. Therefore, we investigated the specific functions of essential genes in various species in the context of PPI networks, using the network topology described by specific centrality parameters(Scardoni et al. 2009). Network topological quantification can be combined with other numerical node attributes to provide biologically meaningful node identification and functional classification. Densely connected regions in large PPI networks were found by using the molecular complex detection method(Bader and Hogue 2003), with the large functional protein association network being clustered into distinct and substantially cohesive functional modules. To examine the functions of these subnet modules, we used gene set enrichment analysis(Subramanian et al. 2005). Each subnet module had its own specific characteristics. Similar functional processes were shared by essential genes in mouse and *S. cerevisiae* (Fig. 3B). These analyses suggest that cellular life is

maintained by the similar functions of an ever-changing group of essential genes across species and throughout evolution.

Taken together, our results suggest that essential genes are typically older in age, shorter in length, and higher in transcript count, although the possible causal relationship between these characteristics and gene essentiality remains unclear. High essentiality is not a constant for genes in the course of evolution but is highly species-specific. However, different essential genes in different species serve nearly the same functions to maintain cellular life.

These evolutionary characteristics of human essential genes have several biological and medical implications. First, essential genes are not more conserved, but their expression levels are higher. Inference of the relative importance of genes from their evolutionary conservation is expected to be unreliable, unless there is a large difference in evolutionary rates. Second, essential genes occupy a special position in evolution: they are functionally important genes with relatively short length and old evolutionary age, which encode stable proteins that are expressed at high levels. These features can be applied in biological research. Third, gene essentiality is not a fixed property of a gene between different species but is context-dependent and evolvable in all kingdoms of life. Fourth, essential genes are screened out from unacceptable “gain-of-toxicity” mutations, which is useful knowledge for understanding some genetic diseases. Computational analysis suggests that most disease-causing missense mutations reduce protein structural stability, which would increase the probability of misfolding(Wang and Moult 2001; Gregersen et al. 2006). Mutations of essential genes similarly cause biological toxicity by affecting structural stability. Fifth, the ability to predict pathogenic mutations has become very important

in medical care(Cooper and Shendure 2011). Essential genes could serve as an important reference in this area.

Birth and maturation of essential genes in embryonic development. Cell fate decisions fundamentally contribute to the development and homeostasis of complex tissue structures in multicellular organisms. By examining mammalian embryos during development, we can determine how and why apparently identical cells have different fates(Zernicka-Goetz et al. 2009). The key to answering these questions lies in the emergence of transcriptional programs. As direct outputs of the gene regulatory network, transcriptional programs map the phenotype and fate of cells.

Therefore, we examined the roles of essential genes in mammalian embryonic development. Due to the lack of experimental data on human embryos, we used preimplantation mouse embryos as a common animal model for the study of human development. We analyzed the characteristics of human embryonic development using mouse multiomic data, which were consistent with human data (Fig. 2A, G–I vs. Fig. 4A–E). Essential genes were expressed at much higher levels than nonessential genes (Fig. 4A), with two peaks in expression. The first peak was observed after the 2-cell stage, which straddles the period of zygotic genome activation(Schultz 2002; Schier 2007) and the destruction of many of the maternally provided mRNAs. The second peak was observed after the inner cell mass (ICM) formed on embryonic day (E) 3.5. This stage represents the first fate decision, when cells positioned on the inside (the ICM) retain their pluripotency but cells on the outside develop into extraembryonic

trophectoderm(Zernicka-Goetz 2006; Zernicka-Goetz et al. 2009). In contrast to essential genes, nonessential genes (CS7–CS9) showed decreasing expression levels throughout development and presented a period of maternal mRNA degradation (Fig. 4A).

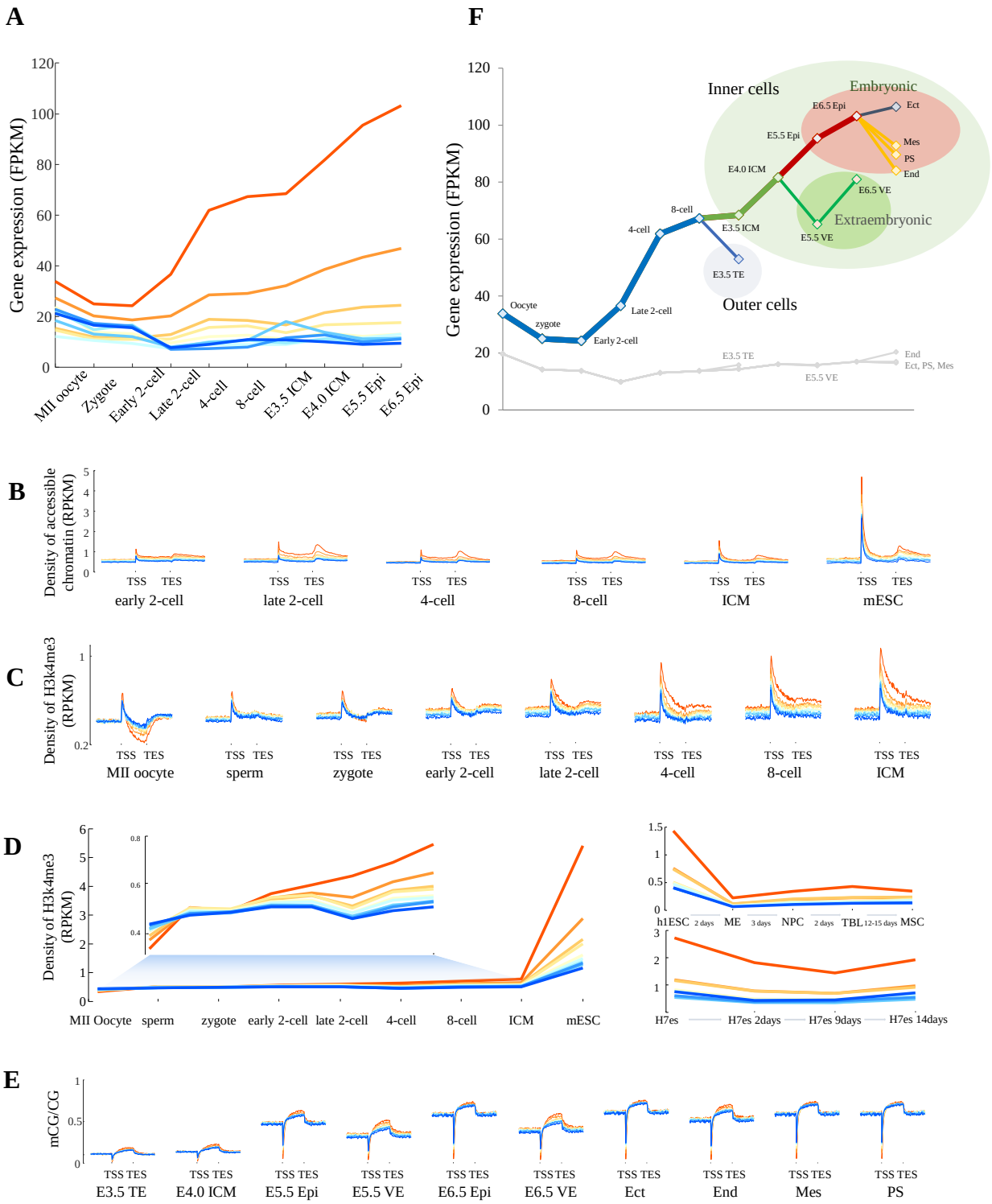

A
Gene expression (FPKM)
MII oocyte
Zygote
Early 2-cell
Late 2-cell
4-cell
8-cell
E3.5 ICM
E4.0 ICM
E5.5 Epi
E6.5 Epi
F
Inner cells
Embryonic
Extraembryonic
Outer cells
Oocyte
zygote
Early 2-cell
Late 2-cell
4-cell
8-cell
E3.5 ICM
E3.5 TE
E4.0 ICM
E5.5 Epi
E6.5 Epi
E5.5 VE
E6.5 VE
Ect
Mes
PS
End
Ect, PS, Mes
B
Density of accessible chromatin (RPKM)
TSS TES
early 2-cell
late 2-cell
4-cell
8-cell
ICM
mESC
C
Density of H3k4me3 (RPKM)
MII oocyte
sperm
zygote
early 2-cell
late 2-cell
4-cell
8-cell
ICM
D
Density of H3k4me3 (RPKM)
MII Oocyte
sperm
zygote
early 2-cell
late 2-cell
4-cell
8-cell
ICM
mESC
h1ESC
ME
NPC
TBL
MSC
H7es
H7es 2days
H7es 9days
H7es 14days
E
mCG/CG
E3.5 TE
E4.0 ICM
E5.5 Epi
E5.5 VE
E6.5 Epi
E6.5 VE
Ect
End
Mes
PS

Figure 4 Birth and maturation of essential genes in embryonic development.

(A) Gene expression level for essential genes (red lines) and other groups at each developmental stage.

(B) Density of accessible chromatin surrounding TSSs and TESs of genes at each developmental stage.

(C) Density of active H3k4me3 modifications surrounding TSSs and TESs of genes at each developmental stage.

(D) Dynamics of H3k4me3 density in gene body in preimplantation mouse embryos (left) and postimplantation human embryos (right).

(E) CG methylation profiles surrounding TSSs of genes at each developmental stage.

(F) Dynamics of gene expression for essential and other genes at each developmental stage. TE, trophectoderm; ICM, inner cell mass; VE, visceral endoderm; Epi, epiblast; Ect, ectoderm; End, endoderm; Mes, mesoderm; PS, primitive streak.

Next, we investigated the rationale for why certain genes are essential in mammalian embryonic development. As a first step, we checked the state of the chromatin surrounding the genes. Previous reports showed strong ATAC-seq (assay for transposase-accessible chromatin using sequencing) signals near TSSs and transcription ending sites (TESs), but much weaker enrichment in somatic cells(Lavin et al. 2014; Maza et al. 2015; Wu et al. 2016). The density of accessible chromatin increased during early embryonic development, and greater accessibility was observed in the TSSs and TESs of essential genes compared to other genes (Fig. 4B). The density of active H3k4me3 modifications surrounding gene promoters increased through early embryonic development and was much higher in essential genes (Fig. 4C). Throughout development of human embryonic stem cells (ESCs), the density of active H3k4me3 modifications in essential genes was higher than it was in other genes, and it decreased after ESCs (Fig. 4D). Nonetheless, the DNA methylation level was much lower in the promoters of

essential genes than in the promoters of other genes (Fig. 4E).

We further checked whether the emergence of essential genes contributes to early lineage specification. As expected, we found that early cell fate commitment was accompanied by the differential expression of essential genes, suggesting that essential genes guide cells to develop into specific lineages (Fig. 4F). At E3.5, cells with higher expression levels of essential genes (so-called “essential gene cells”) developed into ICM. Remaining cells developed into trophectoderm, which will form most of the fetal-origin part of the placenta(Rossant and Tam 2009). At E5.5, essential gene cells developed into epiblast lineage cells, which will give rise to the entire fetus(Bielinska et al. 1999; Rossant and Tam 2009). Remaining cells developed into visceral endoderm, which will form the endoderm and primitive streak. After E6.5, essential genes contributed to the segregation of three layers: the ectoderm, which forms from the anterior epiblast, the mesoderm, and the endoderm, which forms from the primitive streak that develops from the posterior proximal epiblast(Arnold and Robertson 2009) (Fig. 4F). Taking the other genes as background, we observed a slight change in gene expression and found that lineage specification was ambiguously decided (grey line in Fig. 4F).

Our findings showed that chromatin accessibility and epigenetic modifications contribute to the formation of transcriptional programs in essential genes. Moreover, our analysis provides unprecedented spatiotemporal views for the emergence and development of transcriptional programs in essential genes during embryonic development, and indicates that essential genes can be useful as markers in lineage segregation.

Potential applications of human essential genes in cancer research and treatment.

Given their fundamental roles, it is not surprising that essential genes are the targets of many antimicrobial compounds(Roemer et al. 2003; Lu et al. 2014; Paul et al. 2014). However, the therapeutic implications of essential genes are not fully realized. Combined analysis with the Online Mendelian Inheritance in Man (OMIM) Disease Database revealed that very few human diseases have been correlated significantly with essential genes (Supplementary Table 1). For example, only 202 essential genes were associated with diseases in the KEGG DISEASE database, and congenital disorders were the most common diseases to be associated with essential genes.

Interestingly, we found five essential genes that were obviously related to cancer: *BRCA1*, *BRCA2*, *MYC*, *EZH2*, and *SMARCB1*. *BRCA1* and *BRCA2* are tumor suppressors involved in the DNA damage response. All five genes share a common trait that affects chromatin stability, remodeling, and modification. This observation is consistent with our previous analysis. These five genes are not the direct cause of cancer, but they can affect chromatin in different ways to alter biological function, so that the cell becomes disordered or dysregulated. Thus, we speculate that essential genes may be important in human cancer development.

Hart et al.(Hart et al. 2015) focused on essential genes in tumor cells that could serve as therapeutic targets for cancer. We further explored this possibility and compared essential gene lists with cancer-related gene lists from different reports(Kandoth et al. 2013; Vogelstein et al.

2013; Lawrence et al. 2014; Schroeder et al. 2014; Zehir et al. 2017). Cancer genes differed largely among studies, and there was little intersection between cancer gene and essential gene lists (Fig. 5A). This finding suggests that the current cancer research is far from sufficient, and that there is no consensus on the identification and understanding of cancer genes. Of the 260 cancer genes identified by Lawrence et al.(Lawrence et al. 2014), 54 were essential genes, corresponding to 20.8% of all cancer genes but only 10.4% all human protein-coding genes (Fig. 5B). These observations indicate that essential genes are closely related to cancer and are more likely than nonessential genes to be key genes for cancer research.

