Specific Aims

We propose a *Center for Functional Validation and Evaluation of ENCODE Enhancer regions*. We will employ assays to evaluate both sufficiency and necessity of candidate regulatory elements on transcription. Broadly speaking we will perform two types of assays. First, we will use regulatory element (enhancer) reporter assays to test for *sufficiency* of candidate non-coding DNA sequences to modulate gene expression, relying on STARR-seq for high throughput implementations and on genomic integration of smaller numbers of predicted elements to be tested in biological models. Second, we will use mutation/genetic-engineering assays to test for *necessity* of predicted enhancer sequences for normal expression of endogenous genes, relying on CRISPR technology with quantitative RNA measurements in both populations of cells and at the single-cell level. Several related variations on each of these assay types will be applied.

In addition to existing ENCODE data and cell lines, the Center will study two biological systems relevant to human disease. The first biological system, implemented by the White and Roggin labs at the University of Chicago, uses freshly resected primary and malignant pancreatic tissue to grow organoids in three-dimensional culture. The second system, developed by the Quertermous and Snyder labs at Stanford, uses iPS cells that are differentiated into smooth muscle cells grown and studied in a disease phenotypic state (phenotypic modulation). These two different biological systems were chosen for the relevance to inherited traits (coronary artery disease) and to somatically acquired traits (pancreatic cancer). Genetic variation plays a major role in both of these biological systems, with mounting evidence for the role of non-coding cis-regulatory genetic variation. We will also use ENCODE immortalized cell lines and cancer cell lines for comparison, and as tools for whole genome and rapid assay of putative enhancers and the effects of genetic variation. We envision the Center effort being equally divided among four parts: (1) coronary-artery and smooth-muscle studies, (2) pancreatic-tumor/normal organoids, (3) ENCODE cell lines, and (4) applying our approaches to a common set of elements decided on by the ENCODE consortium and tested across all of the Centers (as designated in the RFA).

Aim 1. Using ENCODE and other public data sets to identify regulatory elements for testing

First, we will identify high-confidence active regulatory regions from ENCODE and other public datasets for downstream functional analysis. Second, we will use GWAS and whole genome sequencing data available for samples related to coronary artery disease (CAD) and pancreatic cancer. Finally, in conjunction with the ENCODE Data Analysis Center and Analysis Working Group, we will select tools to deploy for further computational analysis to refine the definition of enhancer elements, including those developed by Gerstein and colleagues. Using these three sets of computational predictions as a guide, in Aims 2 and 3 we will investigate candidate enhancers, as well as the effects of inherited and somatic variation, using the CAD and pancreatic-cancer models, respectively.

Aim 2. Testing for enhancer sufficiency using enhanced STARR-seq

For candidate enhancer *sufficiency* we will use variations of the STARR-seq high throughput reporter assay in cell lines, in human 3-dimensional tissue models of pancreatic cancer, and in human smooth muscle cell models of coronary heart disease. These variations include the use of whole genome screening, capture-based screening, and site-directed mutagenesis to assess the impact of synthetic or naturally occurring mutations predicted to effect enhancer function.

Aim 3. Testing for enhancer necessity using CRISPR mutagenesis

For candidate enhancer *necessity* we will use CRISPR mediated mutation of candidate enhancers in their endogenous chromatin state. Two main variants of CRISPR mediated putative enhancer mutation will be pursued. In the first variant we will generate mutations in putative enhancers using a 96-well plate format and qRT-PCR of nearby genes to generate quantitative transcriptional read out. In the second variant we will use Drop-seq paired to CRISPR in order to get a high throughput, single cell resolution assessment of enhancer mutations. This second variant has the potential to create an “all-by-all” matrix of enhancer-by-transcription unit effects. We will also coordinate with other Centers who are taking similar approaches with complementary assays in other biological systems.

Aim 4. Testing selected human enhancers *in vivo*

We will use mouse transgenic models to test a limited set of enhancers that are validated in Aims 2 and 3, and that are also implicated in genetic risk for coronary artery disease or in recurrent mutations in pancreatic cancer. For CAD we will test both high risk and low risk alleles for differential function, and for pancreatic cancer we will test both normal and mutated variants of each candidate enhancer assayed.

Research Strategy

A. Significance

Great strides have been taken in the last 30 years toward understanding the regulation of genes. Concepts developed originally to describe the DNA sequences that control gene expression in prokaryotes and in viruses, such as promoters and enhancers, have been extended and refined to create working models of how eukaryotic genes are expressed. The decoding of model organism and human genomes, coupled with technology developments over the last 15 years, have radically altered our ability to systematically map and interrogate candidate sequences in the non-coding genome for gene regulatory functions. Scientists today are presented with unprecedented opportunities to discover and validate the candidate regulatory elements that drive the expression of genes in each biological cell type, tissue, normal or diseased system of interest.

The ENCODE project has seized upon these technological advancements, and in doing so has created high resolution mappings of chromatin modifications, transcription factor binding sites, chromatin accessibility, gene expression, and other necessary data for the genome-wide identification of candidate regulatory elements such as enhancers. Other projects have contributed extensive mapping efforts as well, including the Epigenomics Roadmap and the GTEx consortia – along with countless other individual researchers worldwide who, driven by the goal of unraveling the complexities of gene expression in their particular biological systems of interest, collectively have made even greater contributions to technology development and to detailed maps and functional validations for particular cell types.

However, presently we are at the cusp of another leap in the ability to systematically characterize and validate gene regulatory elements. Technologies have emerged that allow testing of gene regulatory elements, including *STARR-seq* (self-transcribing active regulatory region sequencing) and *CRISPR (*Clustered regularly-interspaced short palindromic repeats) Cas9 - based genomic mutation and engineering. The challenge that lies immediately ahead involves using ENCODE and other related data to accurately predict which candidate DNA elements will have biological activity, and then applying these testing/validation approaches in large scale and efficient assays that can be extended and generalized to fit a wide range of biological applications.

Why focus a center on enhancers?

Encoded within the DNA regulatory elements that drive gene expression are the genomic algorithms that are at the root of each cell's identity. Enhancers are one of the most potent and abundant classes of such regulatory elements. Importantly, much of the specific information that leads to cell type-specificity and that is associated with complex (and some Mendelian) human diseases appears to be encoded in enhancers. However, while hundreds of thousands of candidate enhancers have been predicted by the ENCODE consortium and other investigators, and much genetic variation (inherited and somatic) has been mapped to enhancers, a relatively small fraction of these DNA sequences have been tested and validated for function or biological relevance.

Transcriptional enhancers are often short stretches of DNA (a few hundred to a few thousand base pairs) that are able to modify transcription from the promoter of a target gene. Enhancers were originally discovered in simian virus 40 (SV40) (1, 2), but were subsequently identified within much more complex loci such as the mammalian immunoglobulin genes (3-5). In the 1980s it became widely recognized that such enhancer sequences acted as key determinants of cell-type-specific gene expression, which led to launching of many projects in model organisms and human cells to clone and characterize these sequences. It was discovered that enhancers can exert their effect over long distances of thousands, even millions of base pairs, either from upstream, downstream, or from within transcription units (6-8). Most human genes, and those of other multicellular organisms, are thought to be controlled by several enhancers that dictate expression at different developmental stages, in different cell types and in response to different signaling cues.

The ENCODE Consortium, along with work by many individual labs, has resulted in thousands of data sets for genome-wide transcription factor binding sites (TFBS), and other chromatin-associated factors, in a wide range of human cell types. TFBS sometimes cluster in a non-random fashion, and particularly when these clusters co-occur with certain chromatin states and marks, they are often considered to be good candidates for *cis*-regulatory modules (CRMs)(9). The prefix "*cis*" defines these elements are located in the same DNA molecule as the regulated target. CRMs are organized in a "modular" style, and they can regulate transcription in an additive, or sometimes non-additive, manner (10). Based on their functions and mechanisms of action, CRMs can be classified into enhancers, silencers, promoters, locus control regions (LCRs), and insulators (9). The ENCODE project, and its sister model organism ENCODE (modENCODE) projects, have made great strides in systematically mapping these various classes of CRMs. For example, White and colleagues led the effort in *Drosophila* to systematically identify CRMs in the fruit fly genome (11-14). Gerstein, Waterston and colleagues

produced similar maps in *C. elegans* (15, 16). Snyder, White and Gerstein have been among a large team of researchers in the human ENCODE project who have focused in producing CRM maps for the human genome(17).

Among the various types of CRMs, enhancers were initially defined to regulate the transcription of target genes in a location and orientation independent manner (6, 18). However, there is significant evidence, as quantitative and detailed measurements of enhancer function have been made, that location and orientation may sometimes affect enhancer function (19, 20). Enhancers are predicted to be the most abundant class of CRMs in the mammalian genome(21), and often function in a highly cell/tissue specific way (22) compared to other types of CRMs. They can also reproduce highly restricted temporal and spatial expression patterns *in vivo (23, 24)*, suggesting that enhancers are a major contributor of tissue/temporal specific gene expression patterns that are vital in development, and in human disease (25).

Enhancers are typically a few hundred base pairs long, and are often close to the transcription start site (TSS) of the gene they regulate, but they can also be far away from their target gene(18). For example, the wing margin enhancer in the *Drosophila cut* locus is 85kbp upstream of its promoter (26). In human cells, systematic mapping of chromatin interactions between TSSs and candidate enhancers by Dekker and colleagues has revealed widespread long-range interactions as well, with an average “long-range” candidate enhancer distance of 120kbp (27). How do enhancers regulate genes that are so far away in the genome? Studies have shown that some enhancers can directly interact with the promoter of their target genes through chromosome looping(28, 29), facilitated by utilizing targeted tethering of looping factors(30). One recent study on the human β-globin locus showed that looping factor GATA1 is critical for enhancer interaction with the promoter of the β-globin gene. Without GATA1, enhancer activation of β-globin gene is abolished. However, engineering an artificial zinc-finger to tether the enhancer to β-globin promoter in the absence of GATA1 activates transcription substantially, and the removal of this artificial tethering zinc-finger again abolishes this activation. This study provided direct evidence that the chromosome looping/tethering event is critical in enhancer-promoter interaction and regulation at the β-globin gene expression(31).

Major efforts have been made to map enhancers on a genome-wide scale, including large-scale efforts such as the ENCODE project (22), and the FANTOM project (32) and the Roadmap Epigenome Consortium(33). Most of these studies rely on one or more enhancer markers to identify them. These markers include: chromatin accessibility markers such as DNase I hypersensitive site (DHS) (34)and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)(35); chromatin marks such as H3K4me1, H3K4me2, H3K9Ac, and H3K27Ac(36); transcription factor binding, such as p300/CBP co-activators(37); and enhancer RNAs(32, 38). Enhancers can have many transcription factors bound to them (Highly Occupied Transcription factor binding, HOT, regions) or few (Low Occupied Transcription factor binding, LOT, regions) (12, 15, 17, 39). One common finding from enhancer mapping studies is that comparing to other groups of CRMs, enhancers are highly tissue-specific(22, 24, 40), and temporally dynamic(41). These tissue specific and temporal dynamic patterns of enhancer activities are tightly correlated with their target gene expression patterns *in vivo*, suggesting their important roles of directing expression changes throughout mammalian development and disease (7, 42). These studies established a landscape of active enhancers in human and other organisms, and provided a foundation for further studies in enhancer function.

While the vast majority of candidate enhancers identified in the human genome have not been functionally tested or validated, a rapidly growing body of evidence indicates that variation in human enhancers plays an important role in human disease. For example, the vast majority of genetic variation associated with human complex diseases mapped through genome wide association studies (GWAS) is found in the non-coding genome (43). Accordingly, chromatin marks and states associated with candidate enhancers have been shown to be pervasively abundant in GWAS loci across a wide range of diseases (44), useful in fine-mapping of complex traits (e.g. (45)), and to be associated with gene expression variation (e.g. (46)). Enhancers have already been functionally implicated in GWAS loci for coronary artery disease(47-52), obesity(53), and diabetes(54, 55), and cancer(56-59). Furthermore, in some cases, such as TCF21 in CAD, expression quantitative trait variation mapped to target genes has been shown to be significantly enriched for variants with low P-values in the GWAS analyses, suggesting a possible functional interaction between TCF21 binding and causal variants in other CAD disease loci (60). With the growing body of evidence implicating gene regulatory sequences in complex human disease, it has become increasingly important to develop methods and approaches to characterize the potential functions of enhancers and the effects of inherited variation contained within them.

Intriguingly, the identification of recurrent somatic mutations in enhancers has emerged as a key driver of cancers, a discovery enabled by data from large-scale and whole genome sequencing projects. Initial

examples of CRMs frequently mutated in cancer involved promoters, such as activating mutations in the TERT gene promoter in myelomas (61, 62). However examples of enhancers soon followed. For example, in a subset of T-cell acute lymphoblastic leukemias a specific recurrent mutation creates a novel MYB binding site upstream of the *TAL1* oncogene (63). This novel MYB binding site results in creation of a "super enhancer" that recruits chromatin acetylation and gene activation transcriptional co-factors, thus driving the over-expression of *TAL1*. Subsequently large numbers of whole genomes have been scanned for recurrent non-coding regulatory mutations, including initial work by Gerstein and colleagues using early versions of algorithms described in subsequent sections (64), by Weinhold et al and Fredriksson et al. across hundreds of whole genomes and more than a dozen tumor types (65, 66), and subsequently accompanied by functional assessments of enhancer mutations by Snyder and colleagues (67). In addition to the frequently mutated non-coding regulatory sequences, many studies have shown that the epigenomic patterns associated with enhancer activity are altered in cancers, reflecting the altered gene expression states. For example, signatures of colon cancer have been derived from profiling of epigenomic enhancer marks genome-wide (68). Thus characterizing enhancers in the context of cancer is of critical importance both for understanding the altered gene expression patterns that typically accompany cancerous lesions, and for understanding the mechanisms by which somatic mutations in enhancers lead to cancer development and progression.

B. Innovation

The proposed Center will require innovation on several levels. First, each Center contributor will build and extend upon conceptual and methodological approaches that their individual laboratory has advanced over the course of the last decade or longer. Over the last three years White's lab has made important refinements on the STARR-seq technique in collaboration with Stark's lab, and the Center represents an opportunity to extend and hone their preliminary work on optimizing whole genome STARR-seq and variants of capture STARR-seq on more complex models of disease than have been examined to date. Similarly, very little work has been done on utilizing the CRISPR Cas9 system for high throughput candidate enhancer validation. Advances by our team in creating ENCODE cell lines expressing Cas9 promise to open up the opportunity to scale this methodology, while the incorporation of Drop-seq with new bead chemistry and pooled gRNAs in Aim3 we believe is a highly innovative approach that stands to revolutionize our ability to screen candidate enhancer mutations for transcriptional consequences at the whole genome level. Additionally, Gerstein's group will apply the cutting-edge tools that they have developed as part of the ENCODE project to nominate the candidate enhancers that will be tested experimentally by the other Center investigators. Nobrega brings the latest methods for testing human enhancers in mice for the handful of disease relevant candidate enhancers that pass successfully through our tests in human cellular and organoid models of disease.

