

# The atlas of human cardiac promoters and enhancers reveals an important role for regulatory elements in heart failure

**Ruslan Deviatiiarov**

Kazan State University

**Anna Gams**

George Washington University

**Roman Syunyaev**

Moscow Institute of Physics and Technology

**Tatiana Tatarinova**

University of La Verne

**Ramesh Singh**

Inova Heart and Vascular Institute

**Palak Shah**

Inova Heart and Vascular Institute

**Oleg Gusev**

RIKEN Center for Integrative Medical Sciences

**Igor Efimov** (✉ [efimov@gwu.edu](mailto:efimov@gwu.edu))

George Washington University <https://orcid.org/0000-0002-1483-5039>



---

## Article

**Keywords:** heart disease, regulatory elements, heart failure

**Posted Date:** October 28th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-1010746/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Nature Cardiovascular Research on January 16th, 2023. See the published version at <https://doi.org/10.1038/s44161-022-00182-x>.

# Abstract

A continuous increase in the prevalence of heart failure and the lack of adequate therapy highlight poor understanding of the underlying genetic regulatory mechanisms involved in heart failure pathogenesis. Growing evidence has demonstrated a significant contribution of non-coding genome regulatory elements towards transcriptomic changes in heart disease. Thus, there is a pressing need for a comprehensive resource of the human cardiac regulatory network in healthy and failing states. We applied cap analysis of gene expression sequencing to directly measure the expression of RNA associated with enhancers and promoters. Based on this data, we constructed the atlas of transcribed cardiac regulatory elements from 21 healthy and 10 failing (ischemic and non-ischemic cardiomyopathy) human hearts. In total, we have sequenced 109 samples from the left and right atria and ventricles, identifying 17,668 promoters and 14,920 enhancers associated with 14,519 genes. Leveraging this atlas, we provide insights into functional and structural regulatory changes between healthy and failing hearts. Healthy atria and ventricles had distinct pathway enrichment and transcription factor binding patterns, significantly remodeled by heart failure. Using the advantages of deep sequencing that allow effective analysis of *cis*-regulatory elements-derived RNA, we found that heart failure is associated with the expression of transcripts derived from alternative promoters and a specific set of transcribed enhancers. Furthermore, we identified a high prevalence of single nucleotide polymorphisms associated with cardiovascular diseases within the regulatory regions highlighting their importance in disease pathogenesis. This open-source atlas will serve the cardiovascular community to improve understanding cardiac regulatory network and facilitate the development of novel therapeutics.

## Introduction

Cardiovascular disease is the leading cause of morbidity and mortality in the United States and worldwide. Gene expression, which governs protein production and other cell processes, is aberrant in heart failure<sup>1</sup>. The network of regulatory elements, such as promoters and enhancers, modulates the gene expression by interacting with transcription factors in a 3D chromatin space<sup>2,3</sup>. Both promoters and enhancers are *cis*-regulatory elements meaning they are located on the same chromosome and near their target gene<sup>3</sup>. Promoters usually contain a transcription start site (TSS), transcription factor binding site (TFBS), and TATA-box which facilitate gene transcription<sup>3</sup>. Enhancers are short regions of DNA that bind transcription factors and co-activators which then interact with gene promoters to enhance the rate of gene expression<sup>4</sup>.

The expression of every gene is regulated by several regulatory elements in an upstream region. Due to the complexity of the regulatory mechanism, precise genomic locations of promoters and enhancers are not yet well characterized. Mutations in promoter and enhancer regions can change the transcription factor binding affinity, resulting in altered gene expression<sup>2</sup>. Co-localization of regulatory elements with cardiovascular disease-associated variants can explain those transcriptional changes<sup>2</sup>. A comprehensively annotated library of cardiac regulatory elements can give meaningful insights into clinically relevant genes and potentially explain the causes of gene expression remodeling during disease pathogenesis.

Cap analysis of gene expression (CAGE) is a high throughput RNA expression analysis approach for TSS identification at one base pair (bp) resolution. This approach is used in FANTOM5 and ENCODE projects for

accurate 5' ends annotation in human, mouse, and other plant and animal genomes<sup>5</sup>. Deep sequencing of non-amplified CAGE libraries allows for precise estimation of expression of coding and non-coding transcripts. Moreover, since the number of mapped tags directly depends on the mRNA amount in the sample, this powerful approach is also useful for differential expression analysis of the mapped clusters. The CAGE technique allows identifying alternative promoters that might affect the final protein structure or even estimate short TSS shifts linked to regulatory network features<sup>6,7</sup>. Enhancers are identified by bidirectional CAGE signal profiles, which are usually functional at a much higher frequency than non-transcribed enhancers if compared with DNase I hypersensitive sites or histone modification signals<sup>8</sup>. The bidirectional profile of active enhancers occurs due to RNA transcription from the center of enhancer to opposite directions<sup>4</sup>. Transcribed enhancer RNAs interact with several regulatory complexes and are transient in nature. Such behavior requires higher sequencing depth to detect low expression of enhancer RNA. Additionally, the transcribed enhancers accumulate causative mutations and could be associated with promoters within a complex regulatory network<sup>8</sup>.

We sequenced myocardial tissue samples from healthy donor hearts and hearts from ischemic and non-ischemic cardiomyopathy (ICM and NICM) patients obtained at the time of heart transplantation to utilize CAGE sequencing for regulatory network detection and its remodeling. Despite the common manifestations of the diseases during end stage such as ventricular dilation, myocardial fibrosis, and reduced myocardial contractility, genetic differentiation between ICM and NICM could be essential for prognosis and tailored pharmacological therapy<sup>9</sup>. Microarray studies showed genes related to immunologic processes are activated in NICM while immediate-early response genes and cytokine pathways are upregulated in ICM<sup>9</sup>. Yet, those studies only had a subset of genes, and it remains unclear what governs those transcriptional changes.

Here, we present the comprehensive atlas of promoters and enhancers from four cardiac chambers: left and right atria and ventricles of 21 healthy and 10 failing human hearts. This annotated atlas presents 17,668 promoters and 14,920 enhancers associated with 14,519 genes expressed in the human heart. We developed a robust classification pipeline and identified 351 novel promoters and 1,318 novel enhancers not previously described. Furthermore, we identified cardiac chamber-specific differential regulatory element usage between healthy and failing hearts and between ICM and NICM etiology of heart failure. Using the genome-wide association studies (GWAS) database of common cardiac diseases, we located 1,831 single nucleotide polymorphisms (SNPs) in regulatory regions. We identified multiple cases where nucleotide change significantly affects the DNA binding motif of a transcription factor.

## Results

# Map of the human heart genome regulatory network

Our atlas contains 109 individual CAGE libraries prepared from 31 healthy and failing hearts (Figure 1A). We aligned 1025M reads to hg38 genome assembly total with an average mapping ratio of 97.7 % (Figure 1B). Most of the mapped CAGE tags were in promoter regions (Figure 1C). CAGE signal clustering resulted in 55,204 decomposition peak identifiers (DPIs) and 10,254 bidirectional enhancers (promoter and exon regions were masked) (Figure 1D-E). Heart CAGE DPI clusters associated with 14,519 genes corresponding to 27,557

Encode or 20,411 Refseq transcripts. Bidirectional enhancers were connected to DPI clusters in the 500kb range: 1,142 bidirectional enhancers have significant associations with 1,491 DPI related to 469 genes (correlation thresholds  $r \geq 0.5$  and  $p\text{-value} < 0.05$ ).

Next, we performed classification of CAGE peaks by training TSSClassifier on control 2kb sequences such as eukaryotic promoter database (EPD), reference transcription start site (refTSS), promoter-like sequences (PLS), proximal enhancer-like sequences (pELS), distal enhancer like sequences (dELS) from Encode, and FANTOM5 enhancers. This step resulted in 30,036, 37,965, 31,008, 12,768, 15,062, and 5,616 matched heart CAGE clusters, respectively (**Supplementary Figure 1A**). Heart CAGE clusters were checked to have RNA-seq signals using Encode and other available sources<sup>10,11</sup>. This gives an option to filter out CAGE peaks without an RNA-seq signal to identify new markers. In total, 75.8% of DPI clusters were associated with transcription, but in the case of bidirectional enhancers, only 14% had overlap with RNA-seq signal. Other Encode libraries, including DNase-seq, assay for transposase-accessible chromatin sequencing (ATAC-seq), RAMPAGE, and chromatin immunoprecipitation sequencing (ChIP-seq), were used for overlapping and classification (**Supplementary Figure 1B-C, Supplementary Table 2-3**). This step allowed to filter out CAGE peaks without known ATAC/RAMPAGE/DNase/ChIP-seq signals to identify new targets. For example, ChIP-seq based classification confirmed the total of 10,154 active promoters and 4,533 active enhancers combined between DPIs and bidirectional enhancers. There were 29,332 CAGE peaks overlapping with significant PolR2A signal. Most of the DPI CAGE peaks had an “active promoter” label as expected, while ChIP-seq based classifier for bidirectional enhancers showed just a slightly higher number of “active enhancers” compared to “active promoters.” This can be explained by the insufficient sample diversity of ChIP-seq libraries. Most of the TSS (65%) showed the canonical YR initiator motif, 2.4% showed alternative YC, and ~30% had other dinucleotide frequency with the uncertain functional role<sup>6</sup> (**Supplementary Figure 1D**). This “other” motif had enriched G nucleotide which could be related to CAGE library preparation bias or other elements with the uncertain functional role. In addition, we validated our bidirectional enhancers by overlapping them with regions from other databases (**Supplementary Figure 1E**).