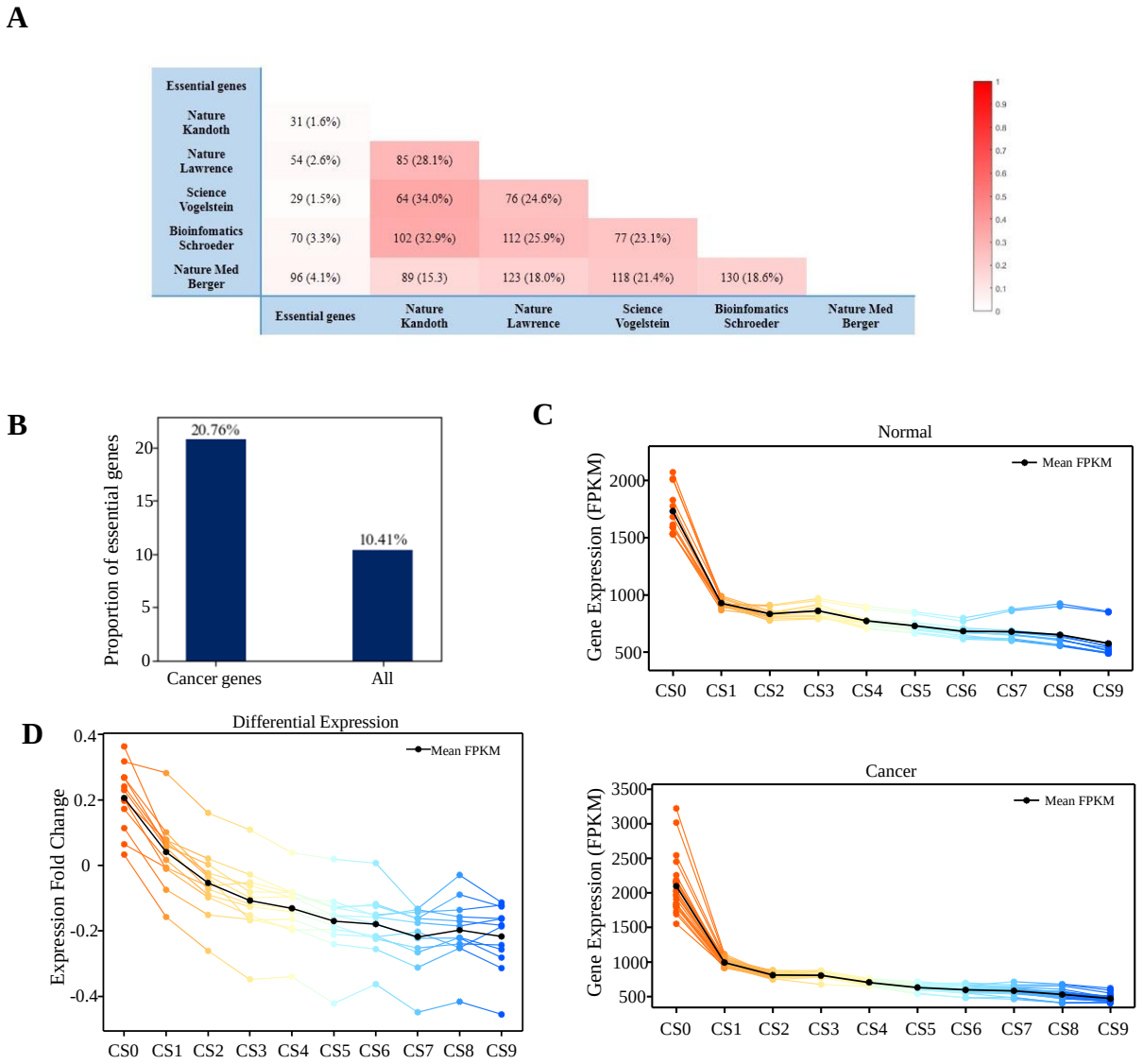


Figure 5 Relationship between human essential genes and cancer.

(A) Comparison of sets of human essential genes and cancer genes from different studies.

(B) Proportions of human essential genes in the sets of 260 cancer genes and all human genes.

(C) Gene expression in normal and tumor tissues from TCGA database, based on gene essentiality. Each line represents one tissue.

(D) Differential expression of genes in normal and tumor tissues from TCGA database.

We calculated the average expression levels of all 260 cancer genes described by

Lawrence et al.(Lawrence et al. 2014) and all of 18,166 protein-coding genes in the H1-hESC cell line and The Cancer Genome Atlas (TCGA) database. The 260 cancer genes were expressed at significantly higher levels than the protein-coding genes (Supplementary Fig. 11A). Essential genes are also expressed at relatively high levels and, therefore, act very similarly to cancer genes at the expression level. We analyzed the top 1,000 most highly expressed genes in the H1-hESC cell line and TCGA tumor tissues for cancer genes, essential genes, and nonessential genes. The H1-hESC cell line expressed fewer of the 260 cancer genes than most of the TCGA tumor tissues, indicating that the overall expression of the 260 cancer genes was higher in tumor tissues than in embryonic stem cells (Supplementary Fig. 11B). Similar numbers of cancer genes belonging to the essential gene set were found in the H1-hESC cell line and TCGA tumor tissues. Thus, a considerable number of cancer-related essential genes are expressed at high levels in both tumor cells and in the H1-hESC cell line (Supplementary Fig. 11C). Significantly fewer nonessential genes were expressed in the H1-hESC cell line than in TCGA tumor tissues (Supplementary Fig. S11D). Therefore, the number of highly expressed nonessential genes could significantly differentiate between tumor tissue and the H1-hESC cell line. Higher expression of nonessential genes is a major feature of cancer cells, suggesting that the difference between cancer and embryonic stem cells can be studied from nonessential genes.

TCGA gene expression data were analyzed to study the relationship between essential genes and cancer. Essential genes were expressed at significantly higher levels than nonessential genes in normal and cancer tissues (Fig. 5C). The high expression of essential genes is consistent with the above analyses. For each cancer, we analyzed the differential expression of essential

genes between cancer and normal tissues. Although expression levels differed markedly between different cancer types, essential genes were always expressed at higher levels in cancer than in normal tissues (Fig. 5D). This finding illustrates that essential genes are an important factor in distinguishing normal from cancer tissues and may be an important element in cancer research.

We propose that essential genes can be used to find candidate drugs because these genes have high differential expression in cancer. Using TCGA data with an appropriate screening process, we selected 297 significantly differentially expressed genes as potential tumor markers (Supplementary Table 2). In silico screening of the Drugbank database(Law et al. 2014) identified 135 candidate drugs for these 297 genes (Supplementary Table 3). Some of these candidate drugs have already been used as antineoplastic agents (pemetrexed, decitabine, doxorubicin, mitoxantrone, capecitabine), but many are not yet approved for use in cancer. The list of candidate drugs includes anti-infective agents (trifluridine, fleroxacin), antibacterial agents (enoxacin, pefloxacin, ciprofloxacin), nutraceuticals (pyridoxal phosphate, tetrahydrofolic acid, L-aspartic acid) and experimental drugs (DB04395, DB08617, DB06963). The primary therapeutic use for colchicine is for gout, although it is being investigated as a potential anticancer drug. Trovafloxacin is a broad-spectrum antibiotic that has been withdrawn from the market. Although many of the identified candidate drugs are not used in cancer therapy, they may have a role in cancer treatment and could be considered in experiments aimed at drug discovery and drug repositioning.

HEGIAP: an interactive web server for studying essential genes.

We developed the interactive web server HEGIAP, which integrates abundant analytical and visualization tools to give a multilevel interpretation of gene essentiality for a single gene. HEGIAP provides an overall gene property graph, which shows the gene length, transcript count, and distributions of exons and introns of each transcript. Boxplots are provided for gene properties, including but not limited to gene length, protein length, exon count, and repetitive element count near the promoter region, for the 10 groups of genes (CS0–CS9). The graphs show the corresponding value and group number of each selected gene. For a selected gene, the web server provides histone modification, methylation, and chromatin accessibility profiles and the Hi-C contact map of chromatin structure, all of which have been shown to be significantly correlated with the CS value. Multigene analysis is available to study groups of genes in a comprehensive manner (Fig. 6).

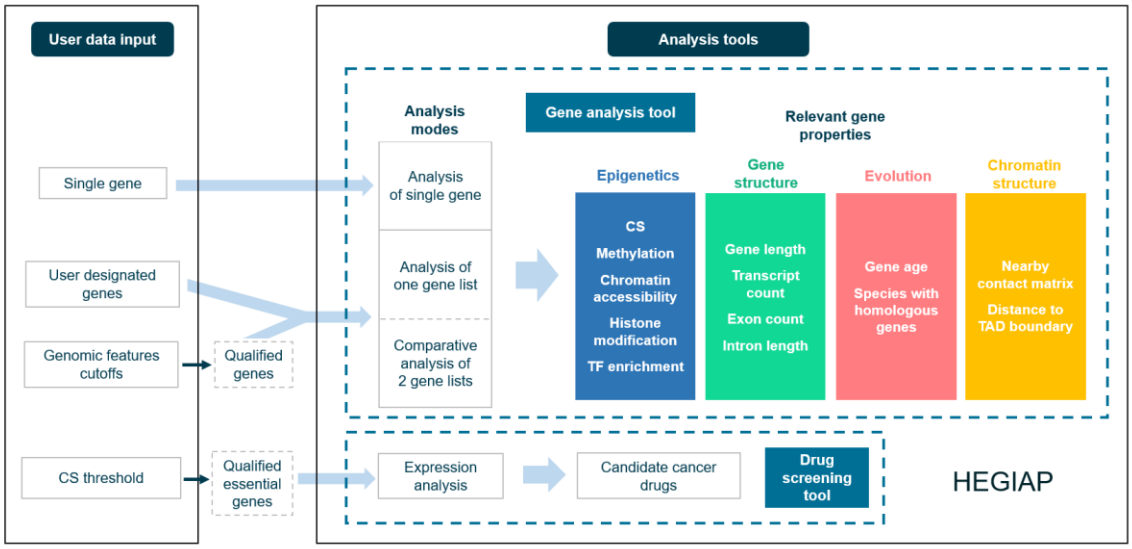


Figure 6 Integrative analysis of individual genes by HEGIAP.

HEGIAP supports both feature- and gene-oriented analyses. In feature-oriented analysis, users can obtain all of the genes that meet chosen screening thresholds for multiple features. They can examine the CS distribution or any other property, enabling exploration of possible correlations between CS and other genomic features. In the classic gene-oriented analysis, users specify their chosen genes and are provided a comprehensive view of their essentiality and genomic features. Comparative analysis of two different gene lists is provided to facilitate free exploration of the variation of genomic features between genes of interest or between genes screened by different degrees of essentiality or any other property. Tools are provided to identify genes with aberrant epigenetic modification levels or genomic features based on their essentiality.

HEGIAP identifies differentially expressed genes in tumor/normal tissues. The web server offers a drug-screening tool, which is based on the assumption that essential genes have predictive power in finding candidate drugs for cancer. Users can set a CS threshold and acquire a list of drugs that significantly suppress expression of cancer-specific highly expressed genes filtered by the CS threshold. The user-friendly interface of HEGIAP was constructed by using R Shiny and requires no plugin installation for users running any popular web browser.

Discussion

Gene essentiality is a key concept of genetics with important implications in both fundamental

and applied research. Systematic research is needed to improve understanding of these genes, which are vital to human life. In this study, we systematically investigated the genetic and regulatory characteristics, including the genetic, epigenetic, and proteomic features, of human essential genes. Our study is a massive-scale effort that aims to discover why essential genes are essential.

From the proteomic perspective, proteins encoded by essential genes have a core function and are the central hub of the PPI network. Thus, these proteins are essential because of their important functions and influence on other proteins. From the genomic perspective, essential genes tend to be at the structural hub and to play important roles in genomic stability. From the epigenetic and environmental perspectives, chromatin near essential genes is highly accessible, with greater influence on epigenetic modifications and cellular environment. Thus, the epigenetic environment of essential genes is more essential.

We studied gene essentiality in terms of evolution and embryonic development. High gene essentiality was related to old evolutionary age but also short gene length, which is more typical of younger genes. The findings indicate that essential genes constitute a distinct category of genes. Gene essentiality was not constant across species or throughout evolution; genes may lose or gain high essentiality due to changes in function. As a result, essential genes comprise different groups of genes across species, but with similar functions that maintain cellular life. At the initial stages of cell development, essential genes are not prominently expressed. Before the 2-cell stage, the active histone modification level near essential genes is low. However, the chromatin of essential genes gradually opens, the active histone modification level gradually

increases, and DNA methylation maintains a low level with development of the cell. Thus, it can be said that the epigenetic regulatory environment creates essential genes.

Our comprehensive study of human gene essentiality was enabled by the development of CRISPR screening technology. Previous results were obtained by integrating the PPI network to improve the quality of RNAi screening results, which are very noisy due to off-target effects and low knockdown efficiency(Kaplow et al. 2009; Tu et al. 2009; Wang et al. 2009; Ma et al. 2014). Consequently, PPI network data should improve the quality of CRISPR screening results. X. Shirley Liu has carried out relevant research(Jiang et al. 2015). In turn, we analyzed essential human genes through key nodes in the human PPI network. About 70% of essential genes were located in the top 6,000 key nodes in the network, and the top 10,000 key nodes covered 90% of all essential genes (Supplementary Fig. 12A–B). Our results provide a guideline for CRISPR experiments, showing that essential genes can be used as an important reference factor to determine the genetic screening range.