Second, the biological models that are brought to the Center by the participating laboratories are state of the art. The Quertermous and Snyder groups from Stanford will study patient derived iPS cells differentiated into coronary smooth muscle cells as a model of coronary heart disease, while White and Roggin from University of Chicago bring the latest technology in pancreatic cancer organoid culture from freshly resected and patient derived xenograft models. Both the Stanford and U. Chicago groups have extensively studied their biological model systems of disease on a genomic level for chromatin accessibility, chromatin marks, transcription factor binding and transcriptional profiling, providing ample data to combine with ENCODE data for predicting candidate enhancers in these disease models.

Third, we believe that the Center is conceptually innovative. The model systems were chosen expressly as exemplars of the two major types of disease-causing genetic variation in non-coding regulatory regions of the human genome, inherited risk factors and somatic genetic mutation. Besides representing two diseases that have tremendous impact on human health and society, coronary artery disease and pancreatic cancer represent biological systems where either inherited or somatic genetic variation play major roles. Thus by refining our technologies and approaches on ENCODE cells and then intensely studying these two biological model systems of disease, our Center aims to develop generalizable and state-of-art approaches that can be applied by others to the wide swath of complex human disease that include both inherited multigenic diseases and human somatically acquired cancers. In both these categories of genetically rooted disease there is growing and undeniable evidence for a key role of non-coding regulatory genetic variation, including genetic risk-causing polymorphisms and cancer-driving mutations in enhancers.

Finally, for us perhaps the most exciting aspect of the proposed Center is the unique combination of investigators, all of whom have made important contributions to the study of genome-wide study of gene expression in the human genome, most of whom have been participants in the ENCODE project that has

produced the data enabling this Center, and all of whom have a track record of successful projects together in a pair-wise fashion but for the first time will work together in combination as a team focused on the common goal of developing and applying generalizable methods for characterizing the role of enhancers in human disease.

C. Approach and Preliminary Data

The purpose of each ENCODE Functional Characterization Center is "to develop and apply generalizable approaches to characterize the role of candidate functional elements identified the ENCODE project in specific biological contexts". Our proposed Center will focus on characterization of candidate enhancer elements. We will develop, refine and apply experimental methods for functional assays of enhancers. We will use two very different biological models chosen for their high potential to act as generalizable exemplars for the study of enhancers in the context of (i) inherited risk factors for disease, and (ii) somatic mutations involved in cancers. We will also develop and refine our experimental methods in ENCODE cell lines, and we will reserve 25% of our efforts for testing candidate genomic elements that will be studied in common across all of the ENCODE Functional Characterization Centers. Using STARR-seq, and variations thereof, we will test for sufficiency of candidate enhancer elements to modulate gene expression. Using CRISPR-Cas9 methods we will edit the human genome, testing for necessity of candidate enhancer elements in their endogenous context. We will utilize these methods to examine the effects of inherited DNA variation on enhancer function in models of coronary artery disease (CAD), and to examine the effects of acquired somatic DNA mutations on enhancer function in models of pancreatic cancer (Pancreatic Ductal Adenocarcinoma – PDAC). While our approach necessarily requires a bioinformatics component to utilize ENCODE and other existing data sets in order to define the best candidate enhancer elements for testing in the specific biological models we will assay, our Center will be focused on experimental characterization of enhancers, testing different combinations of approaches in order to create extensible and generalizable protocols for systematic and accurate characterization of enhancer function.

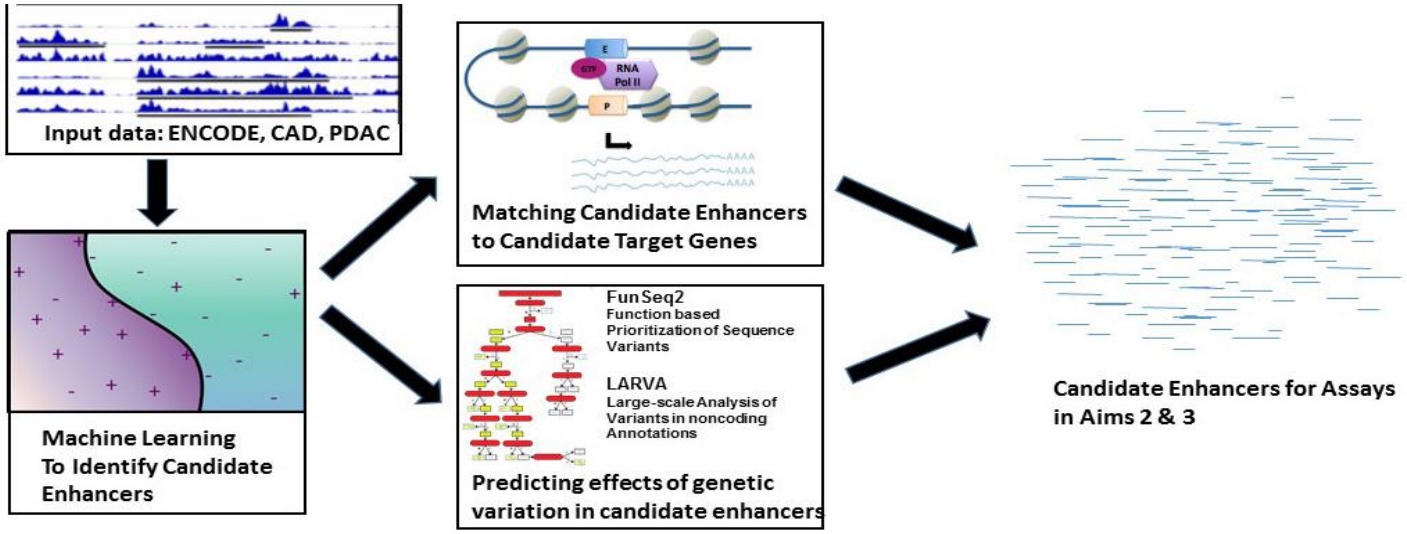
Aim1. Using ENCODE and other public data sets to identify regulatory elements for testing

Intelligent identification of candidate enhancers in specific biological contexts requires the appropriate integration of existing ENCODE data with the appropriate cell-type and disease specific data sets. We expect that our Center will interact and collaborate with the ENCODE Data Analysis Center (DAC) and the Analysis Working Groups (AWGs), and Center investigators Gerstein, White and Snyder have a considerable track record of participating in, and collaborating with, the existing ENCODE DAC and AWG groups. However, there is a need in our Center for a modest level of dedicated bioinformatics effort in order to focus specifically on integrating ENCODE data with datasets from coronary artery disease and pancreatic cancer, and in order to interface with the ENCODE DAC and AWGs. Additionally, logically we must first identify and choose the candidate enhancers we will test, before we perform experimental assays to test them. We also recognize that new data will emerge during the course of the Center grant, both for the disease models we will focus on and from the ENCODE mapping centers. Other work from our own laboratories, as well as from the larger community, will produce more refined maps of CAD and PDAC models that in turn lead to the opportunity for more refined enhancer predictions. Furthermore, the very scale at which we will test candidate enhancers will lead to new data sets that can be used for better enhancer predictions. For example, whole genome STARR-seq data in K562 cells produced by Kevin White's laboratory are already contributing to the algorithms being developed by the ENCODE DAC and functional characterization AWG. Our Center will be well positioned to take advantage of these and future developments to choose the best sets of candidate enhancers for experimental functional characterization in our two disease-relevant models. Therefore, Aim 1 will be focused on candidate enhancer identification for testing in Aims 2 and 3. Initially, we will simply finalize our preliminary results to pick a set of candidate enhancers to experimentally test. However, as additional outside data of relevance is produced, and as algorithmic approaches are improved as part of other efforts, we will apply such data and approaches to refine and improve the quality of the candidate enhancers we will examine in Aims 2-4. We will <u>not</u> develop new computational methods for predicting enhancers as part of the Center, but we will instead apply the latest methods developed as part of other efforts. Identification of targets in Aim 1 is expected to represent 15%, or less, of the total Center effort.

Figure 1 outlines the high-level work flow of our approach for nominating candidate enhancers for functional characterization testing in the Center. ENCODE and related existing data, including from CAD and PDAC, are used as the starting point. These data include histone, chromatin accessibility, expression profiling and transcription factor ChIP-Seq experiments. Machine learning algorithms are applied to identify candidate

enhancers genome-wide. The resulting candidate enhancers are further processed by computational pipelines that identify genetic variants and predict their effects on candidate enhancer function (required for Aims 2 & 3), and by algorithms that match candidate enhancers to candidate target genes (required for Aim 3 assays of effects of CRISPR enhancer mutation on endogenous genes). All of these algorithms are already developed and are routinely applied to ENCODE and other datasets such as whole genome cancer sequences.

Figure 1. Overview of computational pipeline for identifying and annotating candidate enhancers for STARR-seq and CRISPR-Cas9 assays.



1.1.1 Existing datasets:

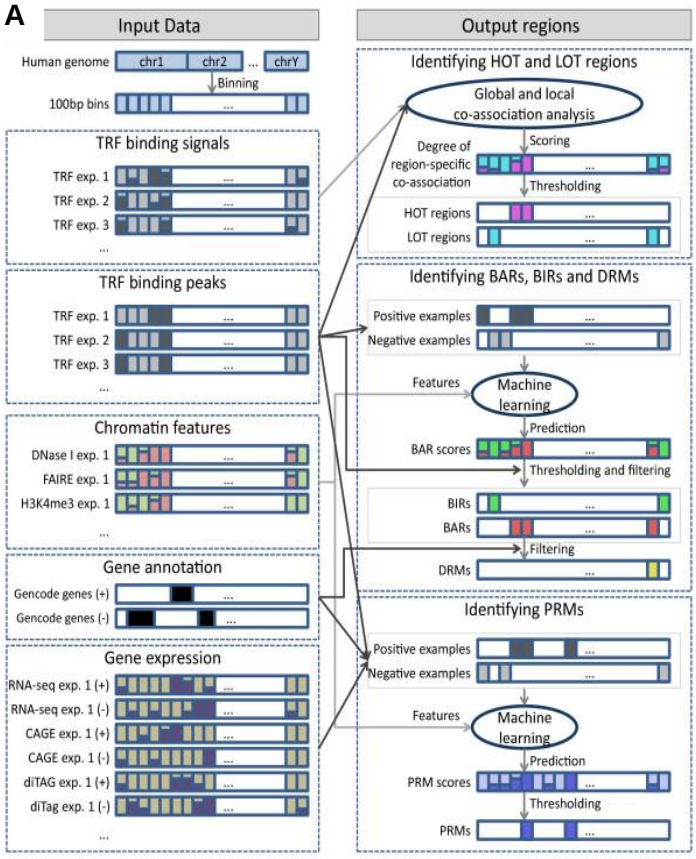
We will begin by integrating existing ENCODE data and results with existing data and results from CAD and PDAC that include (but are not limited to) chromatin modification profiling, chromatin accessibility profiling, transcription factor mapping, whole genome sequencing, GWAS, and eQTL studies. From these analyses we will identify an initial set of candidate enhancers that we will characterize, each annotated with results from one or more of the aforementioned data types. The Gerstein lab, with extensive experience developing and applying computational pipelines to predict elements from sequencing data, will lead this effort. Datasets that we will initially apply are listed in tabular form below (Table 1).

Cell/Tissue Type	K4Me/K27Ac/EP300	Dnase/Faire/5C	RNAseq	GWAS	Roadmap K4me/K27ac/Dnase
GM12878 (ENCODE)	6[b]	5[b]	>10[b]	13[c]	2[a]
K562 (ENCODE)	12[b]	11[b]	>10[b]	20[c]	2[a]
HeLa S3 (ENCODE)	5[b]	5[b]	9[b]	2[c]	0
HEPG2 (ENCODE)	4[b]	5[b]	>10[b]	1[c]	6[a]
HUVEC (ENCODE)	14[b]	5[b]	2[b]	n/a	0
A549 (ENCODE)	4[b]	1[b]	>10[b]	47[c]	4[a]
MCF-7 (ENCODE)	8[b]	11[b]	>10[b,d]	49[c]	1[a]
SK-N-SH (ENCODE)	7[b]	3[b]	>10[b]	4[c]	0
Pancreas	3[b,d]	3[b,d]	10[b]	9[c]	10[a]
Heart	15[b,e]	13[b,e]	>10[b,e]	60[c]	28[a]

Table 1. Identified ENCODE and GWAS datasets for analysis. Existing data generated by ENCODE and GWAS studies have been identified. The number of each type of study, the marks identified and the cells/tissue associated with the data. Datasets are procured from the a) NIH Roadmap Epigenomics Consortia (www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics, Bernstein Nat Biotech 2010), b) ENCODE consortium (www.encodeproject.org, ENCODE Consortium Nature 2012), c) GWAS studies (www.gwascentral.org), and d) internal data from the White and e) Quetermous labs.

1.1.2 Genome-wide identification of candidate enhancers (in support of Aims 2 & 3):
Using the ENCODE and auxiliary datasets outlined in Table 1, we will predict high-confidence candidate enhancers for downstream functional analysis. Prior to inputting the data into our machine learning algorithms for candidate enhancer identification (described below), we have processed the datasets in Table 1 using tools developed by the Gerstein lab, PeakSeq (69) and MUSIC (70) which have been applied by the ENCODE consortium and to ENCODE and Roadmap Epigenomics Consortium (RMEC) data. The main bulk of the datasets that will be used are listed in Table 1, featuring existing functional genomics datasets from ENCODE and RMEC projects. Specifically, we will utilize peaks from histone marks and transcription factors and build a priori probability estimates for localization of the regulatory regions. We will use the activating marks and transcription factors that associate with enhancers (H3K4me1, H3K27ac, H3K9ac, P300, DNase/FAIRE) to build these probabilities. We will also utilize transcription factor binding motif and sequence conservation data as variants in the a-priori estimates of localization.