Based on GENCODE annotation, the DPI clusters were primarily found in promoter regions of genes, while bidirectional enhancers were located in intergenic and intronic regions (**Supplementary Figure 1F**). Usage of the EPD-like cluster classification assigned the location of 98% of DPIs to promoter regions. Initiator motifs for classified DPIs were well represented in promoter-like clusters and look similar to EPD, refTSS, and PLS with YR motif (**Supplementary Figure 2**). A similar motif was also present in enhancer-like clusters (pELS), but the motif is less evident in dELS-like regions.

In total, we defined 17,668 promoter and 14,920 enhancer regions active in the human heart, while 150 regions exhibited ambiguous features. Aggregation plots and heatmaps for Encode epigenetic libraries overlapping heart CAGE consensus regions are shown in **Supplementary Figure 3**.

## **Heart CAGE atlas revealed novel regulatory clusters specific for cardiac tissue**

Compared to other studies, heart CAGE peaks are well supported by other omics data, including histone methylation, acetylation, ATAC-seq, DNase-seq, and ChIP-seq for Pol2 (Figure 2). Defined CAGE peaks have

relatively high GC content and conservation scores, while the total length of consensus regions is the shortest resulting in higher resolution for motif and SNP analysis.

From the overlap with the studies, where heart regulatory regions are available, we identified 3,204 peaks (**Supplementary Figure 4**). These unique peaks contained 351 novel promoters, 1,318 novel enhancers, 8 ambiguous, 500 not classified consensus regions, and 43 alternative peaks. Direct comparison to the FANTOM5 database, where the CAGE technique is used as the primary tool for promoter and enhancer calling, showed 12,670 new DPIs (putative TSS) and 7,381 bidirectional enhancers. Still, EPD-like cluster classification reduced the number of new predicted promoters to 3,238 (**Supplementary Figure 5A**). The largest collection of human enhancers and promoters developed up to date (FANTOM5 project) contains only two heart-derived samples in total: one left atrial and ventricular sample, and through higher number of samples and deeper sequencing in this study allowed us to identify new cardiac peaks. Furthermore, the overlap with FANTOM5 peaks drastically improves the specificity of the epigenetic signal and clearly defines the location of transcribed *cis*-elements (**Supplementary Figure 5B**).

Applying *de novo* motif enrichment, we confirmed the presence of the transcription initiation sequences such as TATA-box and initiator element (INR) in the DPI regions (**Supplementary Figure 6A**). The most representative peak in the bidirectional enhancer region was the transcription factor binding site for the myocyte-specific enhancer factor 2B (*MEF2B*) motif, a muscle-specific gene activator (**Supplementary Figure 6A**). Overall, there was an accumulation of TFBS close to TSS in heart promoters and in the center of the bidirectional enhancers, which is expected from promoter and enhancer regions, thus confirming the accurate classification of promoter and enhancer regions (**Supplementary Figure 6B**).

Besides looking at the CAGE peak overlap with different databases, we also tested how sequencing depth affects comprehensive cluster detection. By increasing the library size, we have observed the sequencing depth reaching a plateau with no additional active genes and transcripts being detected (**Supplementary Figure 7A**). Our samples were saturated at this sequencing depth level in terms of new genes and transcripts active in different parts of the heart. Interestingly, we observed that the right and left atria had ~2,000 more active genes as compared to the left and right ventricles. Compared to FANTOM5 atrial and ventricular samples, we detected additional ~2,200 genes expressed in the human heart chambers (**Supplementary Figure 7A**). These genes participate in important signaling and metabolic pathways such as Wnt, mTOR, and autophagy (**Supplementary Figure 7B**). Libraries appeared saturated at the level of 1-2 million reads. Despite the saturation, deeper sequencing may still uncover the activity of low expressed genes and pseudogenes, including snRNA related to RNA transport and mRNA splicing<sup>12</sup>.

### **Differential expression analysis reveals large differences in transcription by *cis*-regulatory elements between healthy and failing atria and ventricles**

Dimensionality reduction showed a clear difference between atria and ventricles and between healthy and failing states (Figure 3A). However, we noticed a trend that atrial expression from failing hearts becomes similar to that of ventricular samples, with some exceptions. Information about these samples is available on the Zenbu reports platform with an interactive principal component analysis plot. Correlation analysis

confirms those findings by showing a strong correlation within atria and ventricles, with most of the failing atrial samples showing ventricular pattern of expression (**Supplementary Figure 8**).

Differential expression analysis between healthy and failing samples resulted in 1,748 and 915 CAGE clusters significantly different within atria and ventricles, respectively (Figure 3B-C). Of those significant clusters, 645 promoters and 223 enhancers were differentially expressed in healthy versus failing atria, while 291 promoters and 105 enhancers were differentially expressed in healthy versus failing ventricles. Healthy atria showed upregulation in genes related to the upkeep of the immune system surveillance genes, such as several chemoattractants for neutrophils (*CXCL1*, *CXCL3*, and *CXCL8*), cytokines (*CSF3*), and immunoadhesion activators (*SELE*). Genes involved in glucose metabolism (*PFKM*) and lipid utilization (*LPL*) were differentially activated in failing atria. Healthy ventricles also had activation of the innate immune system (*FCN3*, *LCN6*), while failing ventricles had defense response activation (*SPP1*, *ITLN1*).

Functional analysis of differentially expressed clusters between healthy and failing samples also resulted in enrichment of the immune system-related GO terms, including cytokines, inflammation, chemokines in addition to muscle system processes, collagen, and contractile fiber tags (Figure 3D). Among KEGG pathways most enriched were TNF signaling, complement, and cytokine-related pathways (Figure 3E). Disease ontology enrichment analysis resulted, as expected, in mostly heart-related diseases such as cardiovascular system disease, cardiomyopathy, and other muscle and connective tissue-related diseases (Figure 3F). Connecting top differentially expressed clusters with heart diseases identified inactivation of cardiac structure-specific genes (*MYH6*, *TNNT2*) and upregulation of metabolism (*LPL*, *APOA1*) and immune response genes (*CXCL10*, *CD36*) in failing hearts (Figure 3G).

Functional annotation from the FANTOM5 database, such as cell type, anatomy, and disease ontology, was transferred on heart CAGE clusters (**Supplementary Figure 9A**). The cell type annotation revealed that TSS in human heart chambers had features of epithelial cells, myoblasts, fibroblasts, smooth muscle cells, electrically active cells, endothelial cells, and blood vessel cells. Such precise identification of cardiac cell types in clusters obtained from bulk tissue demonstrates the cell level specificity of regulatory elements. Organ (UBERON)<sup>13</sup> ontology enrichment resulted in similarity to FANTOM5: heart, vessels, artery, and circulatory system-related samples, but also to embryo and lateral plate mesoderm samples. Disease ontology showed cardiovascular and heart-related diseases, but also cancer and disease of cellular proliferation. Similar patterns occurred in promoters and enhancers taken separately, confirming a functionally coherent behavior between the two types of regulatory elements. Then, to take a closer look into the functional differences between healthy and failing hearts, we annotated statistically significant differentially expressed clusters (**Supplementary Figure 9B**). The annotation showed a considerable involvement of the immune system and various embryogenesis-related pathways supporting previous observations of dedifferentiation of cells to the embryonic state during heart failure.

Differential expression analysis of ICM versus NICM identified disease-specific clusters, 323 for NICM and 255 for ICM (Figure 4A). Functional annotation with GO and KEGG highlighted several immune pathway activations in NICM (Figure 4B-C) and enrichment of lipid metabolism in ICM (Figure 4D-E). Then, we select several genes strongly associated with heart failure<sup>14</sup> to investigate their TSS usage. Variations in TSS

selection reflect the diversity of preinitiation complexes and can impact post-transcriptional RNA fates. We noticed no significant shift if ICM is directly compared to NICM. However, we observed significant differential remodeling of some clusters when ICM and NICM are compared separately to healthy samples. For example, the sarcomere gene, *TTN*, had an upregulated cluster in ICM, while NICM had an upregulated cluster in the junctional membrane gene, *JPH2* (**Supplementary Figure 1**). Cytoskeleton-related gene, *FLNC*, had a differential usage of the TSS between NICM and ICM, the former having one cluster inactivated while the latter having two clusters significantly inactivated compared to a healthy state. These observations demonstrate that different disease etiology could cause activation of specific regulatory elements of the genome.

Differential expression analysis for sample groups, considering chamber, disease type (ICM or NICM), and sex allowed to determine specific groups of genes and their functional roles using GO and KEGG pathway enrichment (**Supplementary Figure 11-12**). Failing heart tags included channel inhibitor activity and metabolism-related pathways. Failing ventricles showed pathway enrichment related to structural and metabolic activity while failing atria pathways were associated with neuronal activity. Heart failure, in general, was accompanied by *IRF* and *STAT* transcription factors (**Supplementary Figure 13**). Detailed heatmaps are available on the Zenbu reports platform

(<https://fantom.gsc.riken.jp/zenbu/reports/#heart%20CAGE%20GO>,

<https://fantom.gsc.riken.jp/zenbu/reports/#heart%20CAGE%20KEGG>,

<https://fantom.gsc.riken.jp/zenbu/reports/#heart%20CAGE%20TFBS>,

<https://fantom.gsc.riken.jp/zenbu/reports/#heart%20CAGE%20heatmaps>).