Essential genes were expressed at significantly higher levels in tumor than in normal tissues and, therefore, should be actively considered in cancer research. Essential genes could facilitate drug discovery in cancer, offer promise as markers in cancer research, and may be useful for identifying clinical therapeutic applications.

Several limitations of this study are worth mentioning. First, we carried out a systematic data analysis and theoretical exploration for human essential genes, but it was difficult to validate all of the aspects experimentally. Second, the reliability and integrity of the data have some limitations because of the technical level and data quality, although we tried to choose highly

reliable data and literature sources. For example, Hi-C data were obtained by using the latest technological method, but few such data were available. Similarly, the field of preimplantation embryonic development is relatively new, and data in this field are limited. Finally, for the evolutionary studies, the essential genes of different species were screened by different experimental methods.

As technology advances, we will continue to consider the latest multiomic data. Such technological advancements, particularly with respect to the nature of human life and the interpretation of DNA structure, will enable an even-more-comprehensive understanding of essential genes in the human genome, including their roles in preimplantation embryos and evolution. Extensive collaborative and in-depth experimental validation analyses will be performed in clinical disease research. We believe that human essential genes are important markers of disease that can bring new insights and breakthroughs for clinical research.

Methods

Dataset. Genes annotated as human were obtained from the GENCODE database (V21). Data on DHSs, DNA methylation, histone modifications, CTCF, and conservation were downloaded from the ENCODE project and RoadMap database. Hi-C data were from Ren(Jin et al. 2013) (GSE43070). The TAD boundary of the Hi-C data refer to the boundary described in Schmitt et al.(Schmitt et al. 2016) Essential genes from different species were obtained from DEG(Luo et al. 2014) (http://www.essentialgene.org/). Data on transcription factors were downloaded from

Transfac and Jaspar databases. Data on ncRNAs were downloaded from the NONCODE database(Zhao et al. 2016) (http://www.noncode.org). Data on gene expression, accessible chromatin, and histone modifications of preimplantation embryos were obtained from GSE66390, GSE71434, and GSE76687. Cancer data were downloaded from TCGA. Drug data were obtained from the DrugBank database.

Gene expression data of preimplantation mouse embryos and ATAC-seq data on accessible chromatin regions in mouse were obtained from GSE66582, GSE76505 (preimplantation gene expression), and Wu et al.(Wu et al. 2016) (GSE66390, ATAC-seq). Data on histone modification H3k4me3 in mouse were obtained from Zhang et al.(Zhang et al. 2016) (GSE71434) and the ENCODE project. Data on the histone modification H3k4me3 of differentiated H1-hESC cells were obtained from Xie et al.(Xie et al. 2013) Data on the H3k4me3 modification of differentiated H7es cells were obtained from the ENCODE project. Mouse gene annotations were obtained from the Mouse Genome Database(Blake et al. 2017).

Division of genes into groups by CS value. Genes were sorted into 10 groups (CS0–CS9) by ascending CS value, as reported by Wang et al.[(Wang et al. 2015)] There were 18,166 human genes for which CS values were available. Genes were divided into nine groups of 1,878 genes each (CS0–CS8) and one group of 1,264 genes (CS9). Group CS0, which contained genes with the smallest CS values, was defined as the set of essential human genes.

Hi-C data processing. Raw contact matrices were constructed, and normalized matrices were

obtained by using HOMER (http://homer.ucsd.edu/homer/). These data were used to construct the average contact map near the 10 types of gene TSSs. We conducted other analyses using 10-kb resolution Hi-C contact matrices and loops of GM12878, IMR90, and K562 from Rao et al.(Rao et al. 2014) The normalized Hi-C contact map was obtained by using SQRTVC normalization.

Construction of PPI network. We used STRING(Szklarczyk et al. 2015) (version: 10.5) to obtain PPI data of proteins encoded by the analyzed genes. We computed specific centrality parameters describing the network topology using Centiscape(Scardoni et al. 2009) and calculated the degree of connectivity of every node in the network from the PPI network data. Densely connected regions in large PPI networks were found by using the molecular complex detection method, which is based on vertex weighting by the local neighborhood density and outward traversal from a locally dense seed protein(Bader and Hogue 2003). Biological process functions were obtained from DAVID. We used P-value < 0.001 and count > 2 as thresholds for filtering.

Profiling of markers surrounding TSSs and TESs of genes. We counted the reads of accessible chromatin (Fig. 4B), H3k4me3 (Fig. 4C), and CG methylation (Fig. 4E) in 150 bins, including fifty 200-kb bins in the 10-kb region upstream of TSSs, fifty 200-kb bins in the 10-kb region downstream of TESs, and fifty equally sized bins in the gene body. RPKM values were calculated for each bin.

Enrichment analysis of gene sets with the OMIM Disease Database. We used the Enrichr tool(Kuleshov et al. 2016) to determine the top 10 disease-associated genes to be enriched in the essential gene set. However, the calculated significance levels for these associations, whether the *P*-value or combined score, were not high (Supplementary Fig. 11). The *P*-value was computed from the Fisher exact test, a proportion test that assumes a binomial distribution and independence for probability of any gene belonging to any set. Ranking was derived by running the Fisher exact test for many random gene sets to compute a mean rank and standard deviation from the expected rank for each term in the gene-set library. A z-score was calculated to assess deviation from the expected rank. The combined score was computed by multiplying the log of the Fisher exact test *P*-value by the z-score of the deviation from the expected rank.

Screening of cancer-related candidate genes. We analyzed 14 types of TCGA samples containing expression data of tumor and normal tissues. For each cancer type, we calculated the difference in expression between tumor and normal tissues, and screened 1,000 essential human genes having significantly higher expression in tumor than in normal tissues. We selected the first 5,000 genes with the most significantly different expression between tumor and normal tissues in the TCGA samples. Genes in both of the screened gene sets were defined as candidate genes for each cancer type. Finally, we identified genes that were present in the candidate gene lists of at least eight types of cancer as the ultimate cancer-related candidate genes.

Acknowledgments

We gratefully acknowledge funding from the Major Research plan of The National Natural Science Foundation of China (No. U1435222), the Program of International S&T Cooperation (No. 2014DFB30020), the National High Technology Research and Development Program of China (No.2015AA020108).

Author contributions

X.B. and H.L. conceived the project. H.C., Z.Z. and S.J. designed all experiments. H.C., Z.Z., S.J., H.L., R.L. and W.L. performed the experiments. All authors analysed the data and contributed to manuscript preparation. Z.Z. and S.J. wrote the manuscript.

Conflict of Interest

None declared.

References

Alba MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol Biol Evol* **22**(3): 598-606.

Argueso JL, Westmoreland J, Mieczkowski PA, Gawel M, Petes TD, Resnick MA. 2008. Double-strand breaks associated with repetitive DNA can reshape the genome. *P Natl Acad Sci USA* **105**(33): 11845-11850.

Arnold SJ, Robertson EJ. 2009. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat Rev Mol Cell Bio* **10**(2): 91-103.

Bader GD, Hogue CW. 2003. An automated method for finding molecular complexes in large protein interaction networks. *Bmc Bioinformatics* **4**.

Bernardi G. 1989. The isochore organization of the human genome. *Annual review of genetics* **23**: 637-661.

Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**(2): 169-181.

Bielinska M, Narita N, Wilson DB. 1999. Distinct roles for visceral endoderm during embryonic mouse development. *Int J Dev Biol* **43**(3): 183-205.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes & development* **16**(1): 6-21.

Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ, Grp MGD. 2017. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res* **45**(D1): D723-D729.

Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van Diemen FR, Olk N, Stukalov A et al. 2015. Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**(6264): 1092-1096.

Branzei D, Foiani M. 2010. Maintaining genome stability at the replication fork. *Nat Rev Mol Cell Bio* **11**(3): 208-219.

Calo E, Wysocka J. 2013. Modification of Enhancer Chromatin: What, How, and Why? *Mol Cell* **49**(5): 825-837.

Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, Jones RS, Zhang Y. 2002. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**(5595): 1039-1043.

Chen WH, Minguez P, Lercher MJ, Bork P. 2012a. OGEE: an online gene essentiality database. *Nucleic Acids Res* **40**(D1): D901-D906.

Chen WH, Trachana K, Lercher MJ, Bork P. 2012b. Younger Genes Are Less Likely to Be Essential than Older Genes, and Duplicates Are Less Likely to Be Essential than Singletons of the Same Age. *Mol Biol Evol* **29**(7): 1703-1706.

Comings DE. 1978. Mechanisms of chromosome banding and implications for chromosome structure. *Annual review of genetics* **12**: 25-46.

Consortium GT. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**(6): 580-585.

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**(9): 628-640.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding HM, Koh JLY, Toufighi K, Mostafavi S et al. 2010. The Genetic Landscape of a Cell. *Science* **327**(5964): 425-431.

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan GH, Wang W, Usaj M, Hanchard J, Lee SD et al. 2016. A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**(6306).

Davierwala AP, Haynes J, Li ZJ, Brost RL, Robinson MD, Yu L, Mnaimneh S, Ding HM, Zhu HW, Chen YQ et al. 2005. The

synthetic genetic interaction spectrum of essential genes. *Nat Genet* **37**(10): 1147-1152.

Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W et al. 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**(7539): 331-336.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**(7398): 376-380.

Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM et al. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26**(7): 779-785.

Furey TS, Haussler D. 2003. Integration of the cytogenetic map with the draft human genome sequence. *Hum Mol Genet* **12**(9): 1037-1044.

Galperin MY, Koonin EV. 1999. Searching for drug targets in microbial genomes. *Curr Opin Biotech* **10**(6): 571-578.

Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, Osterman A. 2006. Essential genes on metabolic maps. *Curr Opin Biotech* **17**(5): 448-456.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B et al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**(6896): 387-391.

Gregersen N, Bross P, Vang S, Christensen JH. 2006. Protein misfolding and human disease. *Annu Rev Genom Hum G* **7**: 103-124.

Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**(1): 77-88.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F et al. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**(7): 630-638.

Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. 2014. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol* **10**(7).

Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S et al. 2015. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**(6).

Hartman JL, Garvik B, Hartwell L. 2001. Cell biology - Principles for the buffering of genetic variation. *Science* **291**(5506): 1001-1004.

Heath H, Ribeiro de Almeida C, Sleutels F, Dingjan G, van de Nobelen S, Jonkers I, Ling KW, Gribnau J, Renkawitz R, Grosveld F et al. 2008. CTCF regulates cell cycle progression of alphabeta T cells in the thymus. *The EMBO journal* **27**(21): 2839-2850.

Holmquist G, Gray M, Porter T, Jordan J. 1982. Characterization of Giemsa dark- and light-band DNA. *Cell* **31**(1): 121-129.

Hu W, Sillaots S, Lemieux S, Davison J, Kauffman S, Breton A, Linteau A, Xin C, Bowman J, Becker J et al. 2007. Essential gene identification and drug target prioritization in Aspergillus fumigatus. *PLoS pathogens* **3**(3): e24.

Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, Frishman D. 2008. Protein abundance profiling of the Escherichia coli cytosol. *Bmc Genomics* **9**.

Ivessa AS, Zhou JQ, Schulz VP, Monson EK, Zakian VA. 2002. Saccharomyces Rrm3p, a 5 ' to 3 ' DNA helicase that promotes replication fork progression through telomeric and subtelomeric DNA. *Genes & development* **16**(11): 1383-1396.

Ivessa AS, Zhou JQ, Zakian VA. 2000. The Saccharomyces Pif1p DNA helicase and the highly related Rrm3p have

opposite effects on replication fork progression in ribosomal DNA. *Cell* **100**(4): 479-489.

Jacinto FV, Ballestar E, Esteller M. 2008. Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *BioTechniques* **44**(1): 35, 37, 39 passim.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* **411**(6833): 41-42.

Jiang P, Wang HF, Li W, Zang CZ, Li B, Wong YLJ, Meyer C, Liu JS, Aster JC, Liu XS. 2015. Network analysis of gene essentiality in functional genomics experiments. *Genome Biol* **16**.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**(7475): 290-294.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu BF, Lu C, Xie MC, Zhang QY, McMichael JF, Wyczalkowski MA et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* **502**(7471): 333-+.

Kaplow IM, Singh R, Friedman A, Bakal C, Perrimon N, Berger B. 2009. RNaiCut: automated detection of significant genes from functional genomic screens. *Nat Methods* **6**(7): 476-477.

Koonin EV. 2000. How many genes can make a cell: The minimal-gene-set concept. *Annu Rev Genom Hum G* **1**: 99-116.

-. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* **1**(2): 127-136.

Kouzarides T. 2007. Chromatin modifications and their function. *Cell* **128**(4): 693-705.

Krijger PH, de Laat W. 2016. Regulation of disease-associated gene expression in the 3D genome. *Nature reviews Molecular cell biology* **17**(12): 771-782.

Krogan NJ, Dover J, Wood A, Schneider J, Heidt J, Boateng MA, Dean K, Ryan OW, Golshani A, Johnston M et al. 2003. The Paf1 complex is required for histone h3 methylation by COMPASS and Dot1p: Linking transcriptional elongation to histone methylation. *Mol Cell* **11**(3): 721-729.

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan QN, Wang ZC, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**(W1): W90-W97.

Laird PW. 2010. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* **11**(3): 191-203.

Lartigue C, Glass JI, Alperovich N, Pieper R, Parmar PP, Hutchison CA, Smith HO, Venter JC. 2007. Genome transplantation in bacteria: Changing one species to another. *Science* **317**(5838): 632-638.