These data will be input into enhancer prediction algorithms. A variety of enhancer prediction methods have been employed by the ENCODE consortium to examine existing data sets. For example, Ren and colleagues developed a Random-Forest based algorithm, RFECS (Random Forest based Enhancer identification from Chromatin States)(71). Park and colleagues developed a supervised machine learning method to identify and classify enhancers using chromatin marks across multiple metazoan species studied in ENCODE and modENCODE(72). As part of the ENCODE and modENCODE projects, Gerstein, Snyder and colleagues have developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers(73), which they have partially validated(74) (Figure 2). In addition, the Gerstein lab has also developed a tool that utilizes the pattern within the histone marks to predict active regulatory regions in each tissue or cell line. The performance of some of these enhancer prediction algorithms was compared by the DAC as part of the ENCODE Enhancer challenge and the pattern recognition-based algorithm developed by the Gerstein lab was one of the top performing algorithms for enhancer prediction in mouse forebrain.



A. The left side of panel A shows input data types. The right side shows how these datasets were used to identify the regions. We constructed statistical models that capture the genomic features of three paired types of regions by machine-learning methods: firstly, regions with active (BARs) or inactive binding (BIRs); secondly, those with extremely high or low degrees of co-binding, termed HOT and LOT regions; and finally, regulatory modules proximal (PRMs) or distal (DRMs) to genes. From the DRMs, we developed computational pipelines to identify candidate enhancers, many of which were validated experimentally in (20). We further associated the predicted enhancers with potential target transcripts and the transcription factors involved.
B. Low resolution, genome-wide view of DRMs predicted as candidate enhancers from 5 ENCODE cell lines. 70% of the candidate enhancers are cell-type specific, with 30% shared between 2 or more cell lines. Inset (i) and (ii) show examples of regions of chromosomes that differ between cell lines in DRM density.

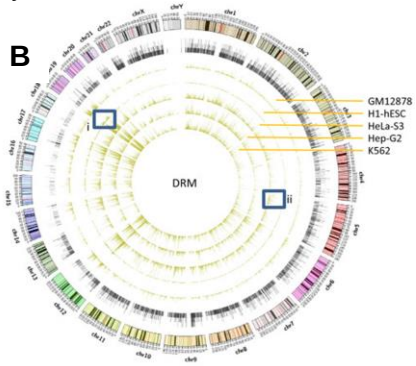


Figure 2. Overview of the pipeline for identifying CRMs and separating out candidate enhancers.

Some of the top performing methods for enhancer prediction in the ENCODE enhancer challenge will be used to predict enhancers initially. We will assess the datasets for each ENCODE cell line individually, and all in combination, and then we will compare the locations of candidate enhancers to the locations of candidate enhancers similarly identified in our CAD and PDAC models. From these analyses we expect to identify both cell/tissue-specific candidate enhancers and candidate enhancers that are shared across cell types, including among ENCODE cell lines and our CAD and PDAC models. For each cell line or tissue type we can expect tens of thousands of candidate enhancers, and in some cases more than 100,000, based on previous results(71, 72, 74). We also expect 50-70% of these enhancers to be cell type specific based on previous results (22, 74)**.**

Although we will initially use the approach outlined above for genome-wide candidate enhancer identification, it is worth noting that presently the ENCODE Functional Characterization AWG is evaluating this and other approaches for enhancer identification, with candidate enhancer validation experiments being performed by the Kevin White lab using whole genome STARR-seq and by the Len Pennacchio lab using in vivo mouse reporter assays. Our Center will apply the statistical and bioinformatics approaches used to nominate the candidate enhancers to be tested, we will use ensemble approaches to combine predictions from the best performing methods based on the results from the current ENCODE Functional Characterization AWG, and from other future Functional Characterization Centers in the ENCODE consortium with whom we will test a common set of candidate elements.

1.1.2 Annotating and classifying genetic variation in candidate enhancers (in support of Aims 2 & 3):
A key goal of our Center is to develop and apply functional assays for assessing the impact of genetic variation, both inherited and somatically acquired, on candidate enhancers. We therefore need to annotate and classify variation in candidate enhancer regions we intend to test.

Previously we have extensively analyzed patterns of variation in noncoding regions, along with their coding targets, creating the tool ncVAR for assessing genetic variation in TFBSs (75). In recent studies (62), we have integrated and extended these methods to develop a prioritization pipeline called FunSeq (and subsequently FunSeq2). FunSeq prioritizes variants with respect to their deleterious impact on many different types of noncoding functional elements, including TF binding sites, regulatory elements, and regions of open chromatin. It identifies the regions under strong selective pressure as estimated using the variant frequencies computed from the whole genome sequencing data in 1000 Genomes Project and uses these regions as sensitive and ultra-sensitive non-coding regions of the genome. For each noncoding mutation in a regulatory element, FunSeq analyzes the target of the affected regulatory element. Then it scores the impact of the variants and prioritizes them based on a number of factors like network connectivity and motif disruption. It identifies deleterious variants in many noncoding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. We will determine linkage of variants identified in candidate enhancers with variants identified in the GWAS datasets for CAD in Table 1. We will thus use FunSeq to annotate candidate enhancers.

We will next characterize the regulatory elements in terms of their association with human diseases. We will analyze GWAS and whole genome sequencing data available for diseases related to coronary artery disease and to pancreatic cancer, respectively. On this front, we have developed LARVA that can identify recurrently damaging non-coding mutations and prioritize them with respect to their significance. To estimate significance, LARVA utilizes models that estimate background mutation frequencies in non-coding elements using as features the functional genomics datasets from ENCODE and RMEC projects (Table 1). We will use the whole genome sequencing datasets from 1000 Genomes Project, and polymorphism datasets from dbSNP and Exome Aggregation Consortium (ExAC) projects as reference backgrounds to filter out the non-causative mutations. We will integrate the tissue specific expression quantitative trait datasets (eQTL) from GTex Project to generate evidence for the causal variants generated by the mutation STARR-Seq experiments. Using these computational predictions and characterizations as a guide for Aims 2 and 3, we will investigate examples of disease-related variation in the human genome that represent both inherited variation and somatic variation, using coronary artery disease (CAD) and pancreatic cancer respectively. Extensive tissue-specific allele-specific expression, vascular cell epigenome mapping, and in vitro functional studies identifying CAD causal variants and enhancers will be employed to validate the bioinformatic findings.

1.1.3 Predicting which candidate enhancers are potentially associated with which genes (in support of Aim 3):

In Aim 3 we will assay endogenous gene expression (using qRT-PCR and using high-throughput single cell sequencing) in response to CRISPR-Cas9 mediated mutations in candidate enhancers. To assist in formulating hypotheses about which genes are regulated by which candidate enhancers in each cell type we will examine, we will computationally characterize the regulatory regions by linking them with their putative targets.

We have previously developed computational pipelines for identification of targets of candidate regulatory elements, including methods that can successfully identify targets for gene-distal and gene-proximal regulatory elements. Our methods utilize the correlation between the gene expression levels and the activity of the regulatory element to identify significantly correlating activity. We will utilize sets of chromatin conformation datasets, generated from experiments including 4C, 5C, Hi-C, and ChIA-Pet. Furthermore, the Gerstein lab recently developed a method, named ENGINE, for utilizing these datasets in a machine learning framework for assigning targets to regulatory elements. ENGINE is a new version of the component of FunSeq that performs enhancer-target matching (http://papers.gersteinlab.org/papers/funseq2). Table 1 shows that there are many conformation datasets from ENCODE and RMEC datasets that we can utilize by combining the correlation based target estimation with conformation datasets. In particular, the conformation datasets will narrow down the possible targets of candidate enhancers to a subset of regions in which each candidate enhancer interacts. We expect this to dramatically decrease the false positive rate of the correlation. Specifically, ENGINE computes the correlation of the activity at the candidate enhancer region (using ENCODE and RMEC datasets) with the expression levels of the genes that each candidate enhancer region has contact with. Then, ENGINE uses the expression levels and several statistics about the shapes of the histone modification and transcription factor binding signals as additional features and builds a random forest based prediction model to score the candidate target genes. We will utilize the validated sets of enhancer-target gene linkages for training ENGINE. To increase the specificity of the model, we will build random backgrounds for distribution of the correlation levels and signal profiles and estimate significance of the linkage between the candidate targets of the enhancer region. In our previous efforts, we have utilized ENCODE data to build a set of such linkages between candidate regulatory elements and their target genes. For the work within the Center we will aim at stratifying the tissue specific behavior of the novel candidate enhancers using the tissue specific datasets for CAD and PDAC.

1.2.1 Human Pancreatic cancer analysis and preliminary results:

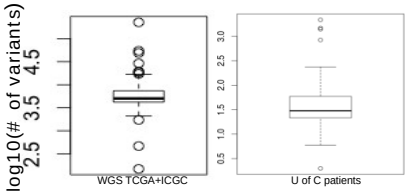
The Chicago Pancreatic Cancer Initiative (CPCI), led by Dr. White and in collaboration with Dr. Roggin in the Department of Surgery at U. Chicago, has a patient data set of 400 tumor/normal pairs at time of submission. Sample sources include ongoing retrospective and prospective studies at the University of Chicago as well as from the University of Rochester and North Shore University Health Systems. Pancreatic ductal adenocarcinomas (PDACs) comprise the vast majority of the pancreatic cancers being studied. Most samples have been collected from paraffin embedded slides and sequenced with a custom panel. However, a subset of samples has been subjected to whole genome sequencing. Importantly, as part of the CPCI we have developed a streamlined experimental pipeline for 3-dimensional organoid cell culture using both primary tumor and normal tissue that has been surgically resected. This organoid system also is routinely used for patient-derived xenograft (PDX) models. This tissue model is where we plan to functionally validate our characterized enhancers.

In addition to the datasets in Table 1, which include PDAC chromatin and histone mark profiling, we have assembled 224 whole genome sequencing (WGS) data sets from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). We have compared the matched tumor-normal pairs to identify the somatic mutations. In total 1,704,431 somatic variants were found in all samples, with an average of 5,041 variants per sample. We applied the same approach to data from the Chicago Pancreatic Cancer Initative and found similar numbers of mutations, indicating that the somatic mutation algorithms are consistent across different data sets. We represent this with a boxplot of the number of variants per sample is shown in Fig. 3A. In addition for the TCGA and ICGC data, we selected around 10k open chromatin regions (with average length at 1kb per region) as test regions for enhancer activity evaluation. In total, we found that 3,012 out of these 10k regions have at least one somatic variant (Fig. 3B). In a typical STARR-Seq experiment, around 20 to 50 percent of the test regions are expected to demonstrate some level of enhancer activity, hence we expect that there are 600 to 1,500 active enhancers with at least one somatic variant.

Furthermore, using the TCGA and ICGC data set we utilized 382 external genomic features to accurately estimate the background mutation rate for precise mutation burden evaluation. We found that our model performs well in pancreatic cancer. Indeed, the Pearson correlation between the predicted and observed

mutation rate can be as high as 0.94 (Fig. 3C). Based on this background mutation rate, we evaluated the mutation burdens on the promoters of protein coding genes, and discovered 11 promoters as highly mutated (Q-Q plots of P values in Fig. 3D). Many of them, such as TP53 and SMAD4, have been well documented as pancreatic cancer related. In the Center we will extend this analysis to distal candidate enhancer regions, and we will test these recurrent mutations in functional assays.

(A)

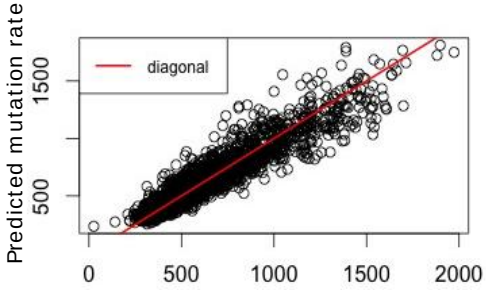


	UC total
Total sample number	75
Mean variant number per sample	121
Median variant number per sample	30
Min variant number per sample	2
Max variant number per sample	2164

(B)

# of Variants	# of regions
1	2327
2	528
3	111
4	37
5	8
20	7

(C)



(D)

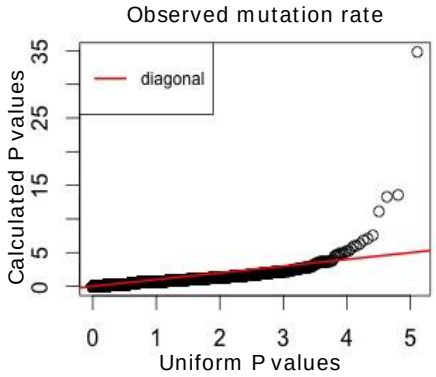


Figure 3. Preliminary analysis of Pancreatic data sets using our pipelines.
(A) Boxplots of ICGC and TCGA somatic mutations (left) and Chicago Pancreatic Cancer initative somatic mutations (right) with mutation calls (lower).
(B) Variants per region in ICGC and TCGA data identified in our pipelines.
(C) Pearson correlation between the predicted and observed mutation rate (ICGC and TCGA) = 0.94.
(D). mutation burdens on the promoters of protein coding genes = 11 promoters as highly mutated (Q-Q plots of P values.

1.2.2 Human smooth muscle cell analysis and preliminary results: The Snyder and Quertermous labs have generated extensive maps of chromatin accessibility, TF binding, enhancer histone modification, and gene expression (under basal and growth factor stimulated states) in primary cultured HCASMC and also in medial layers of normal and atherosclerotic coronary arteries from explanted human hearts (see Table 1). These coronary vascular genomic datasets have enabled us to prioritize mechanisms of candidate causal variants associated with CAD in GWAS meta-analyses and ultimately to generate hypotheses of potential causal signaling pathways that mediate CAD susceptibility in humans. For instance, we identified a candidate SNP rs17293632 (C risk allele) in the SMAD3 locus to create a consensus binding site for AP-1, leading to greater chromatin accessibility, AP-1 binding, and allele-specific expression in HCASMC (Fig. 4). We validated the SMAD3 enhancer activity in cells using gain and loss of AP-1 function in luciferase reporter assays and also in vivo by generating site-specific integrase mediated transgenic mice, which demonstrated allele-specific differences in LacZ reporter expression and localization in the developing vasculature. We also validated our candidate variants in a large eQTL cohort of human atherosclerotic aortic tissue obtained at the time of coronary artery bypass graft surgery, which further emphasizes the context-specific functional roles of these variants during CAD. We have now expanded our HCASMC collection to 60 individuals, which have undergone WGS and RNA-seq profiling for ASE analysis. Nonetheless, it still remains a challenge to identify causal variants for complex phenotypes such as CAD. For instance, the majority of these variants in regulatory regions in HCASMC are not predicted to alter known TFBS or have small effects on candidate gene expression, as determined by a combination of ASE, eQTL or reporter assays. To create highly credible sets of candidate enhancers for both CAD and pancreatic cancer using an unbiased prioritization scheme, we plan to run these overlapping regions and GWAS variants through a custom functional annotation pipeline that includes tools such as FunSeq, and LARVA, as described above.