## SNPs in transcription factor binding sites can alter gene expression

While the most studied SNPs are located within the coding sequences, we also identified a 10% fraction of heart GWAS SNPs, which reside in regulatory regions (Figure 5A), most frequently in the promoter region (Figure 5B). The SNP density in heart CAGE clusters is higher than the genome-wide density, especially in classified consensus regions. The highest frequency of disease-associated SNPs corresponds to familial hypertrophic cardiomyopathy, congenital heart disease, atrial septal defect, cardiomyopathy, and familial atrial fibrillation (Figure 5C). Among National Human Genome Research Institute - European Bioinformatics Institute (NHGRI-EBI) GWAS SNPs, specific features included resting heart rate and glycated hemoglobin levels (**Supplementary Figure 14A**). Promoter regions showed accumulation of reticulocyte related SNPs while enhancers accumulated cardiac electrophysiology related SNPs (**Supplementary Figure 14B-C**).

Differentially expressed CAGE clusters between healthy and failing atria and ventricles accumulated GWAS SNPs related to blood pressure, autoimmune traits, blood protein levels in failing ventricles and electrocardiogram morphology, warfarin maintenance dose in failing atria (Figure 6A, **Supplementary Table 6**). To better understand the role of the SNPs in the regulatory regions, we identified the cases where these SNPs significantly change TFBS. In total, there were 2,679 GWAS SNPs localized in 479 unique TFBS. As an example, we looked at troponin, a known marker for the diagnosis of myocardial infarction and heart failure. The promoter of one of three troponin subunits, *TNNI3*, showed a statistically significant increase in its activity in failing hearts and was linked with SNP in its TFBS (Figure 6B-C). Other cases of SNP in TFBS are

available on the Zenbu reports platform  
(<https://fantom.gsc.riken.jp/zenbu/reports/#heart%20CAGE%20SNP>).

## Discussion

Heart failure is a leading cause of death in the USA and worldwide. Fundamental genetic mechanisms of cardiovascular diseases remain one of the primary targets of cardiovascular science. The genome regulatory network and its connection to cardiovascular disease are not well studied. Here, we present a comprehensive and fully annotated CAGE atlas of human heart promoters and enhancers from four chambers of the healthy and failing human hearts, including both atria and ventricles. We report that atria have 2,000 more regulatory elements versus ventricles which result in a chamber-specific expression pattern and explains higher flexibility of atrial phenotype. For annotation, we overlaid our results with many data sets available for the adult human heart. Deeper sequencing allowed the identification of regulatory elements with low expression levels involved in cell maintenance and embryogenesis. We confirmed the presence of over 17,000 promoters, and due to deeper sequencing and larger sample size, we detected an order of magnitude more enhancers, ~14,000 current state of knowledge in this field<sup>12</sup>. We developed a robust classification pipeline and identified over 3,000 novel regulatory regions not present in the FANTOM5 database and not described in other studies. Additionally, epigenetic markers showed that our dataset has a very concentrated signal of cardiac regulatory element regions compared to other databases. Finally, our enrichment analysis showed that our atlas represents nearly complete landscape of transcribed regulatory elements in atria and ventricles and represent ready-to-use comprehensive source for studies linking non-coding regulatory elements and diseases.

Comparing healthy to failing hearts, annotation of the top differentially expressed clusters revealed enrichment in embryonic processes enrichment suggesting a shift in gene expression due to re-expression of developmental genes in heart failure<sup>12</sup>, which are generally silent in the adult heart. To further understand the shift of regulatory elements in heart failure, we compared healthy to ischemic and non-ischemic cardiomyopathy samples. The most notable differences between healthy and failing samples included immune system responses and structural changes. Two etiologies of heart failure revealed unique features within the genome regulatory network. ICM specifically exhibited metabolic remodeling, while NICM demonstrated a regulatory element shift towards immune response activation. ICM versus NICM distinction is further exemplified by looking at the differential expression analysis at each cluster position. We highlighted the regulatory network of several genes that have substantial evidence in heart failure and genetic cardiomyopathies<sup>14</sup>. In addition to having multiple alternative TSS, genes related to cardiomyocyte structure such as cytoskeleton, sarcomere, and Z-disk functions showed differential TSS activation between healthy and failing hearts. These differences could help identify a novel set of biomarkers for cardiac disease classification and find potential regulatory target regions for development of novel therapeutics. We showed that over 10% of SNPs associated with cardiovascular diseases are located in the promoter regions, outside the protein-coding regions. By using (NHGRI-EBI) GWAS catalog, we found 2,678 SNPs were significantly associated with TFBS, which means they can activate or silence the promoter and thus affect the expression of 455 genes.



## Conclusion

We present an atlas of the genome regulatory elements of the human heart. This is a unique resource for understanding the functional impact of non-coding genetic regions on gene expression in healthy and diseased states. Aside from adding to the functional and genetic understanding of cardiac promoters and enhancers, CAGE data can be used to distinguish between healthy, ischemic, and non-ischemic hearts. The precise location of cardiac disease-related SNPs within the regulatory regions and their correlation with TFBS offers a novel understanding of the genetics of heart failure.

## Study Limitations

Due to a limited number of samples, this study could not investigate the effects of race and ethnicity on the cardiac regulatory network. Additionally, not all hearts had all four chambers sequenced usually due to the poor quality of extracted RNA. It remains unclear why some failing atria samples clustered with ventricles, and some did not.

Even though we included sinoatrial node samples in the database, their analysis was out of the scope of this study which primarily focused on the comparison of healthy and failing atrial and ventricular samples.

Another manuscript is being prepared to report on sex differences in the cardiac regulatory network.

## Methods

### Human heart procurement and demographic

Healthy de-identified human hearts were procured from the Washington Regional Transplant Community (Falls Church, VA), which were not acceptable for transplantation. The cause of death was non-cardiogenic. All protocols were approved by the George Washington University Institutional Review Board. Failing hearts were collected at INOVA Hospital (Falls Church, Virginia) during cardiac transplantation surgeries.

Demographic breakdown of the procured hearts was as follows: healthy hearts (N = 21) included 10 males and 11 females; failing hearts (N = 10) comprised of 7 males (3 ICM and 4 NICM) and 3 females (1 ICM and 2 NICM) (**Supplementary Table 1**).

### Tissue preparation, RNA extraction, and sequencing

Explanted non-diseased human hearts were arrested at the time of procurement with the cold cardioplegic solution and kept at +4°C during transportation to the research laboratory, typically less than 2 hours. Samples from the left and right atrial free wall, the right and left ventricular base, and inferior and superior sinus node were preserved in RNAlater (Invitrogen) overnight at +4°C and stored at -80°C until RNA extraction. Failing heart biopsies from all four chambers were obtained during surgery, immediately frozen in liquid nitrogen, and then delivered to the research laboratory on dry ice. To minimize RNA degradation during extraction, flash-frozen samples were transferred to RNAlater™-ICE Frozen Tissue Transition Solution (ThermoFisher Scientific). According to the manufacturer's protocol, total RNA was extracted from 40 mg of

tissue with the RNeasy Fibrous Tissue Mini Kit (Qiagen). RNA purity and yield were checked with Qbit and Agilent Bioanalyzer to satisfy RIN > 7 and total RNA amount 3-5 µg. This study was conducted in multiple sets: the first set (7 hearts) included samples only from the left atria and left ventricle. This set employed no-amplification non-tagging CAGE (nAnT-iCAGE) library preparation and sequenced on Illumina HiSeq2500 High-Throughput mode with 50nt single end and with 6M reads per sample<sup>15</sup>. After detecting the difference between atria and ventricles, we expanded our sample selection with additional healthy (N = 14) and failing (N = 10) hearts, including all four chambers. The second and third sequencing sets used a more recent CAGE sequencing platform, single strand CAGE (ssCAGE). This methodology is complementary to nAnT-iCAGE with only a technical difference at the adapters ligation stage. This methodology allows for deeper sequencing needed to detect a full range of regulatory elements, especially enhancers with intrinsically lower expression. Samples were sequenced on Illumina HiSeq2500 by one-shot loading with 80nt and over 25 M reads per sample<sup>16</sup>.

## CAGE data alignment, annotation, and classification

Sequenced reads were checked for quality with FASTQC<sup>17</sup>, then trimmed for quality and length with fastx\_trimmer (-Q33), and adapters removed with trimmomatic<sup>18</sup>. Reads mapped to human ribosomal RNA locus (U13369.1) were removed. Clean reads were mapped on hg38 primary chromosomes with Burrows-Wheeler Aligner<sup>19</sup>, and unmapped reads were remapped with HISAT2<sup>20</sup>. Mapped reads were converted into CAGE transcriptional start sites (CTSS) using CAGEr Bioconductor package (sequencingQualityThreshold = 10, mappingQualityThreshold = 20, removeFirstG = TRUE, correctSystematicG = TRUE) to count 5' end of the mapped CAGE reads at single base-pair resolution<sup>21</sup>. Library sizes were defined by the total number of mapped CAGE tags in output CTSS files.