Lavin Y, Winter D, Blecher-Gonen R, David E, Keren-Shaul H, Merad M, Jung S, Amit I. 2014. Tissue-Resident Macrophage Enhancer Landscapes Are Shaped by the Local Microenvironment. *Cell* **159**(6): 1312-1326.

Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu YF, Maciejewski A, Arndt D, Wilson M, Neveu V et al. 2014. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* **42**(D1): D1091-D1097.

Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**(7484).

Leuenberger P, Ganscha S, Kahraman A, Cappelletti V, Boersema PJ, von Mering C, Claassen M, Picotti P. 2017. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* **355**(6327).

Liao BY, Zhang JZ. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *P Natl Acad Sci USA* **105**(19): 6987-6992.

Lu Y, Deng JY, Rhodes JC, Lu H, Lu LJ. 2014. Predicting essential genes for identifying potential drug targets in Aspergillus fumigatus. *Comput Biol Chem* **50**: 29-40.

Luo H, Lin Y, Gao F, Zhang CT, Zhang R. 2014. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res* **42**(D1): D574-D580.

Ma XT, Chen T, Sun FZ. 2014. Integrative approaches for predicting protein function and prioritizing genes for complex phenotypes using protein interaction networks. *Brief Bioinform* **15**(5): 685-698.

Maza I, Caspi I, Zviran A, Chomsky E, Rais Y, Viukov S, Geula S, Buenrostro JD, Weinberger L, Krupalnik V et al. 2015. Transient acquisition of pluripotency during somatic cell transdifferentiation with iPSC reprogramming factors. *Nat Biotechnol* **33**(7): 769-774.

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**(7469): 59-+.

Ng HH, Robert F, Young RA, Struhl K. 2003. Targeted recruitment of set1 histone methylase by elongating pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell* **11**(3): 709-719.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**(7398): 381-385.

Paul MLS, Kaur A, Geete A, Sobhia ME. 2014. Essential gene identification and drug target prioritization in Leishmania species. *Mol Biosyst* **10**(5): 1184-1195.

Rancati G, Moffat J, Typas A, Pavelka N. 2017. Emerging and evolving concepts in gene essentiality. *Nat Rev Genet*.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**(7): 1665-1680.

Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C et al. 2003. Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. *Mol Microbiol* **50**(1): 167-181.

Rossant J, Tam PPL. 2009. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development* **136**(5): 701-713.

Scardoni G, Petterlini M, Laudanna C. 2009. Analyzing biological network parameters with CentiScaPe. *Bioinformatics* **25**(21): 2857-2859.

Schier AF. 2007. The maternal-zygotic transition: death and birth of RNAs. *Science* **316**(5823): 406-407.

Schmitt AD, Hu M, Jung I, Xu Z, Qiu YJ, Tan CL, Li Y, Lin S, Lin YI, Barr CL et al. 2016. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep* **17**(8): 2042-2059.

Schroeder MP, Rubio-Perez C, Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. 2014. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. *Bioinformatics* **30**(17): I549-I555.

Schultz RM. 2002. The molecular foundations of the maternal to zygotic transition in the preimplantation embryo. *Human reproduction update* **8**(4): 323-331.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* **148**(3): 458-472.

Shapiro JA, von Sternberg R. 2005. Why repetitive DNA is essential to genome function. *Biol Rev* **80**(2): 227-250.

Shore D, Bianchi A. 2009. Telomere length regulation: coupling DNA end processing to feedback regulation of telomerase. *Embo Journal* **28**(16): 2309-2322.

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015. *Institute for Systems Biology http://repeatmasker org*.

Song N, Liu J, An SC, Nishino T, Hishikawa Y, Koji T. 2011. Immunohistochemical Analysis of Histone H3 Modifications

in Germ Cells during Mouse Spermatogenesis. *Acta Histochem Cytoc* **44**(4): 183-190.

Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, de Laat W. 2006. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development* **20**(17): 2349-2354.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**(43): 15545-15550.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP et al. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**(D1): D447-D452.

Tu ZD, Argmann C, Wong KK, Mitnaul LJ, Edwards S, Sach IC, Zhu J, Schadt EE. 2009. Integrating siRNA and protein-protein interaction data to identify an expanded insulin signaling network. *Genome Res* **19**(6): 1057-1067.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou SB, Diaz LA, Kinzler KW. 2013. Cancer Genome Landscapes. *Science* **339**(6127): 1546-1558.

Wang L, Tu ZD, Sun FZ. 2009. A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. *Bmc Genomics* **10**.

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015. Identification and characterization of essential genes in the human genome. *Science* **350**(6264): 1096-1101.

Wang Z, Moult J. 2001. SNPs, protein structure, and disease. *Hum Mutat* **17**(4): 263-270.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *P Natl Acad Sci USA* **106**(18): 7273-7280.

Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, Zhang B, Liu B, Wang Q, Xia W et al. 2016. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**(7609): 652-657.

Xie W, Schultz MD, Lister R, Hou ZG, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D et al. 2013. Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. *Cell* **153**(5): 1134-1148.

Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, Srinivasan P, Gao J, Chakravarty D, Devlin SM et al. 2017. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine*.

Zernicka-Goetz M. 2006. The first cell-fate decisions in the mouse embryo: destiny is a matter of both chance and choice. *Curr Opin Genet Dev* **16**(4): 406-412.

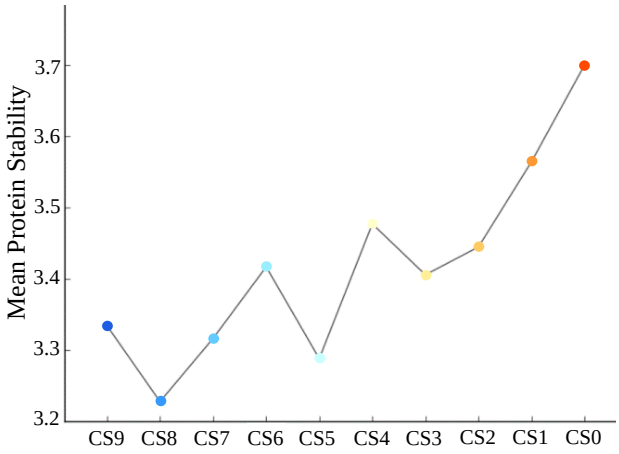
Zernicka-Goetz M, Morris SA, Bruce AW. 2009. Making a firm decision: multifaceted regulation of cell fate in the early mouse embryo. *Nat Rev Genet* **10**(7): 467-477.

Zhang B, Zheng H, Huang B, Li W, Xiang Y, Peng X, Ming J, Wu X, Zhang Y, Xu Q et al. 2016. Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature* **537**(7621): 553-557.

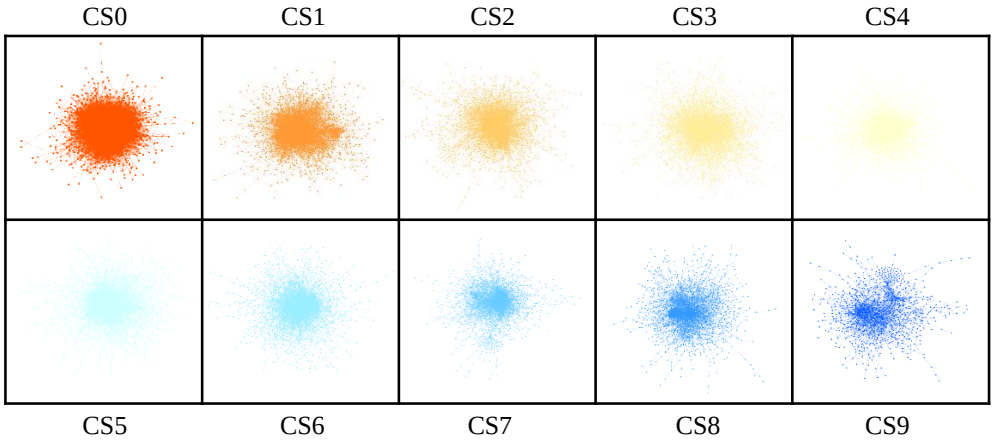
Zhao Y, Li H, Fang SS, Kang Y, Wu W, Hao YJ, Li ZY, Bu DC, Sun NH, Zhang MQ et al. 2016. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res* **44**(D1): D203-D208.

Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL et al. 2013. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**(3): 642-654.

A



B



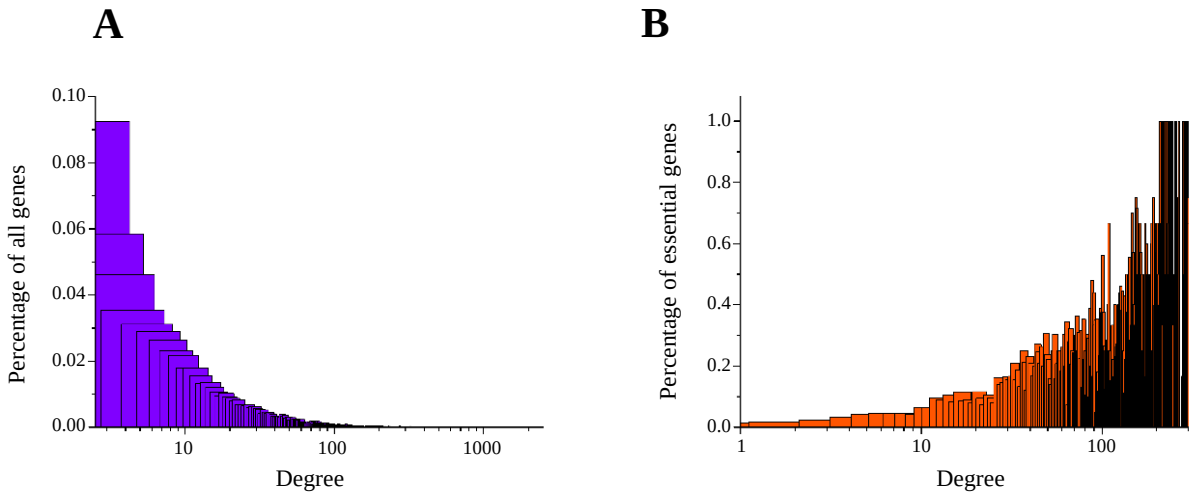
Supplementary Fig. 1. Protein characteristics.

A, Stability of proteins encoded by genes in different gene groups.

B, PPI networks of 10 groups of genes.

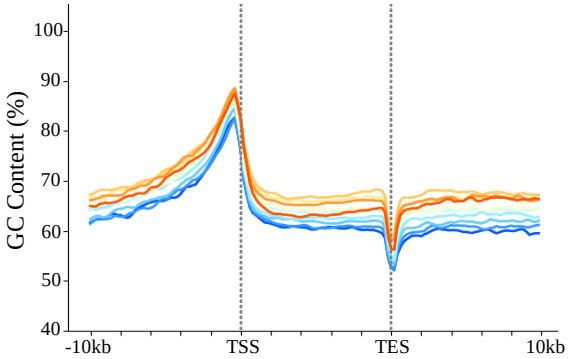
GO term	CS1	CS2	CS3	CS4	CS5	CS6	CS7	CS8	CS9	CS10
rRNA processing	104.48	0	0	0	0	0	0	0	0	0
translational initiation	76.33	0	0	0	0	0	0	0	0	0
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	66.16	0	0	0	0	0	0	0	0	0
mRNA splicing, via spliceosome	65.48	8.03	0	0	0	0	0	0	0	0
translation	64.89	0	0	0	0	0	0	0	0	0
viral transcription	62.88	0	0	0	0	0	0	0	0	0
SRP-dependent cotranslational protein targeting to membrane	60.56	0	0	0	0	0	0	0	0	0
mitochondrial translational elongation	57.98	0	0	0	0	0	0	0	0	0
DNA replication	56.53	0	0	0	0	0	0	0	0	0
mitochondrial translational termination	55.89	0	0	0	0	0	0	0	0	0
transcription-coupled nucleotide-excision repair	52.95	0	0	0	0	0	0	0	0	0
mitochondrial respiratory chain complex I assembly	0	11.63	0	0	0	0	0	0	0	0
mitochondrial electron transport, NADH to ubiquinone	0	9.19	0	0	0	0	0	0	0	0
transcription, DNA-templated	0	4.75	0	0	0	0	0	0	0	0
DNA repair	0	4.50	0	0	0	0	0	0	0	0
transcription elongation from RNA polymerase II promoter	0	4.49	0	0	0	0	0	0	0	0
snRNA transcription from RNA polymerase II promoter	0	4.25	0	0	0	0	0	0	0	0
transcription initiation from RNA polymerase II promoter	0	4.19	0	0	0	0	0	0	0	0
protein import into peroxisome matrix	0	4.00	0	0	0	0	0	0	0	0
mitotic nuclear division	0	3.95	0	0	0	0	0	0	0	0
cell division	0	3.78	0	0	0	0	0	0	0	0
negative regulation of neuron apoptotic process	0	0	3.74	0	0	0	0	0	0	0
transport	0	0	0	4.49	0	0	0	0	0	0
extracellular matrix disassembly	0	0	0	3.31	0	0	0	0	0	0
positive regulation of GTPase activity	0	0	0	0	5.97	0	0	0	0	0
vascular endothelial growth factor receptor signaling pathway	0	0	0	0	3.57	0	0	0	0	0
intracellular signal transduction	0	0	0	0	3.22	0	0	0	0	0
regulation of GTPase activity	0	0	0	0	3.10	0	0	0	0	0
regulation of cell proliferation	0	0	0	0	3.01	0	0	0	0	0
cell fate commitment	0	0	0	0	0	4.41	0	0	0	0
positive regulation of ERK1 and ERK2 cascade	0	0	0	0	0	3.60	0	0	0	0
cell adhesion	0	0	0	0	0	0	4.39	0	0	0
inflammatory response	0	0	0	0	0	0	4.00	0	0	0
lipid catabolic process	0	0	0	0	0	0	3.62	0	0	0
negative regulation of inflammatory response	0	0	0	0	0	0	3.57	0	0	0
membrane repolarization during ventricular cardiac muscle cell action potential	0	0	0	0	0	0	3.52	0	0	0
potassium ion export	0	0	0	0	0	0	3.20	0	0	0
cell-cell signaling	0	0	0	0	0	0	3.19	0	0	0
homophilic cell adhesion via plasma membrane adhesion molecules	0	0	0	0	0	0	0	0	3.70	0
detection of chemical stimulus involved in sensory perception of bitter taste	0	0	0	0	0	0	0	0	0	3.92

Supplementary Fig. 2. GO analysis of 10 groups of genes.