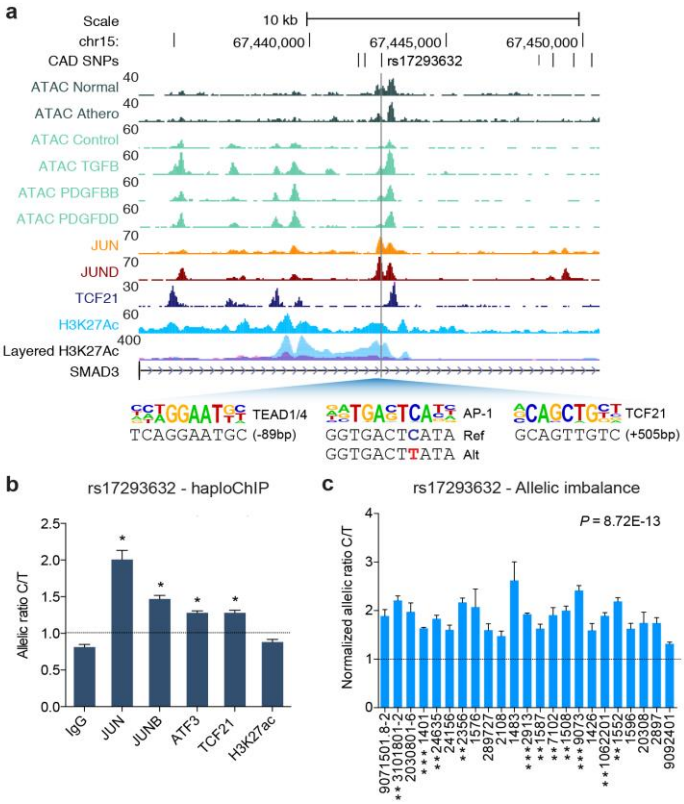


Fig. 4. Epigenome mapping and in vitro studies in HCASMC to identify coronary artery disease causal variation and disease enhancers. (a) UCSC browser screenshot at SMAD3 locus, showing overlap of candidate SNP rs17293632 with ATAC-seq open chromatin tracks in coronary medial tissue and HCASMC treated under various conditions, transcription factor binding ChIP-seq tracks for TCF21, JUN, and JUND, and active enhancer histone modification H3K27Ac ChIP-seq, as well as ENCODE layered H3K27Ac for HUVEC (blue) and NHLF cells (purple). Also shown, motifs in open chromatin regions with alignment to reference sequence and position relative to rs17293632. **(b)** Allele-specific ChIP (haploChIP) for AP-1 proteins (JUN, JUNB, ATF3), TCF21, and H3K27ac in HCASMC heterozygous at rs17293632. Values represent mean ± SEM of triplicates from a representative experiment (n=5). *P<0.01 versus Control, IgG or between two genotypes. **(c)** Allelic expression imbalance for candidate regulatory SNP rs17293632 at SMAD3 detected by TaqMan qPCR in HCASMC pre-mRNA from heterozygous individual donors. P-values shown represent comparison of AEI from all samples versus expected allelic ratio of 1.0 using a Welch's unequal variances t test. **P<0.001, ***P<0.0001.

Aim 2. Testing for enhancer sufficiency using enhanced STARR-seq

Validation of enhancer sufficiency will be investigated using variations of the STARR-seq high throughput assay in ENCODE cell lines, primary cell lines and in human 3D tissue models. Our variations and enhancements of the STARR-seq assay include the use of whole genome screening and capture-based screening to assess the impact of natural variation, and site-directed mutagenesis to assess the impact of naturally occurring or synthetic mutations that affect enhancer function. We have optimized this assay by developing large-scale transfection methods, optimizing transfection efficiency, and using barcoding primers to eliminate PCR duplicates and to increase measurement accuracy.

2.1 Background

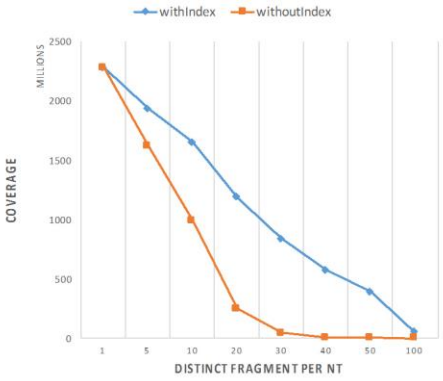
STARR-seq provides direct measurement of genome wide enhancer activity in a high-throughput manner, and was initially demonstrated using the *Drosophila* genome by our collaborator Alex Stark and colleagues (76) (see attached letter).(76)This method is, in principle, similar to massively parallel functional dissection (MPFD) and massively parallel reporter assays (MPRA), but differs by inserting enhancers into the transcript, instead of upstream of promoters in the reporter vector. The enhancer sequence effectively acts as a barcode in high-throughput sequencing. More specifically, genomic DNA is sheared and end-repaired, and subsequently cloned into screening vectors containing a promoter, and expresses a reporter transcript. The enhancers are cloned into the 3' end of the transcript, whereby the reporter transcript will contain the enhancer sequence. This pool of screening vectors is transfected into cells, mRNA is purified and reverse transcribed, and then sequenced using high-throughput sequencing. High copies of the reporter transcripts that contain specific enhancers can identify enhancers that are up-regulating transcription. STARR-seq removes the need for expensive array synthesis of enhancers.

A major limitation of this technique, prior to our optimizations, has been the transfection step: for example screening through the *Drosophila melanogaster* genome required transfection of between 0.5 and 1 billion S2 cells. This makes the direct application of STARR-seq technique to the human genome very difficult and expensive, because the human genome is 20 times larger than the fly genome. Our optimizations of STARR-seq in human cells modifies and builds upon the episomal plasmid library approach, expanding its capabilities. We have overcome the major challenges for scaling up STARR-seq to the entire human genome; namely required library complexity, large-scale transfection of cells, and inaccuracy of the assay due to PCR duplicates during the sequencing step. By optimizing multiple parameters in the candidate element cloning step we have increased complexity while introducing molecular barcodes that allow for PCR duplicate elimination, resulting in a screening library that covers 2.65 Gb of the human genome. Our typical libraries now on average have >50 fragments covering each base pair, given ~250 million post-filtering fragments. This represents a comprehensive screening library, and allows us to effectively screen genomic fragments with enhancer activity in downstream experiments. Using industrial scale transfection protocols, the White lab has now devised a robust technique to screen either the entire human genome or a fraction of the genome that has been captured using oligonucleotide probes. We are now able to produce whole genome STARR-seq data sets at >10X per expressed base pair coverage with 200-300 million paired end 100bp reads, or significantly fewer reads for capture STARR-seq. Table 2 shows the various cell lines and STARR-seq experiments that we have completed thus far.

Cell Line	Type	Replicates	Ave. Reads (millions)
GM12878	Capture	2	100M
GM12878	WG	2	535M
GM12878	WG-Safe	2	188M
K562	WG	1	500M
LnCAP	Capture	8	34M
LnCAP	WG	2	190M
SNU16	Capture	1	16M
OCUM1	Capture	1	29M
MCF-7	Capture	2	29M

Table 2. White lab STARR-seq datasets collected with sequencing depth and number of replicates shown

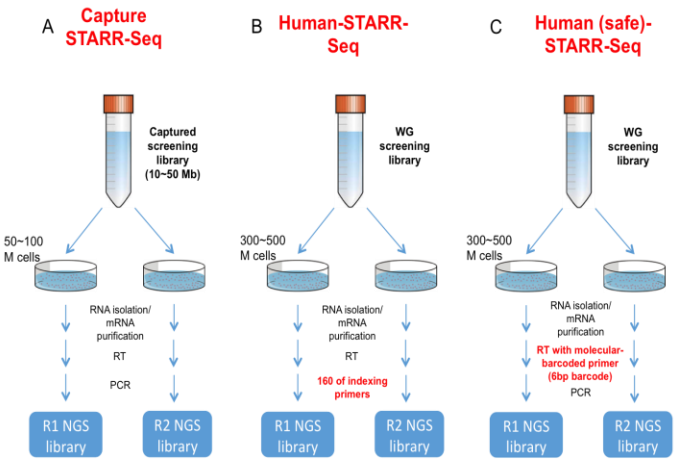
Fig 5. Genome-wide coverage of distinct fragments with (blue line) or without the 160 indexes (orange line).



2.2 Optimization of Human STARR-seq: Developing STARR-seq in human cells requires scaling the experimental and computational workflow to cover a larger genome. The White lab has optimized techniques to improve the experimental protocol and to increase the number of transcripts collected per cell. Collecting data at this scale requires an increase in the number of cells surveyed, an increase in the complexity of the libraries and a reduction in PCR artifacts. To address these three issues we; 1) enabled large-scale transfection using a BTX transfection system coupled with optimized transfection reagents and parameters that have been carefully tuned for each human cell line. Additionally, we introduced a GFP-control plasmid to monitor transfection efficiency and batch effects; 2) increased genomic coverage and library complexity; and 3) we applied two strategies to overcome PCR duplicates. The first PCR duplicate reduction strategy is to introduce additional index sequencing primers, which helps to distinguish the plasmid-transcribed-mRNA copy and PCR duplicates. Our current standard protocol is to introduce 160 different indexes (see Figure 5). The second PCR reduction strategy is to introduce barcoded cDNA primers during the initial reverse transcription phase of building the RNA seq libraries ("safe" seq). This represents the ultimate solution to distinguish the plasmid-transcribed-mRNA from PCR duplicates, since individual mRNA strands are barcoded prior to the reverse transcription reaction, in which every mRNA transcript is barcoded to represent its uniqueness. By also labeling the mRNAs with modified cDNA primers to include six base pair molecular barcodes we can essentially distinguish all individual transcripts. Others have applied similar strategies for traditional RNA-seq (43).

With these strategies, we can create around 100 million distinct fragments from the human-STARR-seq library after transfecting 500 million cells. Positive regions that represent enhancer activity in our assay are called from statistical tests based-on signals from human-STARR-seq library. This scale represents practical and robust

human genome STARR-seq and can be applied to captured genomic fragments, whole genome, or with the above modifications using barcoding for capture or whole genome "safe" sequencing (Figure 6). This powerful technique has been applied on two ENCODE tier-1 cell lines GM12878 and K562, and preliminary results for these cell lines has been collected and analyzed (see Preliminary Results below).



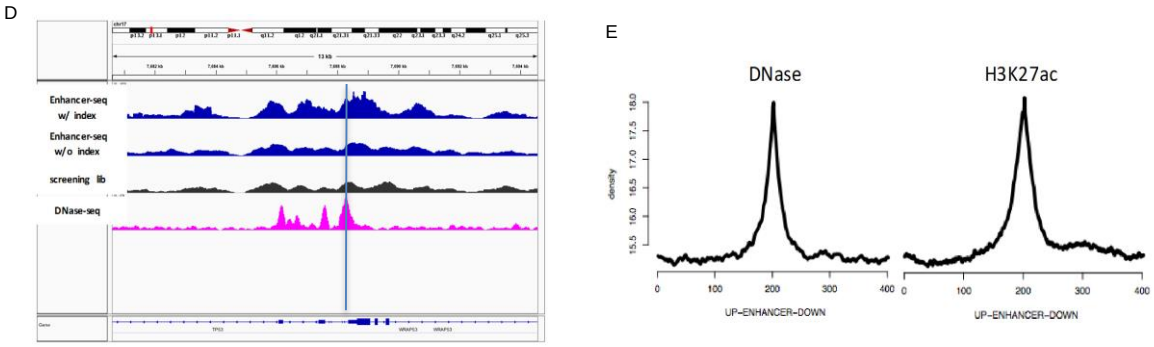


Figure 6. Diagram of White-lab approaches to STARR-seq: (A) Capture panels are created and transfected into human cells to create the screening library for sequencing, note this is not whole genome. (B) Human whole genome STARR-seq is prepared with indexing primers post reverse-transcription. (C) Human whole genome (safe) STARR-seq with barcoding at the time of post reverse transcription. In all cases we create at least 2 biological replicates for each data-set. (D) Genome Browser screen shot from whole genome STARR-seq of ENCODE GM12878 lymphoblastoid cells shows signals piled-up by raw sequencing fragments. 1st row: with index; 2nd row: without index; 3rd row: screening library; 4th row: DNase-seq data in the same cell line. (E) Enhancer-seq peaks align well with DNase-seq and H3K27ac ChIP-seq signals.

2.3 Whole genome STARR-seq data in human cells:

2.3.1 Preliminary data

Thus far we have generated data using whole genome STARR-seq using a diverse set of three different human cell lines. These included two ENCODE cell lines (GM17828 and K562 cells) and one cancer cell line (LnCAP androgen-sensitive prostate cancer cells). Example data from GM17828 cells are shown above in Figure 6 (panels D and E). In these experiments, approximately 500 million GM12878 cells were transfected and collected 24hrs post-transfection. RNA extraction, mRNA purification and cDNA synthesis were performed. In the final amplification step, cDNA templates were evenly divided into 160 distinct PCR reactions, each with a different index. We consider fragments from the same genomic location but with different indexes to be distinct and not PCR duplicates (note Figure 6D shows results plus or minus indexing, the blue line shows an example of a STARR-seq peak that was not apparent without the indexing). We identified 71,160 genomic regions that show enhancer activity in human GM12878, with average length of 213 bp and this represents 15.2 Mbp (~0.5%) of the human genome. A sliding-window based approach was used throughout the entire genome and positive windows were finally merged to give the final peak calls. We also applied the barcoding (safe)-STARR-seq approach to GM12878 cells. Although raw data quality was similar to our standard whole genome STARR-seq data, we found that we required 33% less sequencing depth using Safe-STARR-seq compared to the multi-indexing human STARR-seq (200 million pairs vs. 300 million pairs). For whole genome (safe)-STARR-seq in GM12878, we observed 90,058 genomic regions (covering ~20Mb) with enhancer activities, which is comparable to the 71,160 observed for the multi-indexing human STARR-seq performed in the same cell line. The signal of enhancer activities of these two approaches correlates well (PCC=0.5). Importantly, the whole genome (safe)-STARR-seq allows us to quantitatively estimate the artifacts brought by PCR duplicates. We estimated the copy number of plasmid-transcribed mRNA that is represented by the number molecular barcodes labeling each mRNA, and we found that 1.5% of plasmid-transcribed mRNA fragments were labeled by more than 10 molecular

barcodes given the current sequencing depth (200 million pairs), and 44 of the transcripts were labeled by more than 100 molecular barcodes. We then estimated artifacts from PCR duplicates, calculated from the number of sequenced reads that share the same molecular barcode and aligned to the same genomic location, and estimated relatively few artifacts from PCR duplication, with 78.3% of templates represented as distinct fragments.