To compensate for the heterogeneous profile of the CTSS clusters, the decomposition peak identification (DPI) program, which relies on the independent component analysis, was used to define a set of reference peaks for all samples<sup>5</sup>. This protocol produces permissive (at least 3 mapped reads to the same location) and robust DPIs (tag per million > 1). In the subsequent analysis, we used robust DPIs only. DPIs were linked with genes through a custom-developed CAGE\_peak\_annotation package (+/-500 bp rule) and ChIPseeker package in R<sup>22</sup>. Databases for annotation included Augustus<sup>23</sup>, Genscan<sup>24</sup>, NCBI RefSeq<sup>25</sup>, and Gencode v37 from Gencode<sup>26</sup> and UCSC Table Browser<sup>27</sup>. Bidirectional enhancers were called and linked to genes as described by Andersson et al.<sup>8</sup>. Briefly, divergent CAGE TSS pairs were selected by max separation distance of 400 bp, merged into bidirectional pairs with flanking windows added (200bp), filtered by directionality score, and masked by Gencode promoter and exon regions, +/-500 and 200 bp, respectively. TomeTools TSSClassifier (<http://tomertools.sourceforge.net/>) was used to classify DPI and bidirectional enhancers into promoters and enhancers. As control datasets, we selected eukaryotic promoter database (EPD)<sup>28</sup>, reference TSS (refTSS)<sup>29</sup>, ENCODE sequences such as promoter-like sequences (PLS), proximal enhancer-like sequences (pELS), distal enhancer like sequences (dELS)<sup>30</sup>, and FANTOM5 enhancers, then applied classifier on 2000bp regions of DPI and bidirectional clusters. Optimal specificity/sensitivity thresholds were used for classification. To get consensus regions DPI clusters and bidirectional enhancers were extended up to 400bp (300 upstream, 100 downstream for DPI, and +/- 200 bp for bidirectional enhancers) and collapsed if they

had any overlap. Consensus regions were defined as the highest peak in the 400 bp region belonged to EPD / refTSS / PLS but not to pELS / dELS was classified as a promoter, in the opposite case - as an enhancer, and the rest - ambiguous (**Supplementary Figure 1A**). This classifier was adjusted by the direct overlap with ENCODE regions.

Dinucleotide analysis was performed on DPI clusters centered on TSS. Classification of YR was given in the case of CA/CG/TA/TG and YC for CC/TC, the remaining motifs were called “other” (**Supplementary Figure 1D**).

Vista enhancers<sup>31</sup>, super-enhancers<sup>32</sup>, and *H. sapiens* promoters EPD<sup>28</sup> were obtained from the related sources. UCSC LiftOver tool was applied when hg38 annotation was unavailable (**Supplementary Figure 1E**).

The WGCNA package<sup>33</sup> was applied for co-expression clusters identification on both experimental groups separately with soft thresholds 9 and 14 for groups one and two, respectively.

Overlaps and upset plots of cardiac regulatory elements from other data sources<sup>12,34-37</sup> were done with bedtools v2.28 comparing all heart CAGE clusters (Fig. 2, **Supplementary Figure 4**). We compared the entire set of robust DPI peaks, DPI peaks similar to EPD sequences (by TSSClassifier), and bidirectional enhancers with the FANTOM5 database (**Supplementary Figure 5**).

The number of unique genes and transcripts in the heart was counted by the association of gene models (RefSeq, Gencode, Augustus, Genscan) with DPI clusters by CAGE\_peak\_annotation. Genes or transcripts were considered active if associated with a CAGE cluster with 10 or more tag counts. Saturation curve built on random subsets of selected sizes from original BAM files with mapped CAGE reads. These subsets were clustered independently and connected to gene models to count numbers of genes/transcripts (**Supplementary Figure 7A**).

The heatmap of all heart CAGE samples was based on the Pearson correlation of TPM values for all CAGE clusters (DPI and bidirectional enhancers) (**Supplementary Figure 8**).

## Epigenetic marker overlay with CAGE clusters

ATAC-seq data for the human heart was obtained from Broad Institute’s Cardiovascular Disease Knowledge Portal<sup>38</sup> (**Supplementary Figure 1B**). Signal data for epigenetic marks of H3K4me3, H3K27ac, H3K4me1, together with PolR2A and DNase-seq, were obtained from the ENCODE project<sup>30</sup> (**Supplementary Figure 1B-C, Supplementary Table 3**). For the CAGE clusters which overlap with narrow ChIP-seq peaks, an additional rule was applied to assign class according to Jiang and Mortazavi<sup>39</sup>. RNA-seq data from GSE116250 and GSE128188 were aligned to the genome and clustered using cufflinks<sup>40</sup>. Obtained transcripts were connected to DPI clusters according to the  $\pm 500$ bp rule. Precalculated RNA-seq data from Encode based on Gencode (v24 and v29) was associated with clusters and considered active if transcripts have nonzero FPKM values. Conservation scores phastCons from 100-way alignment and CpG islands regions were accessed from UCSC for hg38 (Fig. 2). Fractions of signals were counted using bigWigToBedGraph and bedtools v2.28 (map, groupBy) with a 5 bp window.

## De novo transcription factor binding site identification for cluster validation

Regulatory sequence motifs we analyzed using Match<sup>41</sup> for known TFBS and cisExpress<sup>42,43</sup> for *de novo* discovery (**Supplementary Figure 6**). Match tool had parameters of matrix similarity with at least 0.95 and core similarity of 1 to calculate the distribution of TFBS density. The clustering of the TFBS was conducted based on the positional density. First, position-specific density profiles were calculated for each TFBS in the TRANSFAC database and the rare motifs (those that appear less than in 1/10 of the analyzed sequences) were excluded. Then, Pearson's correlation coefficients were computed between densities of motifs in the window of 10 nucleotides. Hierarchical clustering was achieved with the Ward method and  $(1 - \text{correlation})/2$  as a distance using hclust<sup>44</sup>. Kruskal-Wallis test, and the Wilcoxon rank-sum test with continuity correction allowed to assess the significance of pairwise differences between clusters of TFBS.

The cisExpress algorithm enabled *de novo* motif discovery. This approach is based on the calculation of the Z-score, showing the influence strength of having a nucleotide sequence  $w$  in the positional window  $k$ . The window size was selected based on the quality of the genome annotation and properties of the organism and the experiment.

$$Z_{score}(w, k) = \frac{e_{with}(w, k) - e_{without}(w, k)}{\sqrt{\frac{Stdev_{with}^2(w, k)}{n_{with}(w, k)} + \frac{Stdev_{without}^2(w, k)}{n_{without}(w, k)}}},$$

where  $e_{with}(w, k)$  and  $e_{without}(w, k)$  were the average gene expression values with and without the word  $w$  in the position  $k$ ,  $Stdev_{with}(w, k)$  and  $Stdev_{without}(w, k)$  were the standard deviations of gene expression values;  $n_{with}(w, k)$  and  $n_{without}(w, k)$  were the numbers of sequences of genes containing and not containing word  $w$  in the  $k^{\text{th}}$  window. Words with Z-scores above a predefined threshold were stored as primary motifs. Groups of similar motifs discovered within one window were merged, resulting in longer and/or more ambiguous motifs. This part of the algorithm produced the motifs in the form of consensus sequences and included the position window where it was discovered.

## Differential expression analysis and pathway enrichment

Identified clusters were analyzed with edgeR for differential expression<sup>45</sup>. Dispersions were calculated with the GLM method, likelihood ratio tests were done with glmFit and glmLRT. Principle component analysis was used for sample dispersion visualization. The main comparison groups were 'healthy ventricle versus failing ventricle' and 'healthy atrium versus failing atrium' (Figure 3B-C). Other comparisons are shown in **Supplementary Figure 11-13** were based on scores calculated as  $z = (x - \mu)/\sigma$ , where  $x$  is the  $\log_2|p\text{-adjusted}|$  (p-values adjusted with Benjamini-Hochberg correction and filtered to be p-adjusted < 0.05),  $\mu$  is the mean, and  $\sigma$  is the standard deviation (**Supplementary Table 4**).

Statistically significant clusters with FDR < 0.05 were further analyzed for gene ontology enrichment, disease, and KEGG pathway enrichment with goana<sup>46</sup>, ShinyGO<sup>47</sup>, and enrichDO and enrichKEGG of the clusterProfiler<sup>48</sup>. Gene - disease network plot was made with cnetplot (enrichplot package) (Figure 3G).

# Transcription factor binding and single nucleotide polymorphism analysis

Heart genome-wide association study (GWAS) data with selected heart diseases for hg38 assembly was obtained from Ensembl BioMart (**Supplementary Table 5**). NHGRI-EBI GWAS catalog was used as another source of SNPs in our study<sup>49</sup>. Enrichment of the features from GWAS studies was made with enricher (clusterProfiler) for 400 (300 upstream, 100 downstream) and 1000 (900 upstream, 100 downstream) bp regions (**Supplementary Figure 15A**). Same heart CAGE flanking regions were used for Fisher's exact test on the selected features (**Supplementary Figure 15B**).

Percentage of heart GWAS SNP overlapping different genomic regions (heart CAGE DPI - upstream 300 bp, downstream 100 bp, bidirectional enhancers center +/- 200 bp) was counted for each chromosome separately (Figure 5). The fraction was estimated in 10M windows with 1000 bootstrap.

Local TFBS enrichment analysis was performed using MEME-Suite CentriMo with JASPAR2018 Pol2 motifs on 400 bp sequences and default parameters for all heart CAGE DPI clusters<sup>50</sup>. SNP effect on TFBS was found with motifbreakR<sup>51</sup> by processing a list of enriched motifs with heart GWAS SNPs located in 400bp regions (Figure 6, **Supplementary Table 6**).