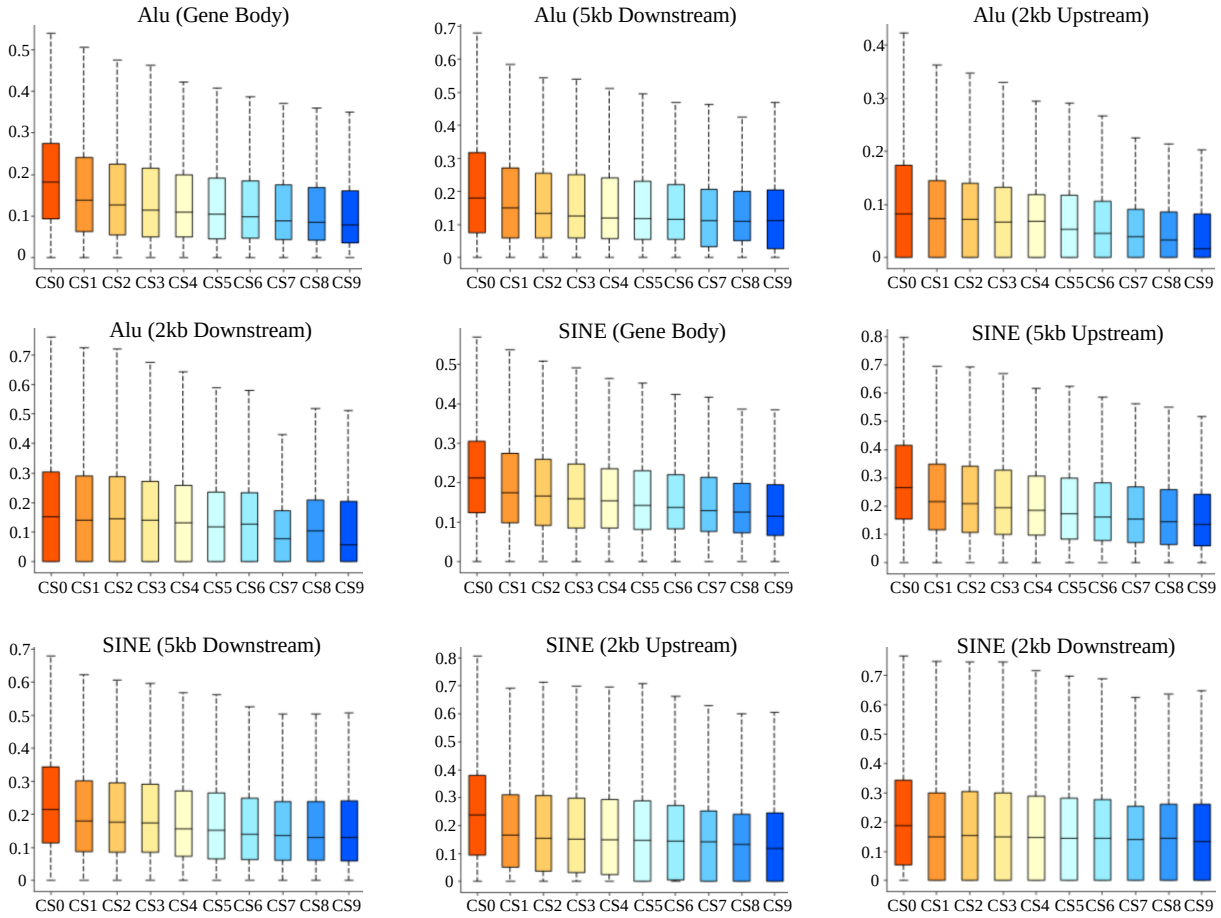


Supplementary Fig. 3. Relationship between the degree of connectivity and the number of all genes (A) and the number of essential genes (B).

A


GC Content (%)
100
90
80
70
60
50
40
-10kb
TSS
TES
10kb

B

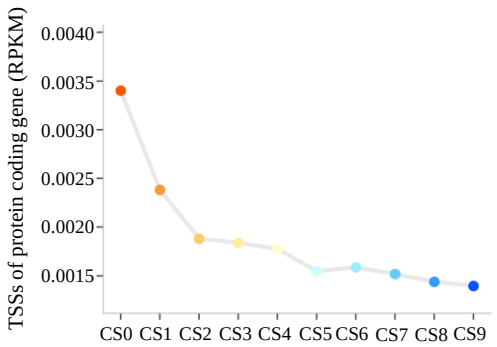

Alu (Gene Body)
Alu (5kb Downstream)
Alu (2kb Upstream)
Alu (2kb Downstream)
SINE (Gene Body)
SINE (5kb Upstream)
SINE (5kb Downstream)
SINE (2kb Upstream)
SINE (2kb Downstream)
CS0 CS1 CS2 CS3 CS4 CS5 CS6 CS7 CS8 CS9

Supplementary Fig. 4. Sequences characteristics.

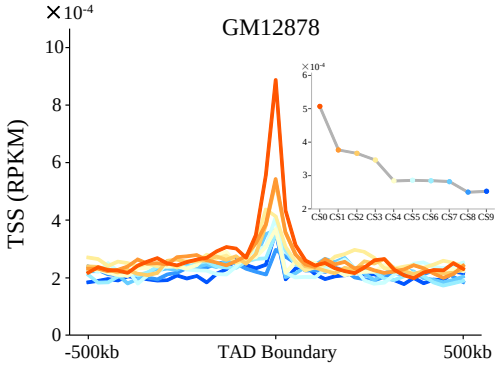
A, GC content of regions 10-kb up- or downstream of gene body. Red line: human essential genes. Color of gene group changes from red to blue with increasing CS value (decreasing essentiality).

B, Repetitive elements of 10 groups of genes.

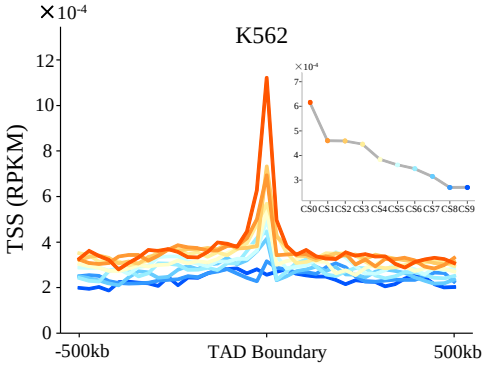
A


TSSs of protein coding gene (RPKM)
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
CS0 CS1 CS2 CS3 CS4 CS5 CS6 CS7 CS8 CS9

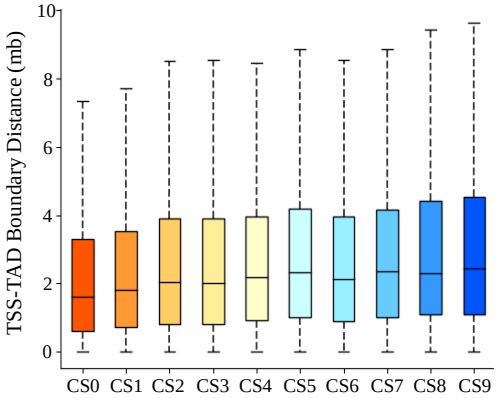
B


GM12878
×10⁻⁴
TSS (RPKM)
-500kb
TAD Boundary
500kb

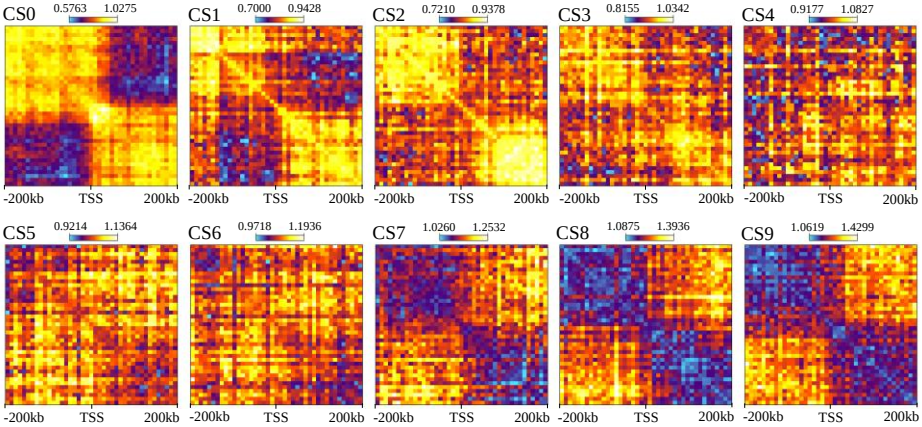
C


K562
×10⁻⁴
TSS (RPKM)
-500kb
TAD Boundary
500kb

D


TSS-TAD Boundary Distance (mb)
CS0 CS1 CS2 CS3 CS4 CS5 CS6 CS7 CS8 CS9

E


CS0 0.5763 1.0275
CS1 0.7000 0.9428
CS2 0.7210 0.9378
CS3 0.8155 1.0342
CS4 0.9177 1.0827
CS5 0.9214 1.1364
CS6 0.9718 1.1936
CS7 1.0260 1.2532
CS8 1.0875 1.3936
CS9 1.0619 1.4299
-200kb TSS 200kb

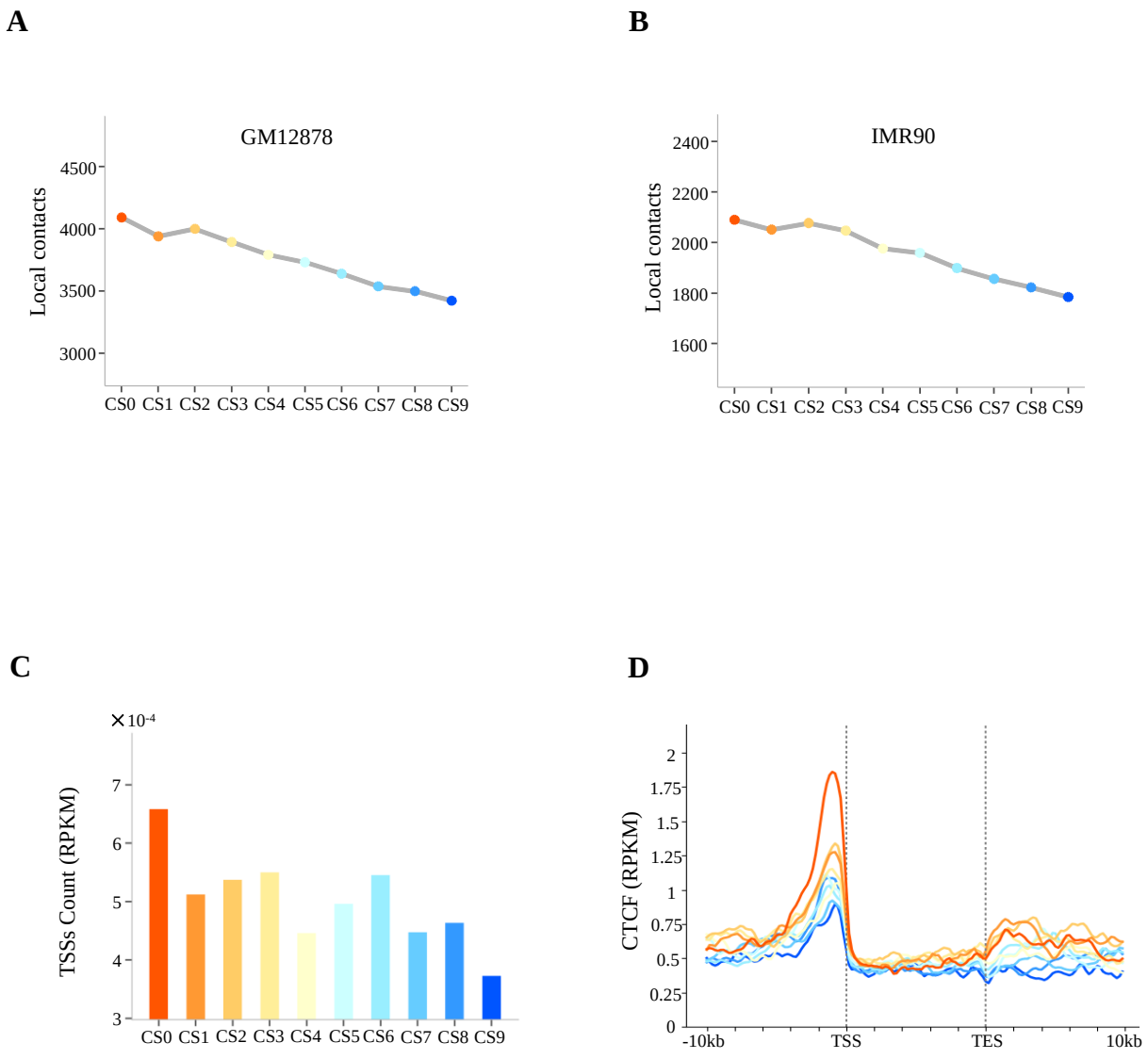
Supplementary Fig. 5. Structural characteristics.