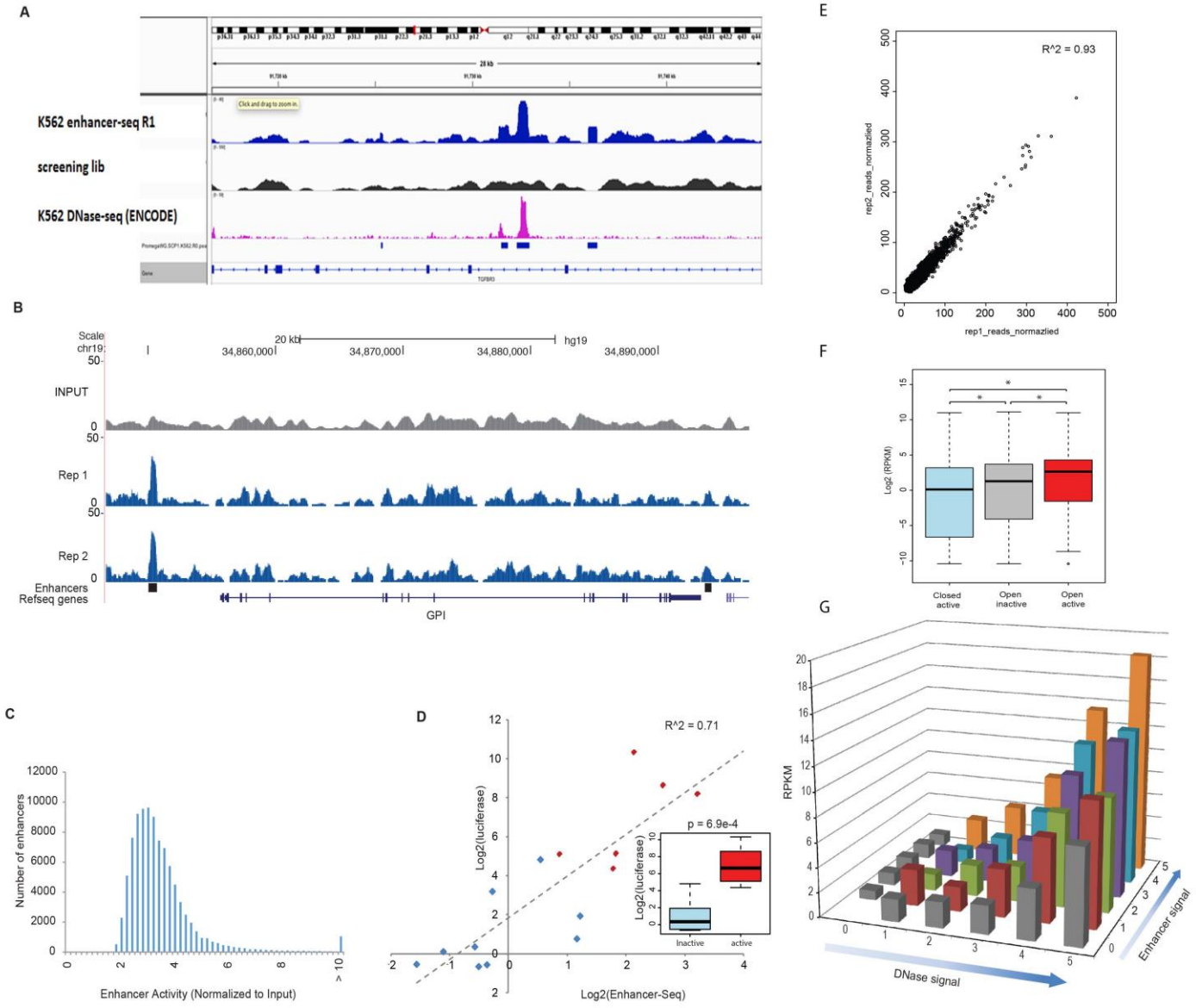


Figure 7. Whole genome STARR-seq (A) Genome browser screen shot shows consistency between K562 human-STARR-seq signal (1st row) and DNase-seq signal (3rd row). **(B)** Genomic snapshot displaying the *GPI* locus region as detected by WG-STARR-seq. There is a strong enhancer region approximately 10-kb upstream of *GPI* and another, weaker enhancer regions in the 3'UTR region. Each blue track signifies normalized enhancer signal of each biological replicate. The gray track represents the normalized input library. **(C)** WG-STARR-seq shows a wide range of enhancer signal strength distribution of all detected enhancers. The median fold change observed was 3.08, with a dynamic range between 1.33 and 119.12. **(D)** The enhancer activity of 6 strong and 9 weak enhancers were validated using traditional luciferase assays in biological triplicates. A strong correlation was observed between luciferase signal and WG-STARR-seq enhancer activity, providing validation of the technique. **(E)** Normalized reads from sequencing were used for reproducibility plots between biological replicates. **(F)** Comparison of expression levels of genes (denoted as RPKM) nearby different groups of enhancers. Statistical significance was calculated using Wilcox Sum Rank test (*p = 2.2e-16). **(G)** Plot comparing expression level of nearby genes in relation to both DNase I signal and enhancer activity. Both DNase I signal and enhancer signals are binned into 6 separate groups according to DNase I signal and enhancer signal rank (0 – 5), respectively.

Similarly we have performed whole genome STARR-seq in K562 and LnCAP cells using the multiple-indexing approach. Figure 7A shows a strong overlap between K562 STARR-seq peaks and DNAse-seq data from the ENCODE project. Whole genome STARR-seq in LnCAP cells (Figure 7B-G) illustrates high reproducibility (panels 7B and 7E), with an r-squared value of 0.93 between replicates. We identified 97,527 enhancer regions that were significantly enriched over the input library (MACS2, q-value < 0.05) and that displayed a median enhancer enrichment activity of 3.1 with most enhancers showing enrichment scores between 2.0 and 5.0 (Figure 7C). In order to validate the function of the active enhancers determined by our whole human genome STARR-seq, we randomly selected 15 regulatory regions previously reported to be nuclear receptor binding sites, and measured their activity using traditional *Renilla* luciferase reporter assays. As seen in Figure 7D, we observe a strong correlation (R^2 = 0.73) between enhancer activity and luciferase reporter signal, with most of the luciferase validated enhancers also having strong WG-STARR-seq enrichment (fold change > 2.0). Alternatively, most regions that did not have positive STARR-seq enrichment did not test positive in the luciferase validation assay (blue dots).

We also have investigated the relationship between chromatin context, enhancer strength and nearby gene expression using whole genome STARR-seq data. We first analyzed nearby gene expression level to enhancers of varying activity and chromatin accessibility. We observed the expression level of genes near active, inaccessible enhancers were lower than the expression of genes near accessible sites lacking enhancer activity (regions lacking WG-STARR-seq signal) (Figure 7F). However, we found the highest expression of genes nearby accessible WG-STARR-seq called enhancers. We then categorized all active and accessible enhancers into 36 separate groups (ranked 0 – 5), based on their DNase I signal and enhancer activity, respectively (Figure 7G). We found that the average gene expression level of genes nearby STARR-seq enhancers increases as one of the variables (DNase hypersensitivity or enhancer activity) stays fixed. However, the highest gene expression values were observed for genes nearby strong STARR-seq identified enhancers that also had high DHS levels. Since the STARR-seq assay is episomal in nature, outside of the native chromatin context, these results suggest that STARR-seq is a sensitive measure of the intrinsic potential of an enhancer to affect target gene expression. These results also indicate that quantification of STARR-seq assay reflects the biological activity of enhancers on their target genes.

2.3.2 Whole genome STARR-seq approach

In years 1 and 2 of the Center, we will continue to focus our whole genome STARR-seq efforts on validating candidate enhancers from the ENCODE project immortalized and cancer cell lines (see Table 1). Thus initially our whole genome STARR-seq efforts will not be focused on the disease models studied by our Center (CAD and PDAC), which we will only attempt to examine using whole genome STARR-seq after we have tested several rounds of capture STARR-seq using a more restricted set of predictions from Aim 1 in these model systems. Our rationale for focus on the ENCODE immortalized and cancer cell lines is two-fold. First we wish to optimize the assay in the CAD and PDAC models using capture STARR-seq due to the lower requirements for transfection efficiency. This two phase approach (capture then whole genome STARR-seq) is similar to how we have approached the cell lines we have assayed thus far. Second, the results from whole genome STARR-seq in ENCODE cell lines will provide valuable data that will help the DAC and overall consortium to develop better predictive algorithms and better genome-wide enhancer annotations. Thus, as the project progresses, results from our Center will positively affect the success of all other Centers that are focused on enhancer characterization as all or part of their work.

Accordingly we will assay the eight ENCODE cell lines in Table 1 during the first two years of the Center. We will perform additional replicates in GM12878 and K562 cells, as these have been “flagship” cell lines of ENCODE where a large and diverse set of data have been produced over the last decade. We will also add MCF-7, HeLaS3, HepG2, HUVEC, A549 and SK-N-SH cells to the repertoire of whole genome STARR-seq data sets. Experiments will be performed as described above for GM12878, K562 and LnCAP cells. At least two replicates will be performed for each cell line, with 4-6 total replicates performed for GM12878 and K562 to examine the statistical power effects of performing deeper replication.

Cell characteristics and media conditions are listed in Table 3. We already have experience growing both adherent and suspension cells in the White lab. Suspension cells are grown in T175 flasks with 100ml of media and split 1:2 when the density hits 1M cells/ml. Adherent cells for transfection will be grown in Hyperflasks (Corning), which enables growth of 300-400M cells per hyperflask. These methods allow for maximal cellular growth to obtain the large numbers of cells needed for transfection.

Table 3. ENCODE Cell Types for analysis: characteristics and media conditions

Cell Name	Tissue of Origin	Cell type	ENCODE Tier	Obtained From	Media
GM12878	Lymphoblast	Suspension	1	ATCC	IMDM + 10% FBS
K562	CML	Suspension	1	ATCC	IMDM + 10% FBS
Hela S3	Cervical Adenocarcinoma	Adherent	2	ATCC	F-12K + 10% FBS
HEPG2	Liver Carcinoma	Adherent	2	ATCC	MEM + 10% FBS
HUVEC	Umbilical Vascular Endothelium	Adherent	2	Lonza	EGM-Plus
A549	Lung Carcinoma	Adherent	2	ATCC	F-12K + 10% FBS
MCF-7	Breast Adenocarcinoma	Adherent	2	ATCC	DMEM + 10% FBS
SK-N-SH	Bneuroblastoma	Adherent	2	ATCC	MEM + 10% FBS

Before beginning STARR-seq protocol, cell types in Table 3 that have not been used in this assay before will undergo transfection tests using GFP, to determine transfection efficiency in our BTX AgilePulse Max electroporation machine (Harvard Apparatus). This device uses electronic waveform pulses instead of single intensity shocks to open cells then force in the plasmids. This method allows for a higher density of transfection (10-100M cells/ml) and high cuvette capacity (2-6mls at once) than other electroporation methods. These tests will not only test transfection efficiency but optimal transfection density. Our current methods include transfection parameters for GM12878/K562 (100M cells/ml) and MCF-7 (30M cells/ml). We also have methods for transfecting MCF10A, T47D, and HEK293T cells. We will test the cells in two types of electroporation buffer: the BTX electroporation buffer T4 recommended by Harvard apparatus and the Lonza kit buffer that would be used on their nucleofector machine. We have had success using both types of buffer in a cell-dependent manner. Each cell type with transfection efficiency of >60% will first be used for whole genome enhancer-seq. Cells whose transfection efficiency <61% will immediately be used for the capture method.

For whole genome STARR-seq, 500M cells will be used for each replicate. Suspension cells will be collected through centrifugation while adherent cells will be collected via trypsinization. Cells will then be combined with electroporation buffer and whole genome library, electroporated, and plated back in their growth media. The whole genome library consists of the promega human male genomic DNA being sheared into 500bp regions and cloned via Gibson assembly into our screening vector. The 500bp region is placed downstream of a luciferase transgene. This library is then sequenced to assure good coverage of all of the regions within the genome. 24 hours after transfection for suspension cells or 48 hours for adherent cells, the cells will again be collected and frozen in liquid nitrogen. 5M cells from each replicate will be taken for downstream RNA-seq. RNA will be extracted using the RNeasy Maxi kit (Qiagen). mRNA will be isolated using Dynabeads (ThermoFisher) followed by DNAse treatments to rid the sample of genomic and plasmid DNA contamination. cDNA synthesis from the mRNA then occurs using SuperscriptIII reverse transcriptase and a reverse transcriptase primer specific for the luciferase gene, thereby only causing cDNA synthesis of our putative enhancer regions in our screening vector, and not cellular mRNAs. This is also a critical step as the primer also contains a molecular barcode. RNAse is then used to remove all remaining RNA from the samples and cDNA samples are purified with the MinElute PCR purification kit (Qiagen). Finally PCR amplification is used to complete the library.

The key to our system is the molecular barcode added to each individual mRNA. This allows for removal of PCR duplicates after sequencing but affords us the ability to keep each single mRNA that was amplified. We use a 6 bp barcode creating 4,096 possible combinations and our ability to deconvolve up to 4,096 single mRNAs to the same region at single base pair resolution. This barcode also allows us to sequence for fewer reads for the same level of fold change calculations. For each whole genome enhancer-seq we plan to perform a minimum of two biological replicates.

2.3.3 Whole genome STARR-seq analysis

With these whole genome STARR-seq data sets in hand, and working as appropriate in coordination with the DAC and other interested groups in the consortium, we will take a systematic approach whereby we will process the STARR-seq signal profiles with PeakSeq (69) and MUSIC (70) to generate a relaxed set of regions that show significant enrichment. These regions will be very highly sensitive but will contain many false positives. Therefore we will use the large compendium of existing functional genomics datasets from ENCODE and RMEC projects listed in Table 1, utilizing peaks from histone marks and transcription factors to build a priori probability estimates for localization of the regulatory regions. We will use the activating marks and transcription factors that associate

with enhancers (H3K4me1, H3K27ac, H3K9ac, P300, DNase/FAIRE) to build these probabilities. We will also utilize transcription factor binding motif and sequence conservation data as variants in the a-priori estimates of localization. Next we will combine the whole genome STARR-seq results with these probabilities of localization in a Bayesian framework, and we will train generalized linear models for scoring the candidate relaxed list of regions that we identified from STARR-seq. The sorted list of regions will be provided to the experimental groups for further validation. This approach can be used for tissue, cell line, and species-specific STARR-seq scoring models as described using cell line, tissue, and species-specific datasets. For our Center, we will concentrate specifically on the vascular smooth muscle and pancreas for the downstream experimental validations that we will perform on these tissues.