Differentially expressed DPI clusters were filtered to avoid overlap with exons, then extended to 400bp (300 upstream, 100 downstream). In case of overlap with another cluster, the cluster with a higher number of tag counts was selected. Sequences of selected regions were submitted to MEME Suite 5.3.3 AME with default parameters and Jaspasr 2018 CORE non-redundant vertebrate motifs. Bonferroni corrected p-values (p-adjusted < 0.05) were transferred to log scale and used for score calculation and clustering (**Supplementary Table 4**).

## Online access to the atlas with Zenbu reports platform

All heart CAGE data, annotations, and analysis results were integrated using the Zenbu reports platform ZENBU 3.0.1, which is provided here as an open-source database (<https://fantom.gsc.riken.jp/zenbu/reports/#Atlas%20of%20cardiac%20promoters%20and%20enhancers>)<sup>52</sup>. The database contains separate tracks for permissive and robust DPIs, bidirectional enhancers, and classified promoters and enhancers. Results of differential expression are presented in individual tracks with highlights of statistical significance at FRD < 0.05. A comprehensive manual for the database usage is available in **Supplementary Document 1**.

All heart CAGE clusters, sample descriptions, and detailed annotations are available on NCBI GEO portal (accession number GSE150736) and in **Supplementary Table 1-2**, respectively.

## Declarations

## Acknowledgments

# Sources of funding

This study was funded by Leducq Foundation (project RHYTHM to IRE and AG), National Institutes of Health (30T20D023848, R01 HL126802, U01 HL141074 to IRE) and Russian Foundation for Basic Research grant 19-29-04111 (to RS), and University of La Verne Faculty Development fund (to TT).

# Author contributions

AG and IE collected samples; OG performed sequencing; RD, AG, and TT conducted data analysis; RD, AG, RS, TT, OG, and IE wrote and critically revised the manuscript.

# Abbreviations

A- atrium, bg- background, bp- base pair, CAGE- cap analysis of gene expression, CDS- coding sequence, CTSS- CAGE transcription start site, DE – differential expression, dELS- distal enhancer like sequences, DO- Disease Ontology, DOID- Disease Ontology ID, DPI- decomposition peak identification, EPD- eukaryotic promoter database, FANTOM5- functional annotation of the mammalian genome 5th version, FPKM- Fragments Per Kilobase of transcript per Million mapped reads, GENECODE- genetic encyclopedia of DNA elements, GO- Gene Ontology, GWAS- genome-wide association study, ICM- ischemic cardiomyopathy, INR- initiator, KEGG- Kyoto Encyclopedia of Genes and Genomes, LA- left atrium, logFC- log<sub>2</sub> fold change, LV- left ventricle, NHGRI-EBI - National Human Genome Research Institute - European Bioinformatics Institute, NICM- non-ischemic cardiomyopathy, pELS- proximal enhancer-like sequences, PLS- promoter-like sequences, refTSS- reference TSS, ROC- Receiver operator characteristic curves, SNP- single nucleotide polymorphism, TF- transcription factor, TFBS- transcription factor binding site, TIR- transcription initiation region, TSS- transcription start site, V- ventricle.

# References

1. Liu, Y. *et al.* RNA-Seq identifies novel myocardial gene expression signatures of heart failure. *Genomics* **105**, 83–89 (2015).
2. Anene-Nzelu, C. G., Lee, M. C. J., Tan, W. L. W., Dashi, A. & Foo, R. S. Y. Genomic enhancers in cardiac development and disease. *Nat. Rev. Cardiol.* (2021) doi:10.1038/s41569-021-00597-2.
3. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
4. Carullo, N. V. N. & Day, J. J. Genomic enhancers in brain health and disease. *Genes (Basel)* **10**, 43 (2019).
5. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

6. Nepal, C. *et al.* Dual-initiation promoters with intertwined canonical and TCT/TOP transcription start sites diversify transcript processing. *Nat. Commun.* **11**, 168 (2020).
7. Nepal, C. *et al.* Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Res.* **23**, 1938–1950 (2013).
8. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
9. Kuner, R. *et al.* Genomic analysis reveals poor separation of human cardiomyopathies of ischemic and nonischemic etiologies. *Physiol. Genomics* **34**, 88–94 (2008).
10. Thomas, A. M. *et al.* Differentially expressed genes for atrial fibrillation identified by RNA sequencing from paired human left and right atrial appendages. *Physiol. Genomics* **51**, 323–332 (2019).
11. Yamaguchi, T. *et al.* Cardiac dopamine D1 receptor triggers ventricular arrhythmia in chronic heart failure. *Nat. Commun.* **11**, 4364 (2020).
12. Gacita, A. M. *et al.* Altered enhancer and promoter usage leads to differential gene expression in the normal and failed human heart. *Circ. Heart Fail.* **13**, e006926 (2020).
13. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).
14. Jordan, E. *et al.* Evidence-based assessment of genes in dilated cardiomyopathy. *Circulation* **144**, 7–19 (2021).
15. Murata, M. *et al.* Detecting expressed genes using CAGE. *Methods Mol. Biol.* **1164**, 67–85 (2014).
16. Morioka, M. S. *et al.* Cap analysis of gene expression (CAGE): A quantitative and genome-wide assay of transcription start sites. *Methods Mol. Biol.* **2120**, 277–301 (2020).
17. Labsquare Team *et al.* *Fastq 0.2.3: A quality control tool for high throughput sequence data.* (Zenodo, 2017). doi:10.5281/zenodo.824550.
18. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
19. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
20. Wen, G. A simple process of RNA-sequence analyses by Hisat2, htseq and DESeq2. in *Proceedings of the 2017 International Conference on Biomedical Engineering and Bioinformatics - ICBE 2017* (ACM Press, 2017). doi:10.1145/3143344.3143354.
21. Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGER: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, e51 (2015).
22. Yu, G., Wang, L.-G. & He, Q.-Y. CHIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
23. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465-7 (2005).
24. Burge, C. B. Modeling dependencies in pre-mRNA splicing signals. in *Computational Methods in Molecular Biology* 129–164 (Elsevier, 1998).

25. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-45 (2016).
26. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
27. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-6 (2004).
28. Dreos, R., Ambrosini, G., Périer, R. C. & Bucher, P. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Res.* **43**, D92-6 (2015).
29. Abugessaisa, I. *et al.* RefTSS: A reference data set for human and mouse transcription start sites. *J. Mol. Biol.* **431**, 2407–2422 (2019).
30. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
31. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88-92 (2007).
32. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44**, D164-71 (2016).
33. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
34. Dickel, D. E. *et al.* Genome-wide compendium and functional assessment of in vivo heart enhancers. *Nat. Commun.* **7**, 12923 (2016).
35. Lee, D. *et al.* Human cardiac cis-regulatory elements, their cognate transcription factors, and regulatory DNA sequence variants. *Genome Res.* **28**, 1577–1588 (2018).
36. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
37. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017).
38. van Ouwerkerk, A. F. *et al.* Identification of atrial fibrillation associated genes and functional non-coding variants. *Nat. Commun.* **10**, 4755 (2019).
39. Jiang, S. & Mortazavi, A. Integrating ChIP-seq with other functional genomics data. *Brief. Funct. Genomics* **17**, 104–115 (2018).
40. Trapnell, C. *et al.* Erratum: Corrigendum: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **9**, 2513–2513 (2014).
41. Kel, A. E. MATCHM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31**, 3576–3579 (2003).
42. Troukhan, M., Tatarinova, T., Bouck, J., Flavell, R. B. & Alexandrov, N. N. Genome-wide discovery of cis-elements in promoter sequences using gene expression. *OMICS* **13**, 139–151 (2009).
43. Triska, M., Grocutt, D., Southern, J., Murphy, D. J. & Tatarinova, T. cisExpress: motif detection in DNA sequences. *Bioinformatics* **29**, 2203–2205 (2013).
44. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *J. Classif.* **31**, 274–295 (2014).



45. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
46. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
47. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
48. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y)* **2**, 100141 (2021).
49. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
50. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-8 (2009).
51. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).
52. Severin, J. *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nat. Biotechnol.* **32**, 217–219 (2014).

## Supplementary Figure Legends

**Sup Figure 1.** Development of robust classification for CAGE peaks. **A.** Receiver operator characteristic curves (ROC) (left) and a number of classified peaks (right) display the results of CAGE peak classification (DPIs – top, bidirectional enhancers – bottom) with a machine learning algorithm, TSSClassifier. The classifier was trained on the regions from Encode (PLS - promoter-like sequences, pELS/dELS - proximal/distal enhancer-like sequences), FANTOM5 (refTSS), and eukaryotic promoter database (EPD). **B.** CAGE cluster (DPIs – top, bidirectional enhancers – bottom) overlaps with RNA-seq and Encode peaks for DNase-seq, ATAC-seq, Rampage, and POLR2A ChIP-seq narrow peaks. **C.** ChIP-seq based classification of CAGE clusters (DPIs – top, bidirectional enhancers – bottom). **D.** Dinucleotide frequency of CAGE TSS (highest peak in DPI cluster). **E.** Bidirectional enhancers overlap with other enhancer databases. **F.** Genomic location of all DPI CAGE peaks, EPD-like regions, and bidirectional enhancers according to TSSClassifier.