A, Level of overlap between 10 types of gene TSSs and all protein-coding gene TSSs.

B-C, Profiles of 10 types of TSSs near TAD boundary of **(B)** GM12878 or **(C)** K562. Inset: mean values of TAD boundary and near-boundary regions (≤100kb).

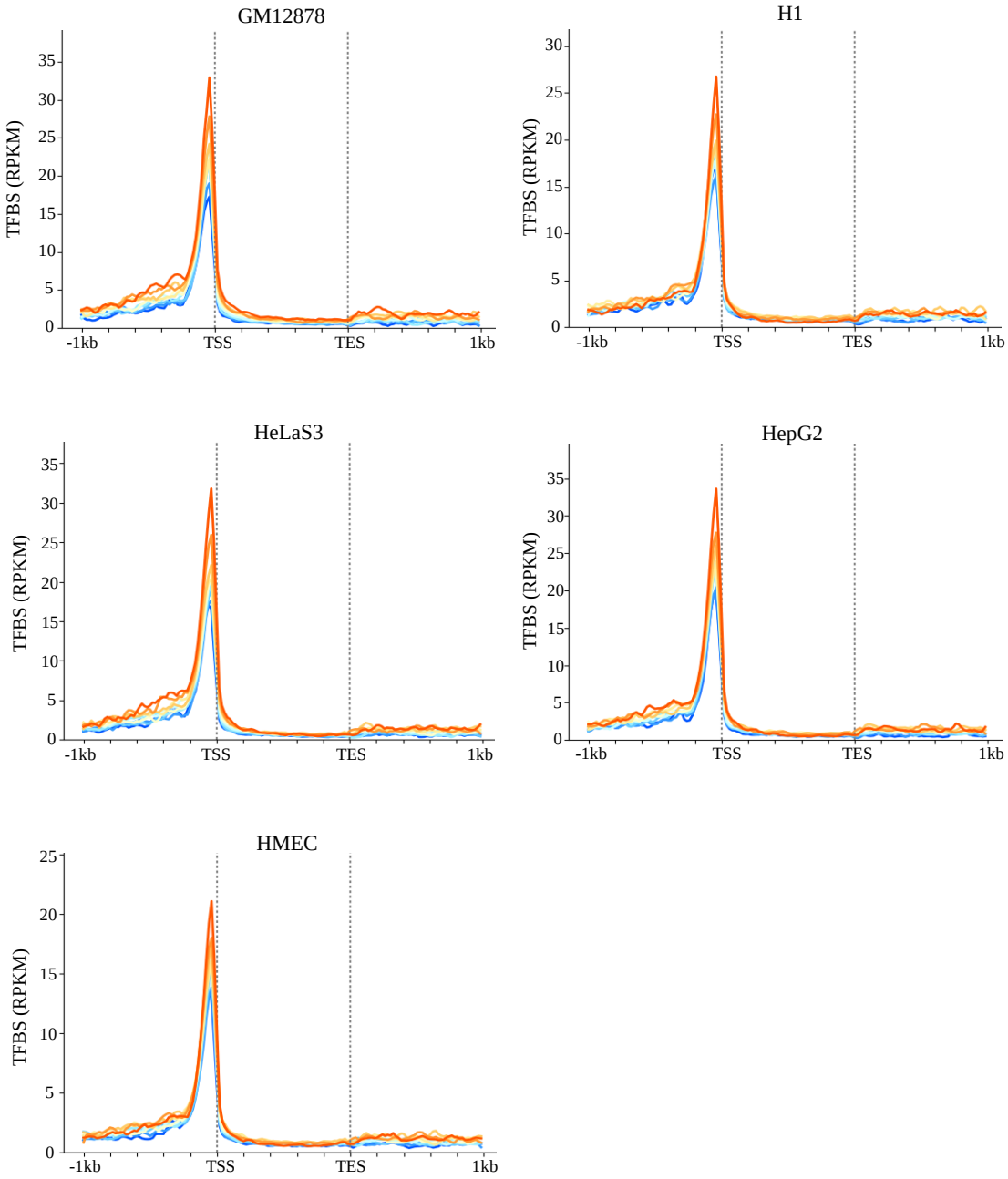
D, Average distance between gene and nearest TAD boundary.

E, Mean Hi-C contact map near 10 types of TSSs. Value for each bin was normalized by dividing by the mean value of all bins of the same linear genomic distance.

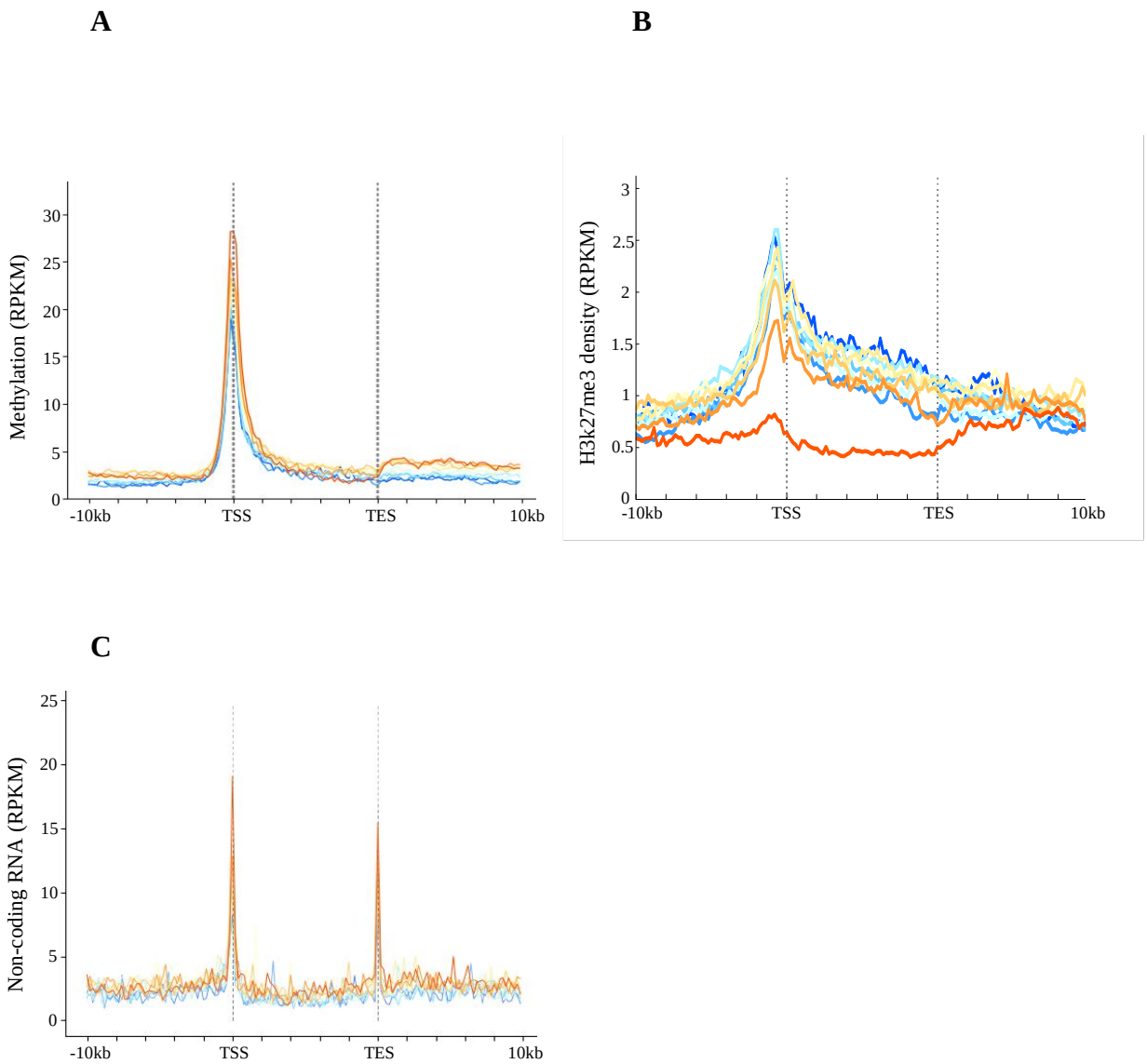


Supplementary Fig. 6. 3D structure

A-B, Mean intra-TAD local (<100 kb) Hi-C contacts for bins containing 10 types of TSSs in (**A**) GM12878 and (**B**) IMR90. For each bin containing a TSS, only contacts with bins in the opposite direction to the nearest TAD boundary were counted. **C**, Level of overlap between 10 types of gene TSSs and GM12878 chromatin loop anchors. **D**, ChiP-seq signal of CTCF in region 10-kb up- or downstream of gene body.

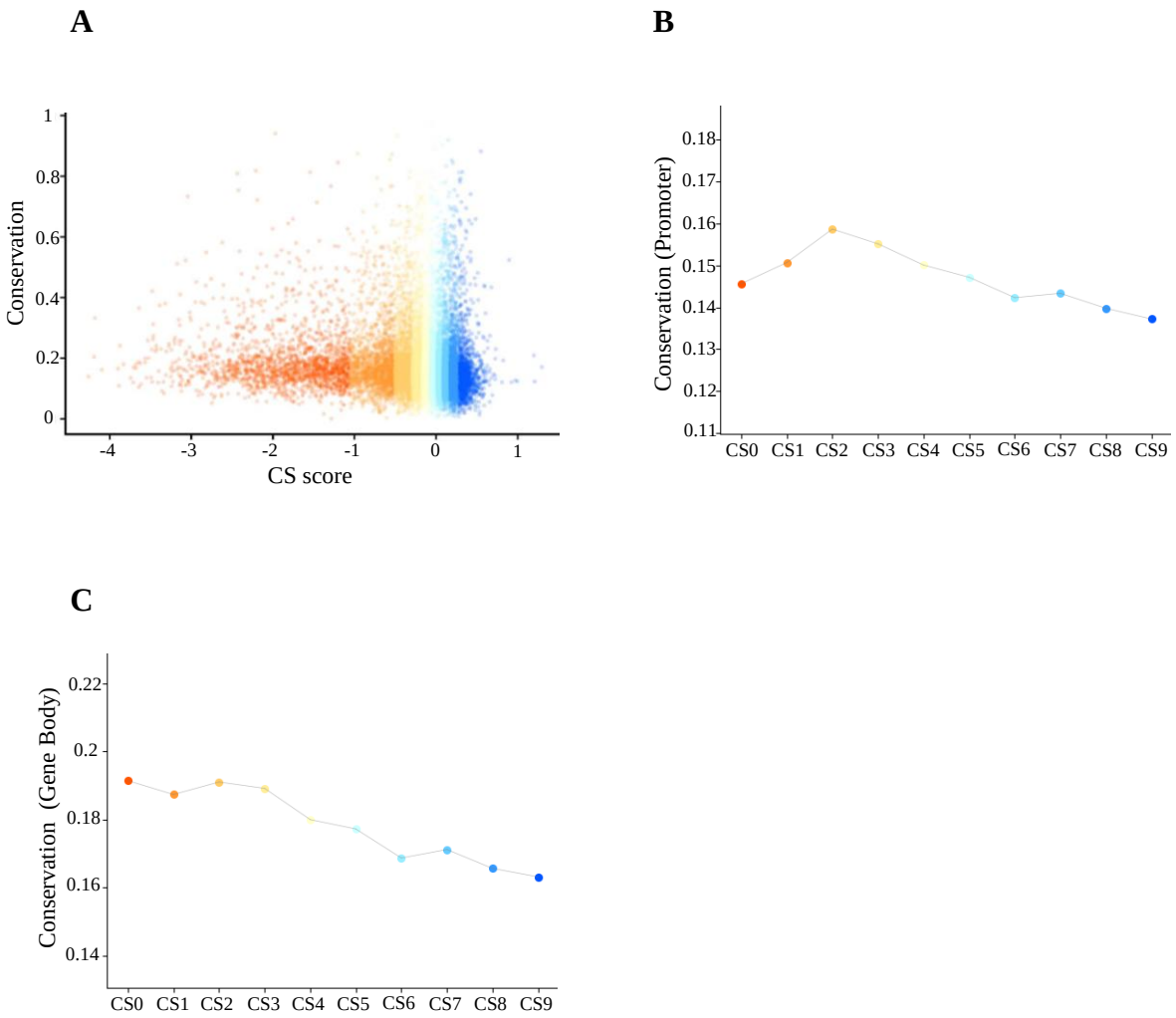


Supplementary Fig. 7. Distributions of transcription factor binding sites in regions 1-kb up- or downstream of gene body from five human cell lines.