Timeline: We expect to complete this component of the project by the end of year 2. We will add additional cell lines, or perform additional experiments as needed by the overall ENCODE consortium in years 3 and beyond.

2.4 Capture STARR-seq:

2.4.1 Preliminary Results

The application of whole genome STARR-seq requires 60% of cells to be transfected; this makes the technique difficult for some applications, such as primary cells and some cell lines. For such cells, we are able to utilize capture libraries (Figure 8). The method we will employ uses a solution-based genomic capture system to enrich for genomic regions that corresponds to putative regulatory elements (identified in Aim 1). This significantly reduces the complexity of the library and allows for fewer cells to be transfected. This method is referred to as Cap-STARR-seq. Thus far we have performed Cap-STARR-seq on six different cell lines including GM12878, K562, LnCAP, MCF-7, SNV16 and OCUM1 cells.

The Cap-STARR-seq library utilizes in-house designed Nimblegen capture probes. After genomic DNA shearing and hybridization to the capture probes to isolate specific regions of the genome, the resulting fragments are cloned into the plasmid. Using this method we typically screen 10-100MB of the genome instead of all 3GB. This technique will be particularly useful for the smooth muscle CAD model and PDAC 3D culture model cells.

The experimental implementation of Cap-STARR-seq similar to whole genome STARR-seq, described above, and cells are processed in the same way. Unlike the whole genome STARR-seq, where the distribution of plasmids entering the cells should be equal, in Cap-STARR-seq it is important to get a more accurate reading of the plasmids present in the population, since our experience has been that slight changes to the library can lead to larger fold change differences and high variability between replicates. It is for this reason that for each Cap-STARR-seq cell type we will perform 3 biological replicates.

One example of how we have used capture-STARR-seq was to test a set of candidate enhancers for estrogen-responsiveness in Estrogen Receptor positive MCF-7 breast cancer cells. In this experiment we probed the regulatory activity of close to 10,000 DNase I hypersensitivity sites (DHSs) in MCF-7 cells. We enriched a total of ~10 Mb regions using NimbleGen SeqCap EZ customized capture probes, and we cloned them into a screening library driven by the SCP1 minimal promoter. The screening library encompassed nearly all the target regions (99.8%) with low off-target rate (11%). Deep sequencing results indicated the library had high complexity. We transfected screening library plasmids into MCF-7 cells followed by 48-hr hormone deprivation and 10 nM E2 or vehicle treatment (40 million cells per replicate per condition), and enriched screening library transcribed mRNA for high throughput sequencing. Sequencing of these libraries required 30 million paired end 100bp reads, which led to an average depth coverage of 300X. We were able to detect more than 1,600 high confident and reproducible STARR-seq peaks from the ~ 10,000 candidates. In addition to calling active enhancers, we were able to further measure estrogen-regulated enhancer activity by comparing enhancer activity in estrogen vs. vehicle treatment. At as early as 45 min post estrogen treatment, which is too early for conventional luciferase assay to detect any change, we were able to determine 245 E2-upregulated enhancers and 180 E2 down-regulated enhancers. E2-upregulated enhancers are significantly associated with E2-upregulated genes (Fisher's exact p = 0.0002) and enriched for ER binding sites (Figure 9).

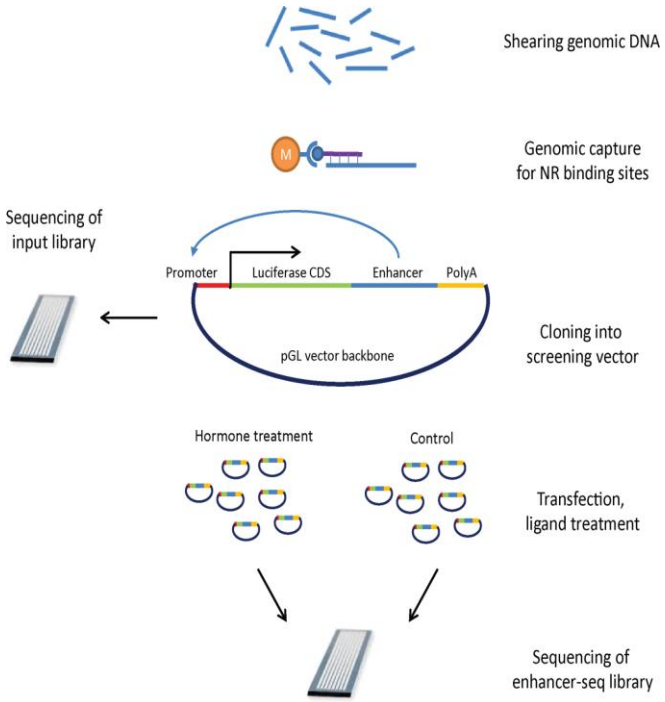


Figure 8. CAP-STARR-seq protocol. Diagram showing process of generating capture libraries for capture based STARR seq. In this example, cells have been differentially exposed to a ligand (see Figure 9)

a

Enhancers	w/ERBS	w/o ERBS	Fisher's exact p
E2-up	86	159	1.86e-7
E2-down	24	156	

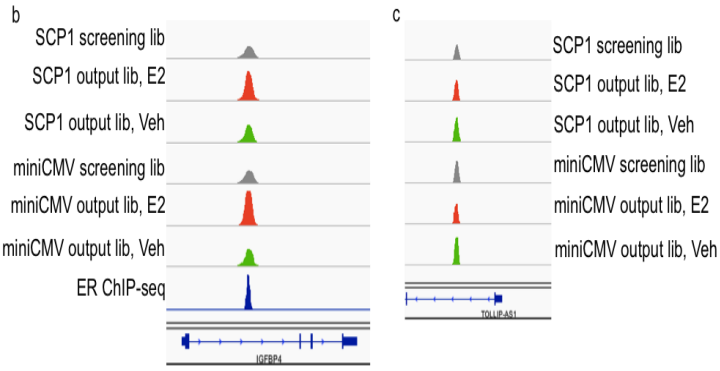


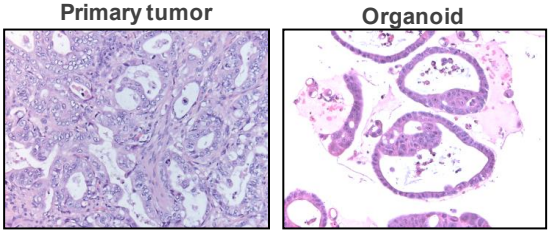
Figure 9. Estrogen regulated cap-STARR-seq enhancer assay. a) ER binding site occurrence in E2-upregulated enhancers and E2-downregulated enhancers. b) Snapshot of E2-upregulated enhancer. c) Snapshot of E2-downregulated enhancer.

2.4.2 Capture STARR-seq approach

Determination of regions captured for our CAD and PDAC biological models will be made based on enhancer predictions in Aim 1 using ENCODE, RMEC and CAD or PDAC data from Table 1, as well as conditioned on results from the whole genome STARR-seq in cell lines as it becomes available. As described in Aim 1, we will also consider recurrent somatic mutation data in non-coding candidate enhancer regions for pancreatic cancer and GWAS data for CAD, based on analyses performed by the Gerstein lab. These candidate enhancers will be captured from total human genomic DNA (we generally capture from a mixed pool of normal DNA isolated from human blood samples), and they will be cloned into plasmids or lentiviral vectors as described in the previous section and below.

Capture STARR-seq in pancreatic normal and cancer organoid models: The Chicago Pancreatic Cancer Initiative (CPCI) is sequencing cancer genomes and developing tools to study the molecular basis of this disease. One of those tools is the pancreatic organoid. The team takes fresh surgical specimens and creates 3-dimensional cell culture models of the tumor and the normal tissue. They also produce patient-derived xenograft (PDX) tumors in mice, which in turn provide a large supply of material for organoid culture. These organoid models provide an ideal substrate for experimental interrogation. In particular, we propose to use these organoids as a tool to test the functions of candidate enhancer regions predicted in Aim 1. Figure 10 shows that these organoid models can have many characteristics in common with primary tumors from the same patients. We histopathologically compared an organoid to the same patient's primary tumor using formaldehyde fixed, paraffin embedded sections that were H&E stained and scored for mitotic index, cytological appearance, and histological grade, or stained for biomarkers of pancreatic cancer, and scored for expression and localization by a board-certified gastrointestinal pathologist. Both the primary tumor and organoids had a mitotic activity of 5-7/10 HPF, a moderately differentiated histologic grade, and irregular nuclear membranes, open chromatin and prominent nucleoli. These findings suggest that organoids are reliable patient-specific pathophysiological models to study pancreatic cancer.

A



B

	Normal Ducts	Primary Tumor	Organoids
CK7	Focal + Cytoplasmic	Strong + Cytoplasmic	Strong + Cytoplasmic
CK20	Negative	Negative	Negative
CEA	Negative	Strong + Cytoplasmic	Strong + Cytoplasmic
p53	Negative	+ Nuclear	+ Nuclear
Claudin-4	Negative	+ lateral membranes	+ lateral membranes

Figure 10 (A). Similar gross morphology of malignant tumor cells and organoid cells from the same patient (B) Both primary tumor and organoid demonstrated diffuse moderate-to-strong positivity for CK7, CEA, p53, and Claudin-4, and focal weak positivity for CK20.

For the proposed experiments we will use both PDX derived organoids and primary tumor/normal organoids, as available. Pancreatic cancer surgical resections provide the fresh precursor material to grow organoid cultures. In collaboration with surgical oncologist Dr. Kevin Roggin, we have refined the procedure to grow and manage normal pancreatic and PDAC organoid cultures. Immediately following surgical resection, pancreatic tissue and tumor are placed in individual 15ml conical tubes containing digestion media for 10 minutes incubation. The tumor is minced into 1-2mm pieces in a petri dish and transferred back to the 15ml conical tubes containing the 10 mL of digestion media, and incubated at 37ºC for 1.5 hours. To break up large pieces of tumor, tissue and media are triturated with a 10ml pipette. Cells are pelleted by centrifugation and pellet is re-suspended in media containing DNase and incubated at 37ºC 10 minutes. The reaction is stopped and cells are collected by centrifugation. The pellet is re-suspended in cold Matrigel (Corning), which is a reconstituted basement membrane containing a rich assortment of laminins, collagen, growth factors, and other basement membrane components. 50μl of the Matrigel suspension is then plated onto a pre-warmed 24 well plates as a droplet, which will form domes upon polymerization at 37°C. The plate is then placed in a 5% CO2 tissue culture incubator for 20 minutes. Finally, cells are bathed in liquid medium. Within 1-2 days, organoids form hallowed spheroids. Media is changed every 2-3 days, depending how fast organoids grow, and organoids are passaged once a week.

We will use a retroviral version of our pGL4.23.ccdB STARR-Seq screening vector, where the ccdB expression cassette is cloned out of pGL4.23-ccdB into the multiple cloning site of pMSCV-loxp-dsRed-eGFP-Puro-WRE, a validated retroviral expression system for use in organoid cultures (77, 78). An advantage of this plasmid is that expression of libraries can be induced by addition of Cre recombinase at any time during organoid development, preventing any possibility of toxicity from overexpression or any growth advantage from non-induced cells. Addition of Cre removes the dsRED cassette, but retains EGFP, allowing for easy visual confirmation of EGFP expression to confirm successful induction by Cre.

Organoids for viral transduction will be prepared in 24-well plates, using one well per transduction. Organoids are transduced by gently combining organoid fragments with the 250μl retroviral supernatant in one well of a 48- well plate. Cells are “spinoculated” by spinning the plate at 600 x g for 1hr at 32°C, then incubating the plate for 6 hours. Fragments are with matrigel and resuspended by gentle pipetting. Matrigel containing infected organoids is then seeded in a new, pre-warmed 24-well plate. The plate is then incubated for 2-3 days. Puromycin selection (1μg/ml) begins after 2-3 days. When the fragments start to form organoids, medium is replaced with medium supplemented with puromycin. Budding structures will be evident within 1-2 weeks. At this time, expression of the libraries is induced with the addition of 4-OHT, and successful induction is determined by the loss of dsRed signal and maintenance of EGFP signal.

Our initial rounds of cap-STARR-seq will be performed in PDX organoids due to their abundant supply. Once the protocol is fully tested, candidate enhancer libraries will be transduced into patient-derived primary 3-D cultured human pancreatic cells (tumor and normal). In this second round of STARR-seq assays we will validate the positive findings from the initial screening, but more importantly we will identify the extent to which there are differences in enhancer activity between normal pancreatic organoids and tumor organoids. Additionally, recent results from Baily et al. analyzing over 450 pancreatic cancers have identified novel molecular subtypes that can be distinguished by expression pattern(79). We are collaborating with the authors of this work to categorize each of our CPCI patient tumors into these molecular subtypes. Our Center will therefore also the question of whether tumors of different molecular subtypes show different enhancer activities. In the mutation STARR-seq assays proposed in the next section of Aim 2 and in Aim 4, we will separately address the question of whether somatic mutations lead to differential candidate enhancer activities.

Capture STARR-seq data for the organoid models will be analyzed similarly to whole genome STARR-seq data, as described above.

Capture STARR-seq in a coronary artery disease model: Atherosclerotic coronary artery disease (CAD) is the world-wide leading cause of death, not only in high-income countries, but also increasingly in developing countries. Through genome wide association studies (GWAS) and the recent meta-analyses of such GWAS data, over 60 associated loci have been validated, and approximately 200 additional loci identified at an FDR<0.05. Despite the phenomenal progress in identifying genetic loci that harbor genes associated with CAD and other complex human disease, there has been limited progress toward deciphering the genetic and molecular mechanisms by which causal variation regulates gene function (80). The majority of disease-associated variation has been found to reside outside of structural genes and presumed to regulate gene expression rather than encoded protein sequence (44, 81-85). The lack of functional annotation of these regulatory regions of the genome accounts for the lack of progress toward identifying and understanding the mechanism of effect for such variation. Given the public health crisis associated with this disease, and the robust genetic association findings that promise to point to disease mechanisms, we have chosen to incorporate the study of vascular cells and tissues in this application.