**Sup Figure 2.** Motif logos for different types of heart CAGE TSS peaks (highest peak in DPI cluster). YR and YC logos were created for TSS classification based on dinucleotide analysis and the remaining TSS were annotated as “other”.

**Sup Figure 3.** Aggregation plots and corresponding heatmaps for heart promoters and enhancers confirm region-specific epigenetic signal enrichment. Epigenomic data for the human heart was obtained from Encode (**Supplementary Table 3**).

**Sup Figure 4.** Upset plot for the intersection of CAGE clusters with different databases. To note, the FANTOM5 database includes the entire set of promoter peaks (robust DPIs and bidirectional enhancers), while other resources are human heart specific.

**Sup Figure 5.** Comparison of heart CAGE clusters with FANTOM5 database. **A.** Venn diagrams for putative promoters (DPI), predicted promoters by TSSClassifier (EPD-like), and bidirectional enhancers. **B.** Aggregation plots of database-specific and overlapping clusters for different epigenetic markers. In total there are four categories - two for specific clusters, and two for overlapping: heart CAGE clusters in FANTOM5 regions and the opposite.

**Sup Figure 6.** Motif analysis of heart CAGE regulatory elements. **A.** Positional enrichment of top enriched motifs: TATA-Box and initiator element (INR) in DPI clusters and EPD-like DPI clusters, and myocyte-specific enhancer factor 2B (*MEF2B*) in bidirectional enhancers. **B.** Frequency of TFBS in regulatory regions. Shown TFBS distribution is a sign of promoter and enhancer function.

**Sup Figure 7.** Deeper CAGE sequencing of heart samples uncovers newly transcribed genes. **A.** CAGE sequencing saturation in comparison to FANTOM5 samples on gene and transcript levels **B.** Functional analysis of genes identified only in heart CAGE data for atria and ventricles.

**Sup Figure 8.** Correlation heatmap of all samples. Heatmap is based on Pearson correlation scores calculated for DPI clusters (TPM counts) for each sample.

**Sup Figure 9.** Functional analysis of heart CAGE elements based on FANTOM5 ontologies. Since many heart CAGE peaks overlap with FANTOM5 peaks it is possible to check the overrepresentation of cell type (CL) / organ (UBERON) / disease (DOID) specific TSS. **A.** Top enriched tags of CL, UBERON, and DOID ontologies in the whole set of heart DPI, predicted promoters, and enhancers (first, second, and third columns respectively). **B.** Top overrepresented ontology tags for differentially expressed DPIs (putative promoters) or EPD-like DPIs (predicted promoters).

**Sup Fig 10.** Differential expression showed ICM and NICM specific usage of CAGE clusters. A curated set of genes related to heart failure was used to demonstrate the distribution and activity of CAGE clusters within each gene.

**Sup Figure 11.** Specific GO terms for differentially expressed genes within selected groups. DE genes were defined as  $|\log_2(\text{FC})| > 1$  and  $\text{FDR} < 0.05$  and submitted to GO over-representation test. Significant GO terms ( $\text{FDR} < 0.05$ ) were selected for each group,  $\log_2(\text{FDR})$  values were used for score calculation. Full clustering information is available in **Supplementary Table 4**.

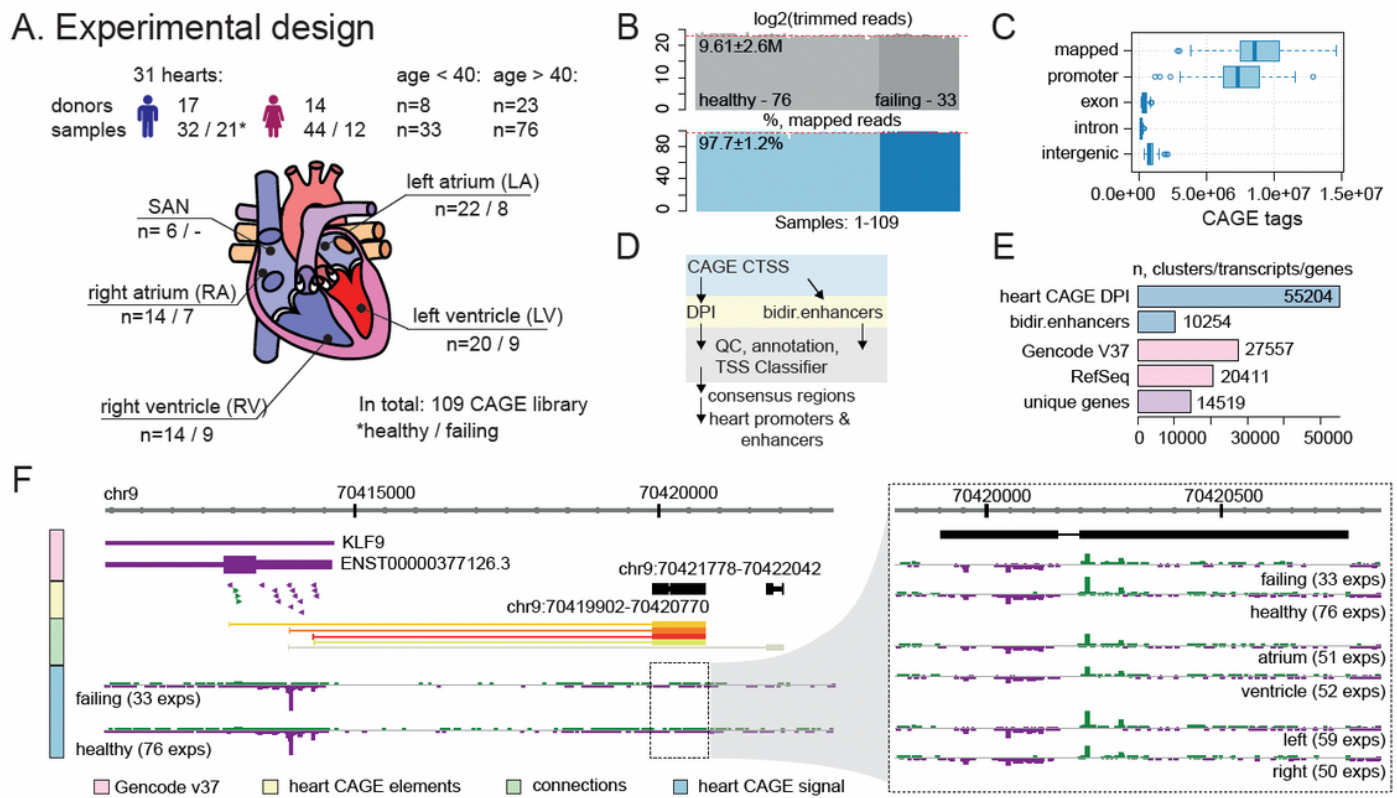
**Sup Figure 12.** KEGG pathway analysis of differentially expressed genes within selected groups. KEGG enrichment analysis of a gene set was applied on differentially expressed genes (similar as in **Supplementary Figure 11**),  $\log_2(\text{FDR})$  used for score calculation and hierarchical clustering. Detailed clustering data is available in **Supplementary Table 4**.

**Sup Figure 13.** TFBS enrichment analysis for selected comparison groups. Differentially expressed DPI cluster regions were extended and submitted to MEME Suite 5.3.3 AME with default parameters and Jaspas 2018 CORE non-redundant vertebrate motifs. Bonferroni corrected p-values were transferred to log scale and used for score calculation and clustering. Clustering data is available in **Supplementary Table 4**.

**Sup Figure 14.** A detailed breakdown of heart GWAS SNPs location from the 10% found in regulatory regions (Figure 5C). Percent of SNPs located in **A.** non-exonic regions to highlight only regions involved in regulatory activity, **B.** promoters, and **C.** enhancers.

**Sup Figure 15.** Variants associated with heart diseases and functionality are enriched in defined regulatory regions. **A.** GO enrichment analysis for heart CAGE consensus clusters with a length of 400bp (left) and 1000bp (right). **B.** Examples of traits with significant association to regulatory regions calculated by Fisher's exact test.