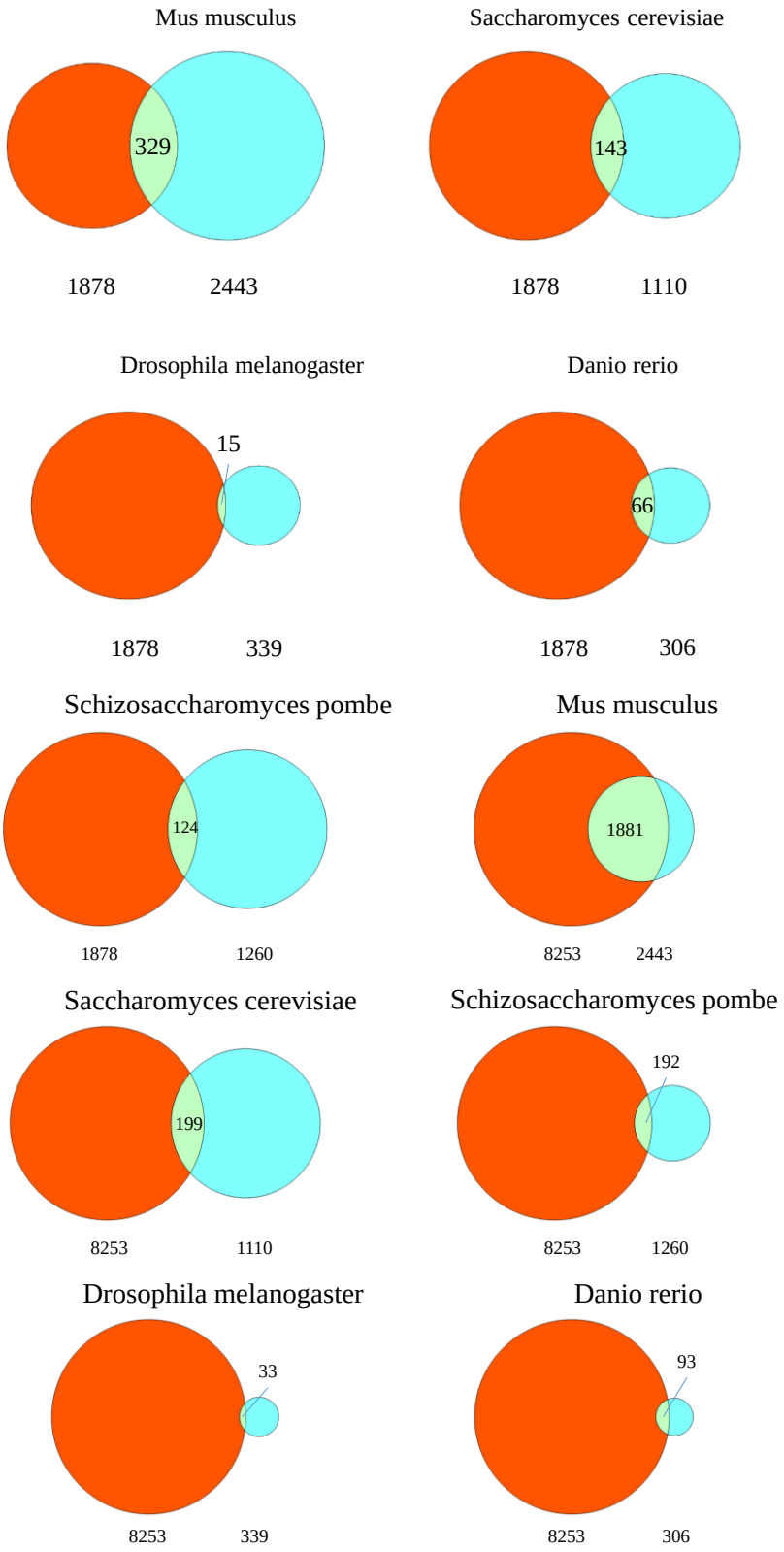
Supplementary Fig. 8. Epigenetic signals.

Methylation level (MRE) (**A**), H3k27me3 density (**B**), and ncRNA distribution (**C**) in regions 10-kb up- or downstream of gene body.

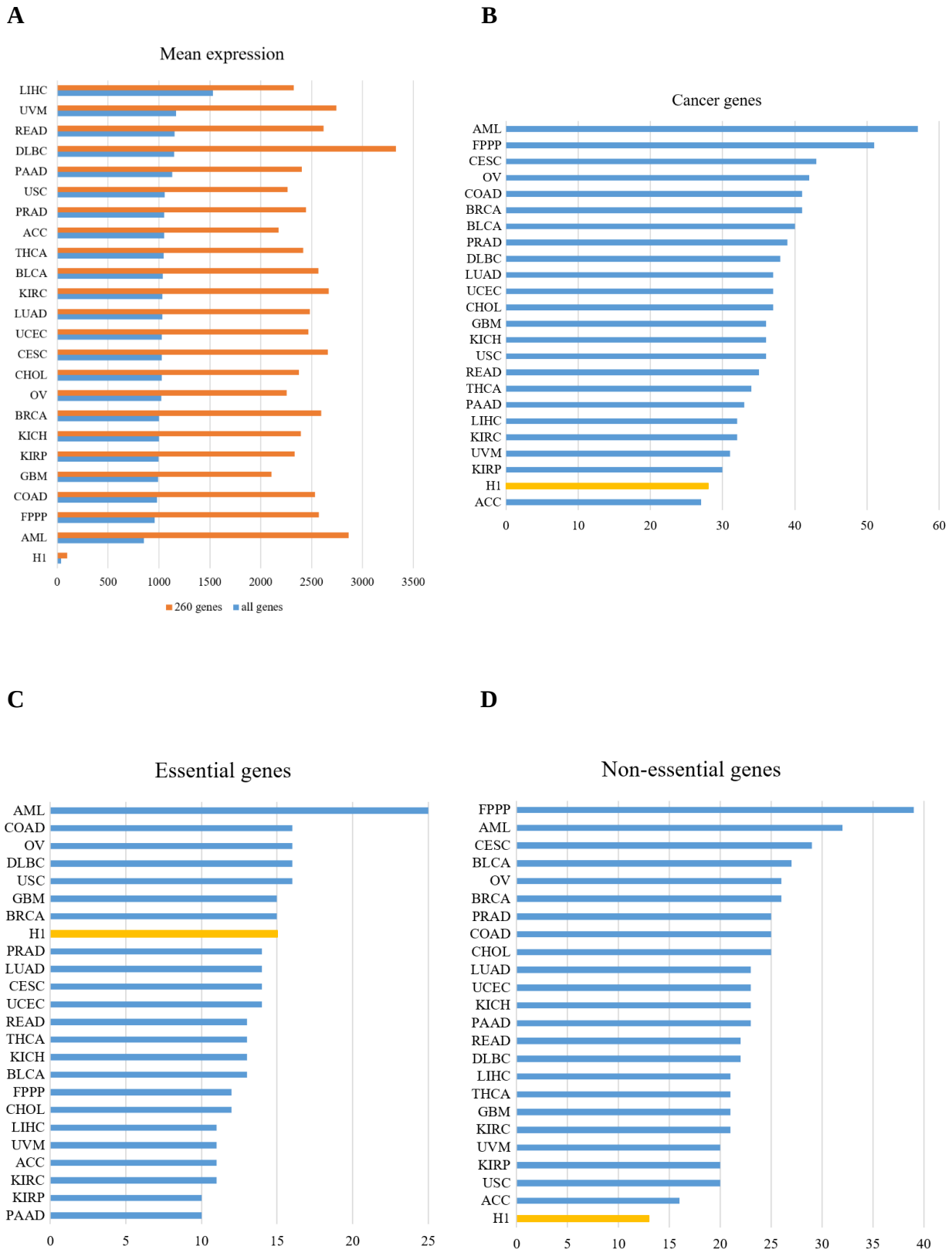


Supplementary Fig. 9. Essentiality and conservation.

A, Relationship between CS value and conservation. **B**, Average conservation of gene promoter region. **C**, Average conservation of gene body regions.



Supplementary Fig. 10. Comparison of homologous essential genes of different species. Two different sets of human essential genes were used for analysis.


A
Mean expression
LIHC
UVM
READ
DLBC
PAAD
USC
PRAD
ACC
THCA
BLCA
KIRC
LUAD
UCEC
CESC
CHOL
OV
BRCA
KICH
KIRP
GBM
COAD
FPPP
AML
H1
0
500
1000
1500
2000
2500
3000
3500
260 genes
all genes
B
Cancer genes
AML
FPPP
CESC
OV
COAD
BRCA
BLCA
PRAD
DLBC
LUAD
UCEC
CHOL
GBM
KICH
USC
READ
THCA
PAAD
LIHC
KIRC
UVM
KIRP
H1
ACC
0
10
20
30
40
50
60
C
Essential genes
AML
COAD
OV
DLBC
USC
GBM
BRCA
H1
PRAD
LUAD
CESC
UCEC
READ
THCA
KICH
BLCA
FPPP
CHOL
LIHC
UVM
ACC
KIRC
KIRP
PAAD
0
5
10
15
20
25
D
Non-essential genes
FPPP
AML
CESC
BLCA
OV
BRCA
PRAD
COAD
CHOL
LUAD
UCEC
KICH
PAAD
READ
DLBC
LIHC
THCA
GBM
KIRC
UVM
KIRP
USC
ACC
H1
0
5
10
15
20
25
30
35
40

Supplementary Fig. 11. A, Average expression of 260 cancer genes and all protein-coding genes in TCGA database and H1 cell line. **B**, Number of cancer genes in the top 1,000 most highly expressed genes. **C**, Number of human essential genes in the top 1,000 most highly expressed genes. **D**, Number of the nonessential genes in the top 1,000 most highly expressed genes.

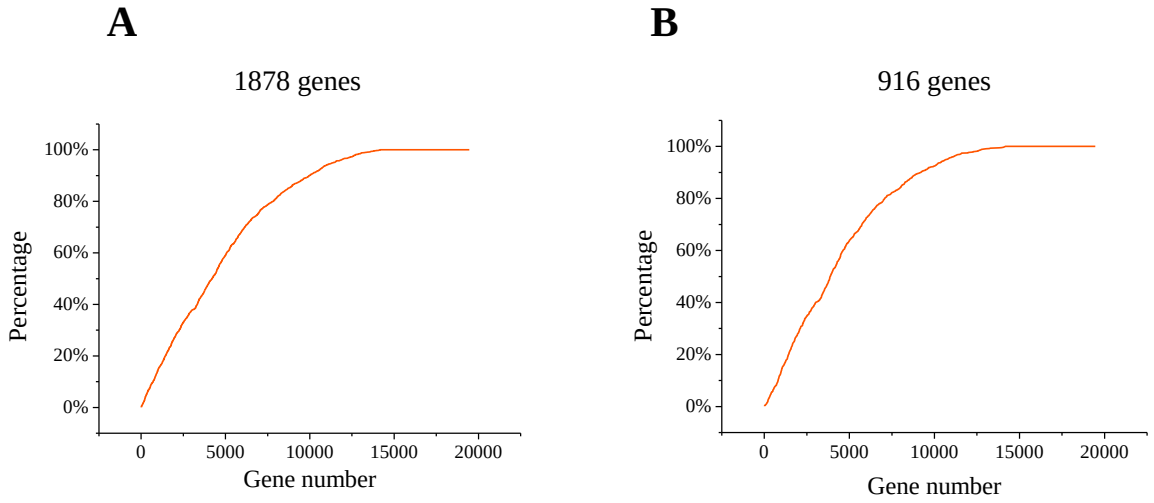


Figure S12. Genes sorted by importance in systems biology. Essential genes use two sets of data: the gene set used in our paper (**A**), and the gene set from the three literatures (**B**). The results are the same.

Supplementary Table 1. A combined analysis of human essential genes with the OMIM disease database. Here is the top 10 disease related with human essential genes.

Index	Name	P-value	Adjusted p-value	Z-score	Combined score
1	anemia	3.66E-07	0.00001612	-1.73	25.68
2	leigh syndrome	0.00001065	0.0002344	-0.84	9.62
3	leukemia	0.005501	0.06051	-1.22	6.34
4	fanconi anemia	0.00006561	0.0009623	-0.33	3.15
5	neuropathy	0.04132	0.303	-0.57	1.8
6	mental retardation	0.9998	0.9998	7.33	0
7	epilepsy	0.9967	0.9998	7.97	-0.03
8	deafness	0.9944	0.9998	7.28	-0.04
9	cataract	0.9903	0.9998	7.92	-0.08
10	ataxia	0.9807	0.9998	6.63	-0.13

Supplementary Table 2. Candidate cancer genes.