Vascular smooth muscle cells (SMC) are increasingly recognized as pivotal contributors to the vascular response to injury and genetic dissection of signaling pathways that modulate the response of this cell type to disease initiating stimuli would contribute significantly to our understanding of human vascular disease (86). The risk of CAD events is inversely correlated to the number of vascular smooth muscle cells (SMC) in atherosclerotic plaque, and it is speculated that this cell type stabilizes the lesion to protect against plaque rupture and myocardial infarction (86-90). In fact, SMC contribute up to 80% of cells in atherosclerotic plaque (91) and express a majority of CAD-associated genes(92). For these reasons, we have focused our studies investigating the mechanisms of genetic CAD risk on loci encoding genes that are fundamentally linked to SMC biology, e.g. previously establishing causality and investigating the SMC biology of CDKN2B at 9p21.3(93, 94) and TCF21 at 6q23.2(47, 48, 92, 95). These cells never fully differentiate but rather possess a phenotypic plasticity that allows them to respond to epigenetic signals and to downregulate the classic SMC contractile marker gene expression program in the setting of stimulation by disease related factors(96). This process is postulated to promote medial SMC downregulation of lineage marker expression, proliferation, and migration to the luminal surface of the atheroma where these cells contribute to the plaque stabilizing fibrous cap(86, 91). This phenotypic switch is dependent on chromatin remodeling, and can be modeled in vitro by growing SMC in the presence of serum (proliferative, undifferentiated, disease state) or absence of serum (quiescent, differentiated, physiologic phenotype)(96). This in vitro differentiation model may be used to bolster in vitro studies aimed at dissecting transcriptional mechanisms by which CAD-associate causal variants contribute to risk. Human vascular cells grown in culture may represent a disease or healthy phenotype, with related chromatin organization and gene expression that may or may not represent the functional environment in which the causal regulatory variants impact disease risk related processes. Evaluating chromatin structure, transcription factor binding and transactivation in SMC under both disease and physiologic conditions should increase the likelihood that in vitro experiments have in vivo relevance, and that disease related pathways can be identified and studied. Thus, because of the significant contribution of this cell type to CAD risk, and the apparent involvement of causal variation with enhancer function, we will investigate through these studies the enhancer makeup of SMC, and correlate findings to our previous epigenetic and mechanistic studies.

As background for the studies proposed here, we have used ATAC-seq to map regions of open chromatin, performed ChIP-seq for histone modifications (H3K4me1, H3K4me3, H3K27ac) and disease relevant transcription factor binding in primary cultured HCASMC, and combined these and ENCODE datasets to promote the identification of causal variants and genes in CAD associated loci (see Aim 1). In fact, ChIP-seq studies for JUN, JUND, JUNB, FOS, CEBPB, and SMARCA4 were conducted in HCASMC as a pilot using ENCODE antibodies and methodology, and these data will be uploaded as part of the ENCODE dataset. We have identified 10 causal variants in enhancer regions in CAD loci, some of which have been validated with in vitro CRISPR studies and in vivo transgenic mice(49). These data will provide an important opportunity for validation of the bioinformatic and computational approaches described in Aim 1, as well as the STARR-seq and other approaches described here.

Although studies with SMC have provided for significant new insights into the mechanisms of gene expression and in particular those associated with risk for complex disease such as CAD, these cells senesce after a limited number of passages and cannot be transfected at sufficient efficiency to allow primary STARR-

seq enhancer screening. An alternative approach that allows the continued access to unlimited numbers of cells from the same human subjects with defined genotype and disease risk profiles are currently provided by the derivation and culture of iPSC. As for a number of other cell types, iPSC can readily be differentiated into SMC populations(97, 98). Also, they can be transfected to >60% efficiency. Importantly, by employing different chemically defined conditions, it is possible to reproducibly differentiate induced pluripotent stem cells (iPSC) into distinct SMC populations that reflect spatial specificity that normally results from different embryonic origins(97). A recently derived method allows for the robust and rapid differentiation of human iPSC into vascular SMC, and we have streamlined this approach to generate SMC in 5 days(98). .

We will source iPSC lines from our collection generated through our GENESiPS study (U01 HL107388) that is part of the NHLBI funded Next Generation Genetic Association Studies (Next Gen) consortium (RFA-HL-11-006). Through this effort, we have generated 3-6 clonal iPSC lines for each of 200 subjects that have been phenotyped for CAD risk factors and undergone whole genome genotyping with imputation to 1000 genomes. The iPSC lines were created from erythroblasts using the non-integrative Sendai virus system, passaged to allow clearance of Sendai virus and growth in feeder free conditions. The lines have been extensively characterized for markers of pluripotency (Tra1-60), sample identity and genomic integrity. We are currently re-consenting GENESiPS subjects with an open consent form according to recent ENCODE recommendations to allow sharing of cells and data. Enhancer mapping using cap-STARR-seq and other methods described below will be performed in iPSC-derived SMC, iPS-SMC, with transfection efficiency approaching 60%.

For differentiation, iPSC are plated as single cells with the Rhokinase inhibitor Y-27632, subsequently differentiated to mesoderm using CP21 and BMP4 for 2 days, followed by treatment with N2B27, PDGF-BB and activin A and then maintained in PDGF-BB plus heparin. We can easily culture the needed number of cells and perform the Cap-STARR-seq transfection employing the Amaxa system with Lonza reagents. In our pilot studies, after 5 days the fully differentiated SMC maintain significant levels of reporter gene (luciferase) expression, and we will further validate that the level of RNA reporter expression is sufficient for STARR-seq. We will thus transfect cells, immediately initiate the differentiation protocol, and begin the assay after 5 days of in vitro differentiation. An alternative approach would be to conduct the transfection with a selectable drug selection marker, drug select pools of transfected iPSC and subsequently put them through the differentiation protocol. We will pilot these two approaches to determine which is the more viable method.

Enhancers identified through the initial Cap-STARR-seq screen will be validated by transfection of relevant reporter constructs (libraries) into patient-derived primary cultured human coronary artery smooth muscle cells (HCASMC). In this second round of Cap-STARR-seq assays we will validate the positive findings from the initial screen, employing an in vitro SMC differentiation paradigm that recapitulates the phenotypic modulation that occurs in vivo in the disease setting. The in vitro disease phenotype is induced by culturing cells in high percentage serum, or including growth factors such as platelet derived growth factor BB in the culture medium. Cells grown in 1-2% serum are quiescent, and express numerous SMC markers including MYH11, ACTA2, and TAGLN. By growth in serum or PDGF-BB, the cells adopt the disease phenotype, by 48hrs downregulating the SMC markers and increasing proliferation rates. While these cells are not suitable for the initial high-throughput STARRseq screen, they can be electroporated with the relevant Amaxa protocol to achieve >30% transfection of the reporter constructs, which is more than adequate for the secondary STARR-seq screens. Evidence for enhancer function in the phenotypic modulated cells but not in the serum starved cells would argue for disease relevance for enhancer function.

Timeline: In years 1 and 2 we will exclusively focus on capture-STARR-seq for the PDAC and CAD model systems. We will begin testing whole genome STARR-seq in year 3 for these model systems, but we will continue the cap-STARR-seq approach throughout the course of the Center grant, continually testing refined predictions from Aim 1. We will also integrate mut-STARR-seq in the next section, starting in year 2.

2.5 Mut-STARR-seq: Combining site-directed mutagenesis with STARR-seq to assay the effects of genetic variation on candidate enhancer funcion: Site directed mutagenesis has long been an invaluable tool in studying specific mutations in the context of protein-structure interactions and mutational effects on target gene expression. This technique is general used to generate DNA sequences that induce mutated codons, insertions or deletions. To generate desired mutations, PCR reactions are conducted using a pair of oligonucleotide primers that amplify a targeted region of interest, designed in such a fashion where mismatching nucleotides are located in the center of the primers. The resulting PCR product is the target region that contains the mutation of interest. This region is subsequently cloned into a vector for downstream experimental analysis.

Unfortunately, current standard of protocols for site-directed mutagenesis require the picking and processing of individual colonies for downstream Sanger sequencing to identify the correct clone. Additionally, these current methods are low-throughout and expensive to conduct.

The White lab has extensive expertise in this area of molecular biology. His team has led the recombinant engineering of transcription factor tags for the ENCODE consortium, for use in ChIP experiments conducted in the White, Farnham and Snyder labs. Our production has focused largely on tagging transcription factors by recombineering to generate the TF-GFP fusion within Bacterial Artificial Chromosomes (BACs). We created over 1,000 tagged constructs using this technique. More recently we have adopted the CRISPR/Cas system for our tagging needs. Using this technique we have tagged 42 pairs of constructs for the ENCODE consortium. Site-directed mutagenesis is very similar to the protocols we have been using thus far. Using site-directed mutagenesis we will generate mutations within our STARR-seq libraries. By using our STARR-seq libraries and mutagenic DNA oligo primers made in batch similar to our Nimblegen capture probles, we will create "batches" of mutants in STARR-seq libraries and screen them in the same way as non-mutated libraries. From these data different mutational variants for the same candidate element can be distinguished computationally when processing sequencing results. These mutations differ in contrast to established random mutagenesis pipelines, since it is not possible to separate one mutant sequence from another in random mutagenesis protocols, and as such the whole pool can only be used from the same assay together. Others(99) have used this type of approach to study proteins. Overall, this approach represents a rapid and cost-effective method to screen mutations. Those mutations can be based on somatic mutations identified in PDAC or on inherited polymorphisms associated with CAD GWAS.

To perform this mutation STARR-seq (mut-STARR-seq) we will use the same candidate enhancer regions, identified and annotated in Aim 1, that are used in the previous sub-Aims. We will systematically test the function of these candidate enhancers when mutated to contain precise somatic mutations identified from primary PDAC or the exact polymorphisms identified in CAD GWAS. In order to quality control this technique, we will perform our STARR-seq enhancer evaluation pipeline up until the step where we will clone our regions of interest in our vector system. Before doing so, we will perform batch site-directed mutagenesis to introduce mutations in our target enhancer regions, and then proceed to cloning them into the vector. This way, we will be testing at the same scale as our enhancer evaluation pipelines, the effect of the mutation on each enhancer region as compared to wild-type. Just as we do for our cap-STARR-seq libraries, we will sequence the resulting libraries to determine the baseline coverage of each candidate enhancer, as well as the frequencies of introduced mutations (we expect a mix of mutated and wild type fragments). This experiment will provide us with an initial assessment of which enhancers may be affected by single base or small indel changes, and as a result may provide a way to priortize further testing of specific enhancers in the functional validation assays we plan to explore in Aims 3 and 4.

We will plan to perform this technique on all of the same CAD and PDAC models we will perform cap-STARR-seq on. Other than the site-directed mutagenesis step, the protocols will be identical to those described in the previous section.

Timeline: In year 1 we will optimize the method, building and sequencing the initial libraries for use in the ENCODE cell lines for testing purposes. In years 2 and beyond we will turn the effort to PDAC and CAD models.

Aim 3. Testing for enhancer necessity using CRISPR mutagenesis

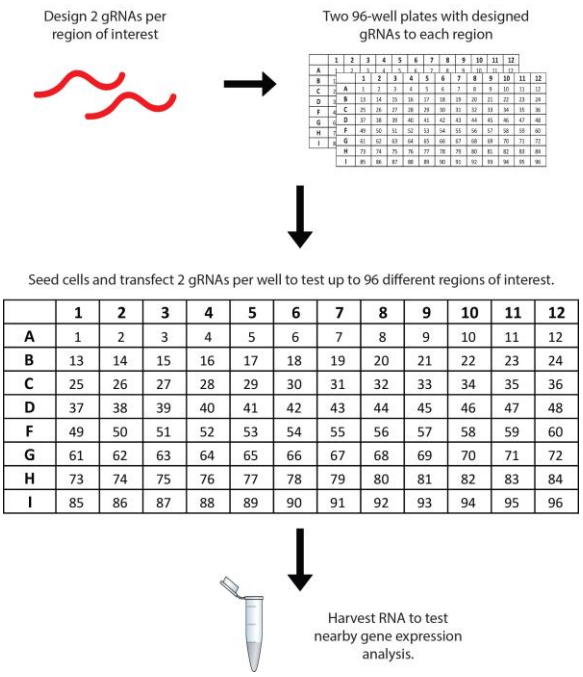
A major caveat of the STARR-seq method is that all assays are done on transfected plasmids that lack the contextual framework of the genomic loci from which they have been derived. The advantage of this method is that it allows us to systematically test thousands of genomic regions for regulatory potential; however, for many loci it can be only a modest indicator of the actual regulatory function within the native genomic context. Based on previous experience, we expect that data collected as described in Aim 2 will narrow the regions of interest from tens of thousands of candidate enhancer elements down to hundreds or thousands of partially validated enhancer elements. We will prioritize lists of functionally-validated enhancer elements based on Aim 1 and other computational methods developed in the Consortium. We will then test these regulatory elements using CRISPR mediated genome editing to determine which regulatory elements show differential function when mutated.

Traditional candidate enhancer validation assays have historically involved cloning candidate sequence of interest into a construct upstream of a reporter gene, in order to determine its regulatory activity by measuring the output of reporter gene expression. However, these methods do not take into account the endogenous chromatin environment the target enhancer resides in. Nor does it address which nearby genes the enhancer

region transcriptionally controls. To obtain a better understanding of the functional role candidate enhancers validated by STARR-seq in Aim 2 have on endogenous transcriptional regulation, we propose to functionally validate a subset of these candidate enhancer regions by endogenously mutating them using the CRISPR/Cas9 targeting system. By utilizing CRISPR technology, we are able to edit the genome using CRISPR and CRISPR-associated (Cas) genes that have been exploited to achieve site-specific DNA recognition and cleavage. (100). In this fashion, not only are we interrogating our target enhancers in their endogenous chromatin context, but we will also be able to obtain a clearer picture on which gene(s) the regulatory element may control.

Testing for enhancer necessity will be approached in two ways; 1) we will generate mutations in putative enhancers using a 96-well plate format and use a qRT-PCR (quantitative reverse transcriptase PCR) assay of nearby gene transcripts to generate quantitative transcriptional read outs (See Aim 3.1); and 2) we will use Drop-seq paired to CRISPR/Cas9 in order to get a high throughput, single cell resolution view of enhancer mutations (See Aim 3.2). This second variant has the potential to create an "all-by-all" matrix of enhancer-by-transcription unit effects. We will coordinate with other Centers who are taking similar approaches with complementary assays in ENCODE cell lines and other biological systems to ensure that we create a robust data set.