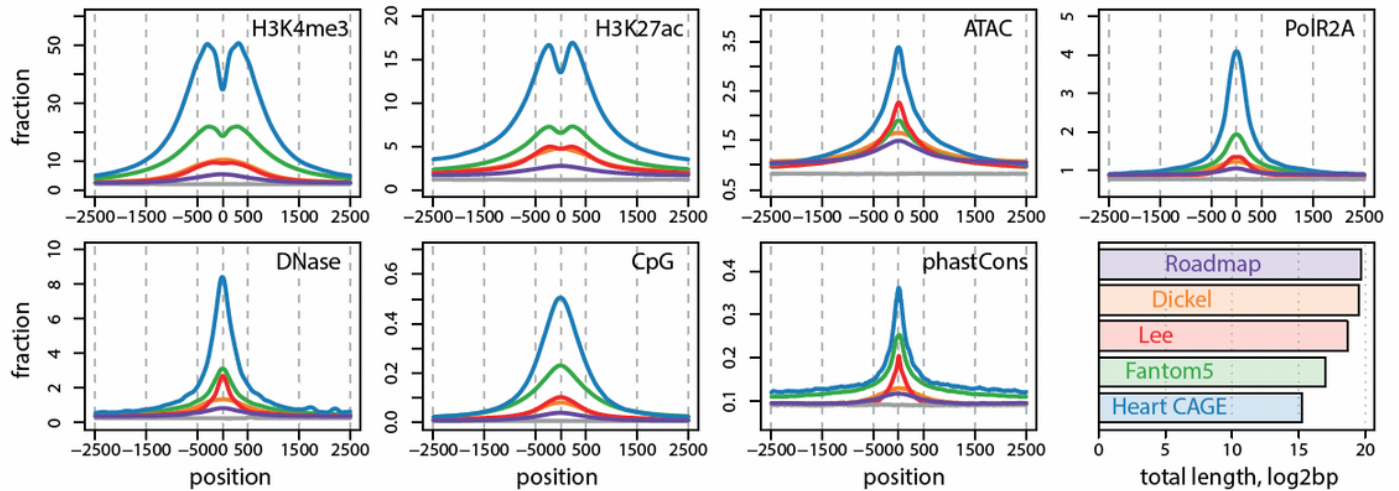
## Figures



**Figure 1**

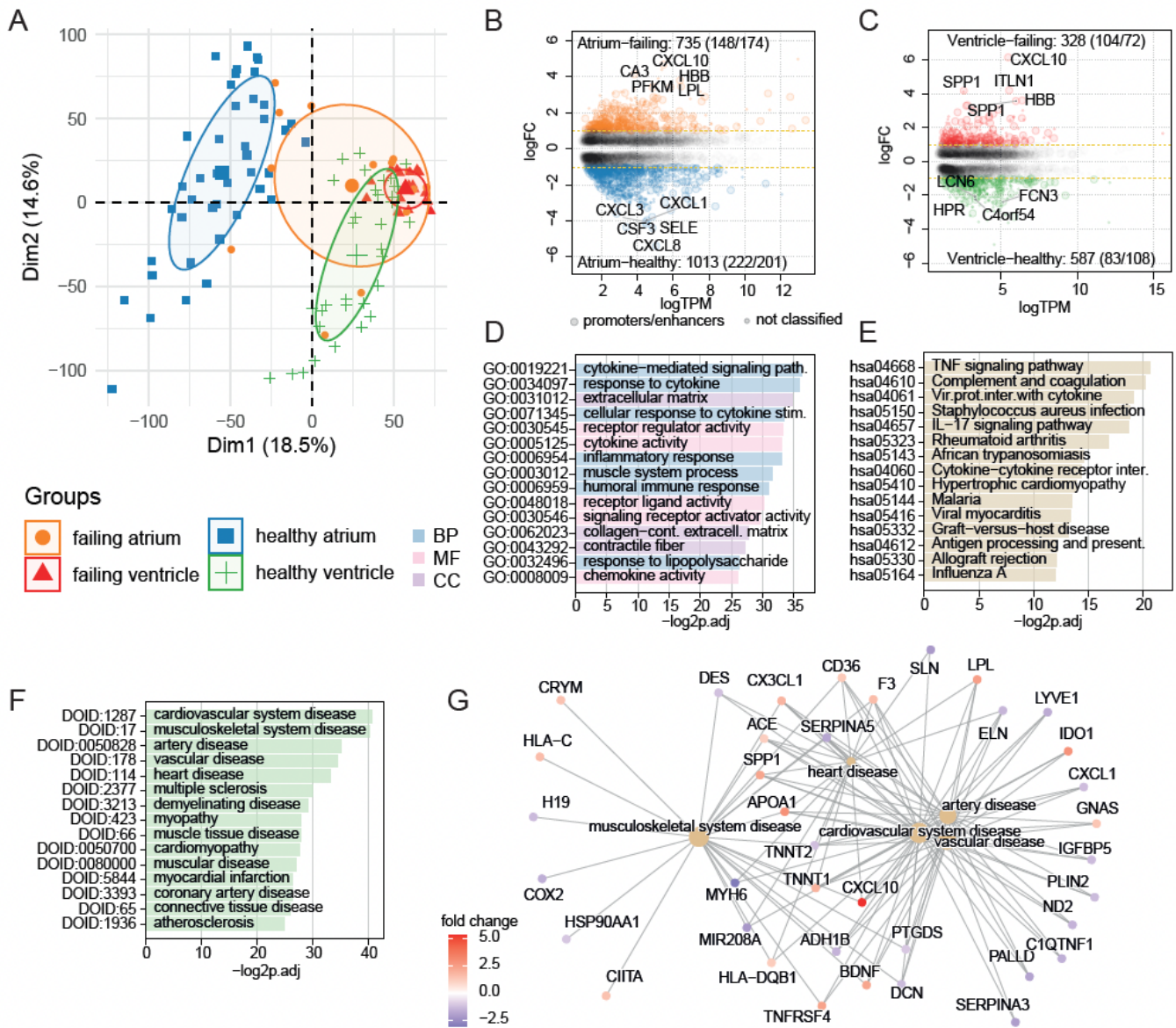
Atlas of transcribed CAGE clusters of healthy and failing human hearts. **A.** Experimental design, showing numbers of samples, donors, hearts, sources of samples. **B.** Sequenced reads were trimmed by quality, against rRNA and adapters. In total >109 reads were aligned to the genome with an average mapping ratio of 97.7%. **C.** Majority of reads were aligned into promoter regions of annotated genes and transcripts by Gencode v37 (on average 7.4M per library), a second large portion of reads was mapped to intergenic regions (~0.9M per library) and likely associated with enhancer activity. **D.** We applied two protocols of CAGE signal clustering using FANTOM5 pipelines to annotate putative promoters and enhancers. Defined clusters were classified using machine learning on reference regions including Encode, FANTOM5, and EPD. Then, clusters were extended and merged into consensus regions - predicted heart promoters and enhancers. **E.**

Total amount of DPI and bidirectional enhancers clusters, transcripts based according to overlap with annotation databases, and total transcriptionally active genes in the human heart. F. Example view showing an expressed bidirectional enhancer connected to DPI CAGE peaks. All peaks with related classifiers are available in Supplementary Table 2 and Zenbu reports platform.



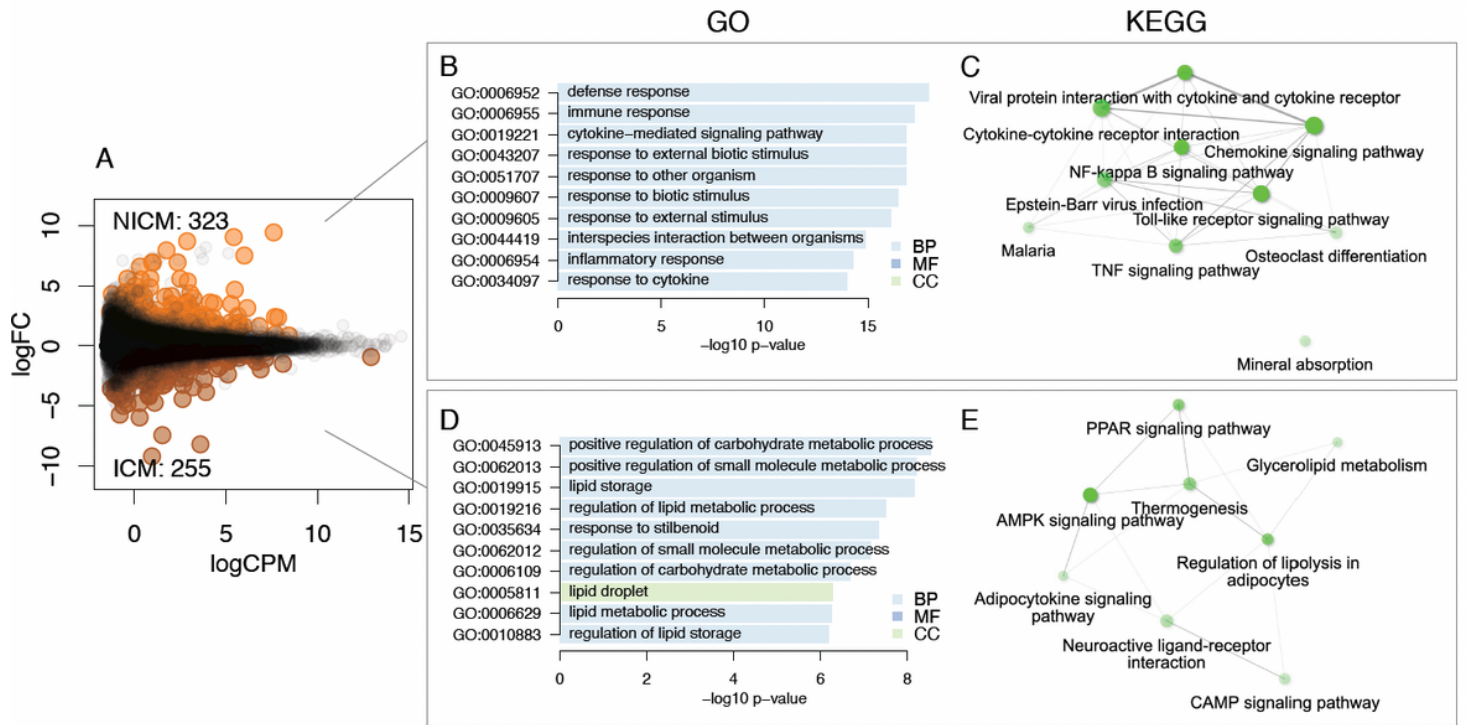
**Figure 2**

CAGE clusters active in the human heart show higher resolution and specificity of epigenetic signals in comparison to other available databases of human heart promoters and enhancers. Data for aggregation plots in the case of ChIP-seq (H3K4me3, H3K27ac, PolR2A), ATAC-seq, and DNase-seq were obtained from Encode. CpG and phastCons values for 100-way genomic alignment were obtained from UCSC.