SNRPD2	DOLPP1	ANAPC11	GMPS	ATAD5	ECT2	CHAF1A
MAD2L2	PSMB2	SMC2	TOMM40	CHAF1B	CDC7	NR2C2AP
MYBBP1A	SF3A2	RFWD3	NUP205	POLE	EXOSC5	AURKB
RFC5	MOCS3	THOC3	MARS	ABHD11	H2AFX	AURKA
CLNS1A	SLC4A5	LSM7	CCT5	MCM4	TRAIP	RACGAP1
NLE1	RPS2	HMBS	RPP21	NCAPD2	GSDMC	FAM72B
PSMB3	EHMT2	PSMD14	SMG7	DTYMK	XRCC3	CHEK1
ADSS	DDX49	SEC61A1	C12orf45	MCM7	MTHFD2	CDK1
CKAP5	ATR	KAT2A	LIN9	TFAP4	SMC4	EZH2
FDXR	CCT2	MRPL17	SHFM1	DNMT1	INTS8	BIRC5
SRM	ZC3H3	DKC1	EIF3C	CPSF1	NOP2	C16orf59
LSM2	CCAR1	RRS1	SRD5A3	PRIM1	KIF11	XRCC2
EBNA1BP2	HSPD1	IER5L	TUBG1	RFC2	BOP1	CDCA8
GALE	HIST2H4A	NUP85	MRPS23	KIF23	DSCC1	SPC24
POLR2H	RFC3	SNRPD1	PFDN6	RCC1	BRCA1	CENPM
FBL	NAT10	CPSF3	SMARCA4	ALDOA	TOPBP1	NCAPG

RPS19	TBC1D3H	FKBPL	CASC5	RFC4	ESPL1	CENPK
CCT3	SLC38A5	SNRPE	RRP1	TYMS	GINS2	KIF20A
SRPRB	PNKP	RTEL1	PALB2	NOP16	GAPDH	DTL
TBC1D3G	UTP20	POLD2	RAN	PCNA	SNRPB	CLSPN
TIMM10	FAM50A	TUBA1B	TSEN54	SOX12	CENPI	HJURP
TELO2	PFDN2	MASTL	TFRC	DNAJC9	RAD51	PKMYT1
ATP6V1F	NCLN	RAD9A	DHX37	MCM3	BUB1B	MCM2
PPAN	PGM3	TUBB	MCM5	TRAF2	LIG1	SPC25
BANF1	SSBP1	MRPL12	POP7	CDH24	FANCI	CDC45
HEATR1	TXN	UBE2I	HIST1H2BN	RUNX1	FEN1	TOP2A
MRPL9	YKT6	MRPS34	PDRG1	GINS1	CENPW	RRM2
ADRM1	RPN2	GSS	FCGR1B	POLR3K	KIFC2	CEP55
RUVBL2	TUBB8	RPP40	WDR12	GINS3	PRC1	CDT1
HAUS8	RPL8	ACTL6A	GNB1L	NUP155	PSMG3	POLE2
AP2S1	PAICS	ILF2	SNRPF	DONSON	BRCA2	CDCA5
CENPN	YARS	BYSL	MRPS12	GARS	UBE2S	NUF2
AATF	EXOSC4	TKT	LSM4	AZI1	CENPE	CCNA2
CSE1L	HAUS5	MYB	TPI1	DSN1	NCAPH	OIP5
NDOR1	ENO1	PGK1	TTF2	PABPC1	SKA3	CDC20
ACLY	PPAT	XPO5	HMGB3	MCM6	SKA1	MYBL2
NAA10	TIMELESS	RANGAP1	BRIX1	RNASEH2A	SPATS2	CDC6
WDR4	POP1	HIST2H3C	RAE1	KIF18A	PLK4	PLK1
DDX52	VARS	CCT6A	HIST1H2BF	FAM72A	NDC80	TPX2
RUVBL1	CDK7	POLA2	FANCG	HIST2H2AA3	ZWINT	
CCDC86	HSPE1	NUP62	INCENP	MTHFD1L	CENPA	
SCD	NOP56	ATIC	SHMT2	GINS4	CHTF18	
C14orf80	SPAG5	PMM2	ATXN2L	TACC3	FAM72D	

Supplementary Table 3. Candidate drugs.

1	(3AR,6R,6AS)-6-((S)-((S)-CYCLOHEX-2-ENYL)(HYDROXY)METHYL)-6A-METHYL-4-OXO-HEXAHYDRO-2H-FURO[3,2-C]PYRROLE-6-CARBALDEHYDE
2	L-ASPARTIC ACID
3	2'-MONOPHOSPHOADENOSINE 5'-DIPHOSPHORIBOSE
4	URIDINE-DIPHOSPHATE-N-ACETYLGLUCOSAMINE
5	DB04395
6	BORTEZOMIB
7	CARFILZOMIB
8	ANISOMYCIN
9	PUROMYCIN
10	L-TYROSINE
11	TYROSINAL
12	N6-ISOPENTENYL-ADENOSINE-5'-MONOPHOSPHATE
13	DB08617
14	L-GLUTAMINE
15	L-VALINE
16	PHOSPHONOTHREONINE
17	COENZYME A
18	EPOTHILONE B
19	CYT997
20	2-MERCAPTO-N-[1,2,3,10-TETRAMETHOXY-9-OXO-5,6,7,9-TETRAHYDRO-BENZO[A]HEPTALEN-7-YL]ACETAMIDE
21	VINCRISTINE
22	VINBLASTINE
23	COLCHICINE
24	VINORELBINE
25	GLUTATHIONE
26	GLYCINE
27	L-CYSTEINE
28	GAMMA-GLUTAMYLCYSTEINE
29	ADENOSINE-5'-[BETA, GAMMA-METHYLENE]TRIPHOSPHATE
30	3-PHOSPHOGLYCERIC ACID
31	DACARBAZINE
32	TETRAHYDROFOLIC ACID
33	PEMETREXED
34	5--MONOPHOSPHATE-9-BETA-D-RIBOFURANOSYL XANTHINE
35	BETA-DADF, MSA, MULTISUBSTRATE ADDUCT INHIBITOR
36	L-GLUTAMIC ACID
37	PHOSPHOGLYCOLOHYDROXAMIC ACID
38	[2(FORMYL-HYDROXY-AMINO)-ETHYL]-PHOSPHONIC ACID

39	PYRIDOXAL PHOSPHATE
40	CLADRIBINE
41	DB03280
42	3'-AZIDO-3'-DEOXYTHYMIDINE-5'-MONOPHOSPHATE
43	P1-(5'-ADENOSYL)P5-(5'-(3'AZIDO-3'-DEOXYTHYMIDYL))PENTAPHOSPHATE
44	DECITABINE
45	PROCAINAMIDE
46	1,6-FRUCTOSE DIPHOSPHATE (LINEAR FORM)
47	CAPECITABINE
48	FLOXURIDINE
49	LEUCOVORIN
50	RALTITREXED
51	6-ETHYL-5-PHENYLPYRIMIDINE-2,4-DIAMINE
52	N-(3,5-DIMETHOXYPHENYL)IMIDODICARBONIMIDIC DIAMIDE
53	10-PROPARGYL-5,8-DIDEAZAFOLIC ACID
54	S,S-(2-HYDROXYETHYL)THIOCYSTEINE
55	GEMCITABINE
56	TRIMETHOPRIM
57	TRIFLURIDINE
58	(5S)-5-(3-AMINOPROPYL)-3-(2,5-DIFLUOROPHENYL)-N-ETHYL-5-PHENYL-4,5-DIHYDRO-1H-PYRAZOLE-1-CARBOXAMIDE
59	(2S)-4-(2,5-DIFLUOROPHENYL)-N,N-DIMETHYL-2-PHENYL-2,5-DIHYDRO-1H-PYRROLE-1-CARBOXAMIDE
60	(2S)-4-(2,5-DIFLUOROPHENYL)-N-METHYL-2-PHENYL-N-PIPERIDIN-4-YL-2,5-DIHYDRO-1H-PYRROLE-1-CARBOXAMIDE
61	(5R)-N,N-DIETHYL-5-METHYL-2-[(THIOPHEN-2-YLCARBONYL)AMINO]-4,5,6,7-TETRAHYDRO-1-BENZOTHIOPHENE-3-CARBOXAMIDE
62	MONASTROL
63	3-[(5S)-1-ACETYL-3-(2-CHLOROPHENYL)-4,5-DIHYDRO-1H-PYRAZOL-5-YL]PHENOL
64	ADENOSINE-5-DIPHOSPHORIBOSE
65	4-(2-AMINOETHYL)BENZENESULFONYL FLUORIDE
66	BLEOMYCIN
67	AT9283
68	HESPERIDIN
69	4-(4-METHYLPIPERAZIN-1-YL)-N-[5-(2-THIENYLACETYL)-1,5-DIHYDROPYRROLO[3,4-C]PYRAZOL-3-YL]BENZAMIDE
70	N-[3-(1H-BENZIMIDAZOL-2-YL)-1H-PYRAZOL-4-YL]BENZAMIDE
71	2-[(5,6-DIPHENYLFURO[2,3-D]PYRIMIDIN-4-YL)AMINO]ETHANOL
72	18-CHLORO-11,12,13,14-TETRAHYDRO-1H,10H-8,4-(AZENO)-9,15,1,3,6-BENZODIOXATRIAZACYCLOHEPTADECIN-2-ONE
73	1-(5-CHLORO-2,4-DIMETHOXYPHENYL)-3-(5-CYANOPYRAZIN-2-YL)UREA

74	(2R)-1-[(5,6-DIPHENYL-7H-PYRROLO[2,3-D]PYRIMIDIN-4-YL)AMINO]PROPAN-2-OL
75	2-[5,6-BIS-(4-METHOXY-PHENYL)-FURO[2,3-D]PYRIMIDIN-4-YLAMINO]-ETHANOL
76	DB07243
77	REL-(9R,12S)-9,10,11,12-TETRAHYDRO-9,12-EPOXY-1H-DIINDOLO[1,2,3-FG:3',2',1'-KL]PYRROLO[3,4-I][1,6]BENZODIAZOCINE-1,3(2H)-DIONE
78	1-[(2S)-4-(5-PHENYL-1H-PYRAZOLO[3,4-B]PYRIDIN-4-YL)MORPHOLIN-2-YL]METHANAMINE
79	N-(4-OXO-5,6,7,8-TETRAHYDRO-4H-[1,3]THIAZOLO[5,4-C]AZEPIN-2-YL)ACETAMIDE
80	2-(METHYLSULFANYL)-5-(THIOPHEN-2-YLMETHYL)-1H-IMIDAZOL-4-OL
81	1-[(2S)-4-(5-BROMO-1H-PYRAZOLO[3,4-B]PYRIDIN-4-YL)MORPHOLIN-2-YL]METHANAMINE
82	1-(5-CHLORO-2-METHOXYPHENYL)-3-{6-[2-(DIMETHYLAMINO)-1-METHYLETHOXY]PYRAZIN-2-YL}UREA
83	(5-{3-[5-(PIPERIDIN-1-YLMETHYL)-1H-INDOL-2-YL]-1H-INDAZOL-6-YL}-2H-1,2,3-TRIAZOL-4-YL)METHANOL
84	2-(CYCLOHEXYLAMINO)BENZOIC ACID
85	(2S)-1-AMINO-3-[(5-NITROQUINOLIN-8-YL)AMINO]PROPAN-2-OL
86	2,2'-{[9-(HYDROXYIMINO)-9H-FLUORENE-2,7-DIYL]BIS(OXY)}DIACETIC ACID
87	N-{5-[4-(4-METHYLPIPERAZIN-1-YL)PHENYL]-1H-PYRROLO[2,3-B]PYRIDIN-3-YL}NICOTINAMIDE
88	SU9516
89	ALSTERPAULLONE
90	HYMENIALDISINE
91	AT7519
92	OLOMOUCINE
93	INDIRUBIN-3'-MONOXIME
94	DOXORUBICIN
95	TENIPOSIDE
96	VALRUBICIN
97	IDARUBICIN
98	EPIRUBICIN
99	MITOXANTRONE
100	DAUNORUBICIN
101	AMONAFIDE
102	ELSAMITRUCIN
103	BANOXANTRONE
104	GENISTEIN
105	FINAFLOXACIN
106	LUCANTHONE
107	FLEROXACIN
108	SPARFLOXACIN

109	OFLOXACIN
110	LEVOFLOXACIN
111	NORFLOXACIN
112	LOMEFLOXACIN
113	TROVAFLOXACIN
114	CIPROFLOXACIN
115	PEFLOXACIN
116	ENOXACIN
117	DEXRAZOXANE
118	AMSACRINE
119	MOTEXAFIN GADOLINIUM
120	IMEXON
121	GALLIUM NITRATE
122	4-[(7-OXO-7H-THIAZOLO[5,4-E]INDOL-8-YLMETHYL)-AMINO]-N-PYRIDIN-2-YL-BENZENESULFONAMIDE
123	2-ANILINO-6-CYCLOHEXYLMETHOXYPURINE
124	(2S)-N-[(3Z)-5-CYCLOPROPYL-3H-PYRAZOL-3-YLIDENE]-2-[4-(2-OXOIMIDAZOLIDIN-1-YL)PHENYL]PROPANAMIDE
125	6-CYCLOHEXYLMETHOXY-2-(3'-CHLOROANILINO) PURINE
126	5-[5,6-BIS(METHYLOXY)-1H-BENZIMIDAZOL-1-YL]-3-{[1-(2-CHLOROPHENYL)ETHYL]OXY}-2-THIOPHENECARBOXAMIDE
127	N-[4-(2,4-DIMETHYL-THIAZOL-5-YL)-PYRIMIDIN-2-YL]-N',N'-DIMETHYL-BENZENE-1,4-DIAMINE
128	1-(3,5-DICHLOROPHENYL)-5-METHYL-1H-1,2,4-TRIAZOLE-3-CARBOXYLIC ACID
129	HYDROXY(OXO)(3-{[(2Z)-4-[3-(1H-1,2,4-TRIAZOL-1-YLMETHYL)PHENYL]PYRIMIDIN-2(5H)-YLIDENE]AMINO}PHENYL)AMMONIUM
130	4-METHYL-5-{(2E)-2-[(4-MORPHOLIN-4-YLPHENYL)IMINO]-2,5-DIHYDROPYRIMIDIN-4-YL}-1,3-THIAZOL-2-AMINE
131	4-(6-CYCLOHEXYLMETHOXY-9H-PURIN-2-YLAMINO)--BENZAMIDE
132	3-(6-CYCLOHEXYLMETHOXY-9H-PURIN-2-YLAMINO)-BENZENESULFONAMIDE
133	3-({2-[(4-{[6-(CYCLOHEXYLMETHOXY)-9H-PURIN-2-YL]AMINO}PHENYL)SULFONYL]ETHYL}AMINO)PROPAN-1-OL
134	4-{[4-AMINO-6-(CYCLOHEXYLMETHOXY)-5-NITROSOPYRIMIDIN-2-YL]AMINO}BENZAMIDE
135	DB06963