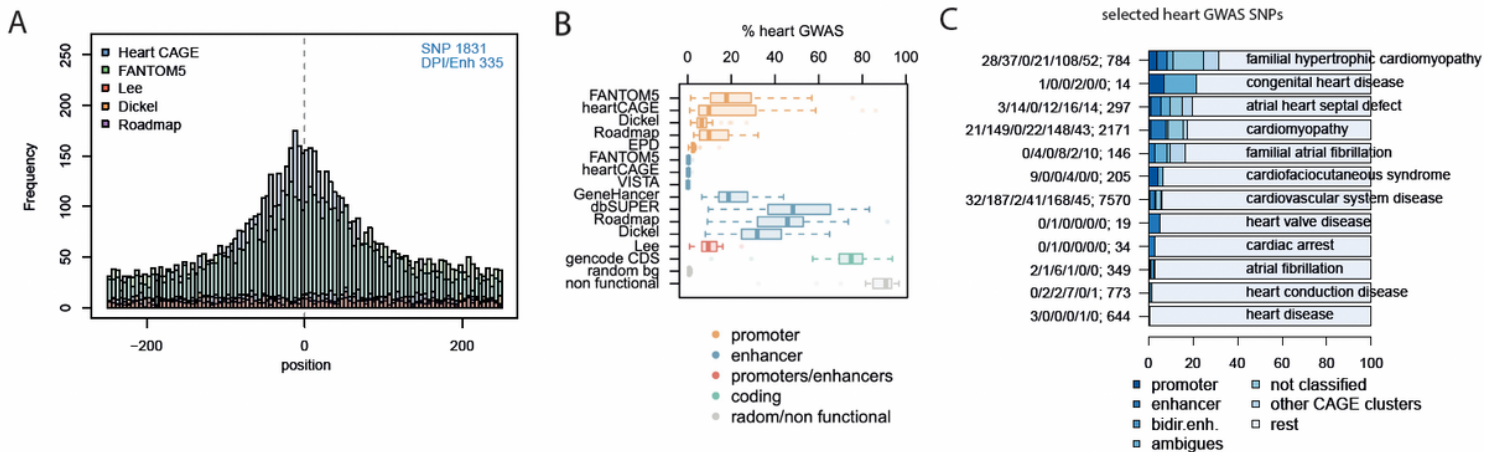
**Figure 3**

Differential expression analysis between healthy and failing hearts reveals main gene markers and immune system activation. A. Principal component analysis of all heart CAGE clusters from all 109 samples. Highlighted are four groups of interest: healthy and failing atria; healthy and failing ventricles. B, C. Differential expression analysis between healthy and failing sample groups - atria and ventricles, respectively. Big dots represent promoters/enhancers (TSSClassifier based), small dots - unclassified. Colored dots have  $FDR < 0.05$ . Legends show a total number of differentially expressed all clusters, promoters, and enhancers respectively. D-F. Top gene ontology, KEGG pathway, and disease ontology enriched tags for differentially expressed genes both in ventricles and atria. G. Gene network associated with top genes enriched in heart diseases.



**Figure 4**

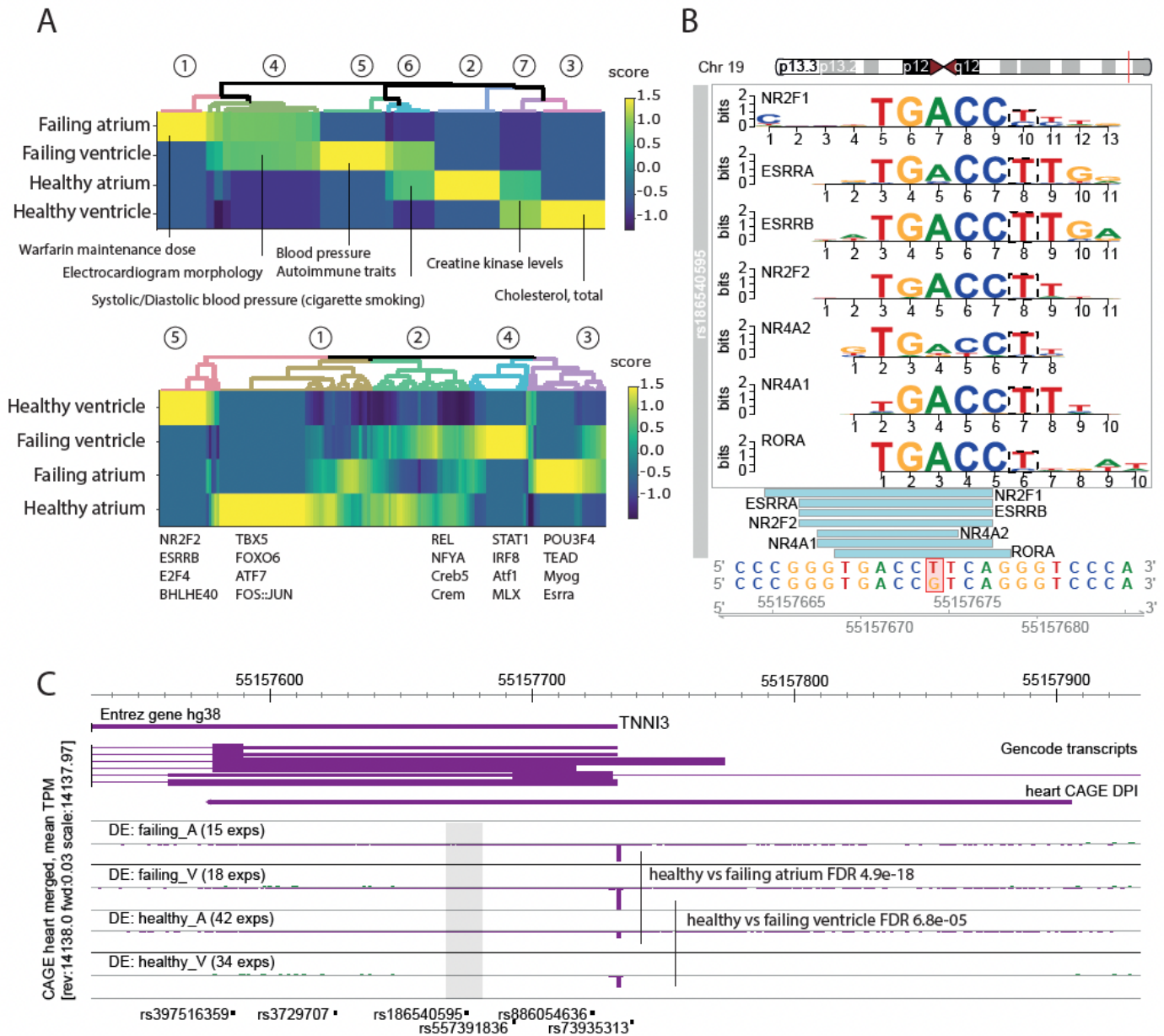
Comparison between ICM and NICM identifies disease-specific regulatory elements and pathways. A. Differential expression between ICM and NICM, numbers highlight the number of statistically significant (FDR<0.05) clusters in each group. B-C. GO and KEGG pathway analysis for differentially expressed clusters in ICM. D-E. The same analysis for NICM.



**Figure 5**

Heart GWAS SNPs are located primarily in promoter regions. A. Frequency of SNP occurrence in TSS region (center of the region in case of enhancers) identified in different databases. In total, 1831 SNPs were found in proximity to 335 CAGE clusters. B. Distribution of heart GWAS SNPs based on their function in the

genome. C. Sorted SNPs detected in CAGE clusters. Shown are the numbers of SNPs found in each region. Category 'rest' represents the SNPs without overlap to hear CAGE. Full list of SNPs is available in Supplementary Table 5.



**Figure 6**

Genomic variants affecting the regulatory network are linked to heart traits and diseases. A. Specific GWAS SNPs (top) and TFBS (bottom) were defined for differentially expressed CAGE clusters between healthy and failing heart samples. Highlighted are representative example features and TFs. More detailed terms are available in Supplementary Table 6. B. An example of enriched motifs located in regulatory regions containing familial hypertrophic cardiomyopathy SNP region, rs186540595 (location is marked with dotted line). The presence of this SNP causes significant changes in the TFBS structure (highlighted in red square). Other cases are available on the Zenbu reports platform. C. Visualization of rs186540595 SNP shows its proximity to differentially expressed CAGE TSS of TNNI3.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryDocument1HeartCAGEdatabaseuserguide.pdf](#)
- [SupplementaryTable1Donordemographics.xlsx](#)
- [SupplementaryTable2Annotationclassification.xlsx](#)
- [SupplementaryTable3Encodelibrariesused.xlsx](#)
- [SupplementaryTable4GOKEGGmotifs.xlsx](#)
- [SupplementaryTable5HeartGWASSNP.xlsx](#)
- [SupplementaryTable6Figure6GWASmotifs.xlsx](#)
- [SupplementaryFigures.pdf](#)