In addition to testing the necessity of our target enhancer regions discovered in Aim 2 by targeting them directly, we will also plan to knockdown/knockout a selected number of transcription factors that are observed to bind our target enhancer regions to see the downstream effect on nearby genes. TFs are widely known to exert their transcriptional regulatory potential via binding to enhancers. Therefore, by targeting TFs themselves, we will obtain similarly valuable insights on enhancer function, necessity, and the trans-effects of specific TFs. In order to determine which TFs we plan on interrogating, we will rely on the combination of computational analysis performed in Aim 1, alongside motif enrichment analysis we will perform on our putative enhancer regions from the various STARR-seq datasets generated in Aim 2. In all, we will plan to test 10-15 highly relevant TFs that are found to be enriched in our target regions for each biological model system we will examine (CAD, PDAC, a limited set of ENCODE cell lines for technique development and comparison to other ENCODE consortium data).



	1	2	3	4	5	6	7	8	9	10	11	12
A	1	2	3	4	5	6	7	8	9	10	11	12
B	13	14	15	16	17	18	19	20	21	22	23	24
C	25	26	27	28	29	30	31	32	33	34	35	36
D	37	38	39	40	41	42	43	44	45	46	47	48
F	49	50	51	52	53	54	55	56	57	58	59	60
G	61	62	63	64	65	66	67	68	69	70	71	72
H	73	74	75	76	77	78	79	80	81	82	83	84
I	85	86	87	88	89	90	91	92	93	94	95	96

Figure 11. Diagram representation of CRISPR qPCR enhancer validation workflow. We will apply 2gRNA's per region to ensure that we obtain efficient disruption to the region in a 96 well plate format. Cells will be harvested and regulated genes identified with qPCR.

3.1.1 High throughput enhancer validation: We propose to perform the aforementioned experiments using a CRISPR-qRT-PCR workflow (Figure 11). For targeting candidate enhancers and TFs, we plan to design 2 gRNAs for each candidate. For screening we will using a 96-well plate format (Figure 11), seeding cells at a density of 1.0e4 – 3.0e4 cells per well, depending on cell type (see Table 3). For ENCODE cell lines, we have already created stably expressed Cas9 (see Figure 12A), which will result in a higher efficiency and success rate of gRNA targeting. Furthermore, the gRNAs will be constructed as gBlocks (IDT). This approach provides us with many benefits, which include: A) it will allow us to bypass difficult co-transfection approaches of both Cas9 and the gRNA, B) it eliminates the need to clone all the gRNAs individually into a vector, C) eliminates lentiviral approaches for gRNA introduction and D) this workflow will allow us to functionally test up to 96 different conditions simultaneously on one plate. We plan to interrogate the expression levels of the closest or most likely 3-5 gene targets of each candidate enhancer (determined in Aim 1) via qRT-PCR. Transfection experiments will be conducted 24 hrs after initial cell seeding, using Lipofectamine 3000 transfection reagent, according to standard manufacturer's protocol. After transfection, cells will then be incubated for 72 hrs at 37°C before harvesting for RNA. High quality RNA will be extracted using a combination of the Trizol Reagent (Thermo Fisher) and Direct-Zol RNA Miniprep kit (Zymo Research). cDNA synthesis will be performed using SuperScript III First Strand Synthesis kit (Thermo Fisher). We will design qPCR primers for the nearest 3-5 genes of each target enhancer region, depending on genomic location and quantitate expression levels using RT-PCR on the

Roche LightCycler 480 instrument, using manufacturer's recommended reagents. The genes to test with each enhancer will be chosen through a combination of two methods, 1) genes that are predicted to be influenced by a given gene (as predicted in Aim 1) and 2) through proximity of gene to nearby enhancer. Using this method, we can test 96 gRNA pairs a month and their putative corresponding genes.

While testing on ENCODE cell lines should give us general insight on how candidate enhancer sufficiency as determined by STARR-seq relates to necessity as determined by CRISPR-Cas9 mutational targeting, it will not tell us whether the specific candidate enhancers identified in our models of CAD and PDAC are necessary in the context of the disease-relevant cell types. Thus a similar approach will be taken for the PDAC organoids and the human coronary artery smooth muscle cells. Unfortunately we will not have the advantage of stably transfected Cas9 in these models. However we will deliver the Cas9 and gRNAs at more modest throughput using the same transfection and lentiviral transduction approaches outlined in Aim 2 for these cell types. Rather per month, we expect to be able to process 20-30 enhancers and 10-15 transcription factors per year in PDAC and CAD models. Combining the human Starr-seq data procured in Aim 2 for each cell type with Fun-seq and LARVA annotation data (Aim 1), will enable us to pick out and target the most disease relevant enhancers to test using this method.

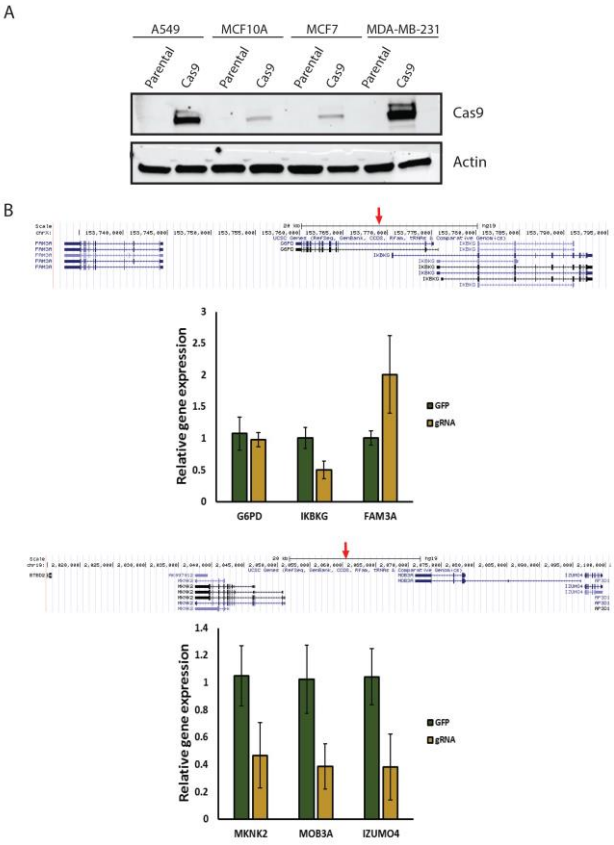


Figure 12. a) Western blot showing stable exogenous expression of Cas9 in multiple cell line backgrounds, including ENCODE tier 2 cell line MCF7 and A549. B) Two gRNAs (as indicated by the red arrow) were designed to candidate enhancer regions that overlap the active enhancers identified in STARR-seq as well as predicted functionally relevant genetic variation from FunSeq analysis. gRNAs were transfected into MCF7-Cas9 cells for 72 hours and RNA subsequently harvested for qPCR analysis. Nearby genes are observed to be down regulated as compared to control, indicating our enhancer validation pipeline is working.

3.1.2 Preliminary data for high throughput enhancer validation: The White lab has created stably expressing Cas9 cell lines in MCF-7, MDA-MB-231, MCF10A, and A549 cell lines (Figure 12A). We have begun testing enhancers using this method in ENCODE MCF7 cells (Figure 12B). We will continue this process with the Tier 1 and Tier 2 ENCODE cell lines (Table 3), as well as with our PDAC organoid and CAD smooth muscle cell models.

Figure 12B shows two candidate enhancer regions from MCF-7 cells that displayed ERα binding, estrogen responsiveness, Cap-STARR-seq activity, and mutations from breast cancer patients based on FunSeq analysis. Results show that by creating CRISPR-mediated mutations at these regions (indicated by the red arrows), nearby gene expression levels are affected (Fig. 12b). We observed that mutations in one candidate enhancer led to down regulation of one nearby gene and up-regulation of another. In the second candidate enhancer mutation, all three nearby genes that were assayed by qRT-PCR are affected. These results illustrate the importance of interrogating multiple nearby genes in their endogenous context instead of only experimenting with the target enhancer region exogenously (i.e., traditional luciferase assays). While these data were as gathered in a relatively low throughput manner, the approach is readily scalable.

Timeline: We will assay approximately 500 candidate enhancers in ENCODE cell lines in years 1 and 2. In each year of the grant we will assay 20-30 candidate enhancers in each biological model (PDAC organoids and CAD smooth muscle cells), as well as 10-15 transcription factors.

3.2.1.CRISPR-Drop-seq (CD-seq): Once targets are positively validated, we will use CRISPR technology to remove regulatory sites and measure the effect using a modified version of the Drop-seq method that has recently been developed in the White lab. In Drop-seq, single cells are encapsulated within nanoliter droplets and a uniquely barcoded is associated with the cell's RNA. Thousands of barcoded transcriptomes are then sequenced simultaneously, and single cell gene expression profiles are constructed. We have combined this method with transcription factor knock-down, and propose to combine it also with our CRISPR constructs to determine the downstream effects of the enhancer in the context of its native chromatin environment.

In CRISPR-Drop-seq (CD-seq), we will use modified beads that capture mRNA (through a poly-dT oligonucleotide) and the expressed gRNAs through a second oligonucleotide complementary to the common portion of each gRNA (Figure 13). Using this method, we will be able to see specific gRNA effects in single cells through use of RNA-seq, creating a high throughput validation of the functional effect of mutating candidate enhancer elements.

The overall workflow for CD-seq is similar to Drop-seq. A microfluidic device is used to encapsulate cells within nanoliter-sized droplets along with barcoded beads. The device will co-flow two aqueous solutions across an oil channel to form a water-in-oil emulsion. One flow contains a cell suspension; the other contains barcoded beads suspended in lysis buffer. The flows is precisely controlled using syringe pumps, such that laminar flow prevents mixing of the two aqueous solutions prior to droplet formation. A flow rate of ~4000 uL/hr will be used for the aqueous flows, and ~15,000 uL/hr will be used for the oil flow, resulting in the formation of droplets with a diameter of ~100 um. Droplet diameter will be measured for a given set of flow rates and cell/bead concentrations will be determined based on droplet size, with higher cell/bead concentrations used for smaller droplets and vice-versa. Overall, the number of droplets generated will be much larger than the number of beads or cells injected, so that a droplet will generally contain zero or one cell/bead. Immediately after droplet formation, the two aqueous flows will mix, resulting in cell lysis and release of mRNA, which will then captured by hybridization to the primers on the bead surface. Droplets generated by flowing in 1 mL each of cell and bead solutions are collected together. The droplets are broken by the addition of perflourooctanol that disrupts the oil-water interface, into a large volume of aqueous solution, which quenches further hybridization. The beads are then be collected by centrifugation, washed and reverse-transcribed. Next, the beads are treated with exonuclease I to remove any primers that did not capture an RNA molecule, washed, counted, aliquoted into PCR tubes and PCR amplified. The PCR reactions are purified and pooled, and the amplified cDNA quantified. Finally, the cDNA is fragmented and amplified using primers that allow amplification of only the 3' ends, then processed into RNA-seq libraries and sequenced.

CD-seq can be performed by mutating one candidate enhancer or one transcription factor at a time. Although our initial results are encouraging (see Preliminary results below), we will validate this method in year 1 by choosing 5-10 of the same candidate enhancers and transcription factors as we will assay in Aim 3.1. While we expect to get insights into cell heterogeneity and to address the question of whether results in cell populations are reflective of results in single cells, the most exciting prospect for the technique is its application to pooled gRNAs that get decoded along with the captured polyA RNA from each cell.

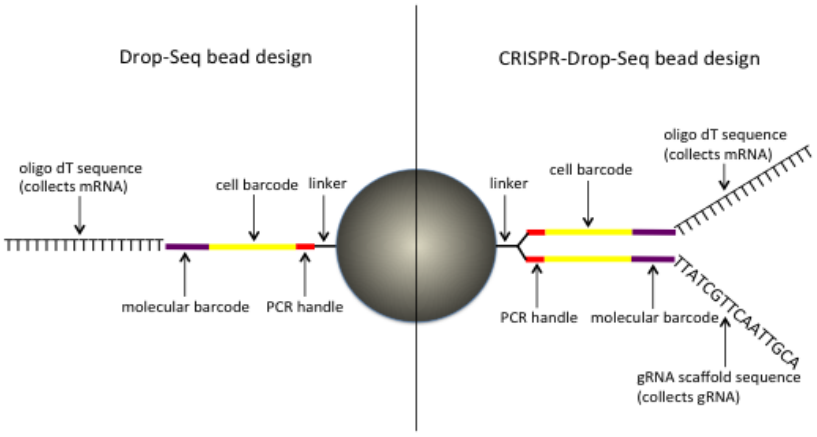


Figure 13. Bead comparison of Drop-Seq (left) and CRISPR-Drop-seq (right). The primary difference in these techniques is the beads. In CD-seq, the oligo coated beads have been enhanced to collect gDNA and mRNA.

Since Drop-seq allows us to individually sequence the transcriptomes of thousands of cells at a time, we plan to design a library of gRNAs that target our regions of interest based on the CRISPR validation assays and the computational analyses discussed in Aim 1. We will introduce this pooled library into our cells at a multiplicity of infection that ensures that only 1 (or zero) gRNA will enter each cell, and then we will use our CD-seq approach as described above. This will allow us to construct single cell expression profiles of each of the mutated (or normal) cells individually, and consequently, to determine the effect each gRNA mutation will have on transcriptome-wide RNA expression. We will run the CD-seq pipeline in a given cell line multiple times (to get 100,000s of cells) to ensure reproducibility of the data. Consequently, we will be able to identify and correlate target genes with their respective enhancer regions more robustly than what current methods such as single validations using qPCR or RNA-seq, would provide. More intriguingly, if the pooled gRNA version of CD-seq using the modified bead chemistry shown in Figure 13 (which we are obtaining through a collaboration with the oligo bead source company) is successful, it will open up the possibility of assaying thousands of enhancer mutations in a single experiment that can be run in a few hours by one investigator.

While we will rely on the high throughput candidate enhancer validation described in Aim 3.1 for the bulk of the work proposed in the Center, we feel that this will be well complemented by the CD-seq method, particularly due to its potential for developing into an ultra high throughput, low cost method for target element mutation that could be useful to other Centers in the consortium as well.

3.2.2.CRISPR-DROP-seq (CD-seq) preliminary results:

Figure 14 shows proof of concept of knocking down a transcription factor and our ability to measure its effects at a single transcriptome level. Here the transcription factor KLF4 is knocked down in MCF7 cells, followed by expression profiling using both conventional RNA-seq as well as Drop-Seq. (A) shows the correlation between fold changes of a subset of genes calculated from RNA-seq (y-axis) or Drop-seq (x-axis). (B) is a heatmap showing single cell expression levels of the same genes for 100 individual MCF7 knockdown and control cells. These results demonstrate that CD-seq can recapitulate traditional assays using pooled cells from culture.

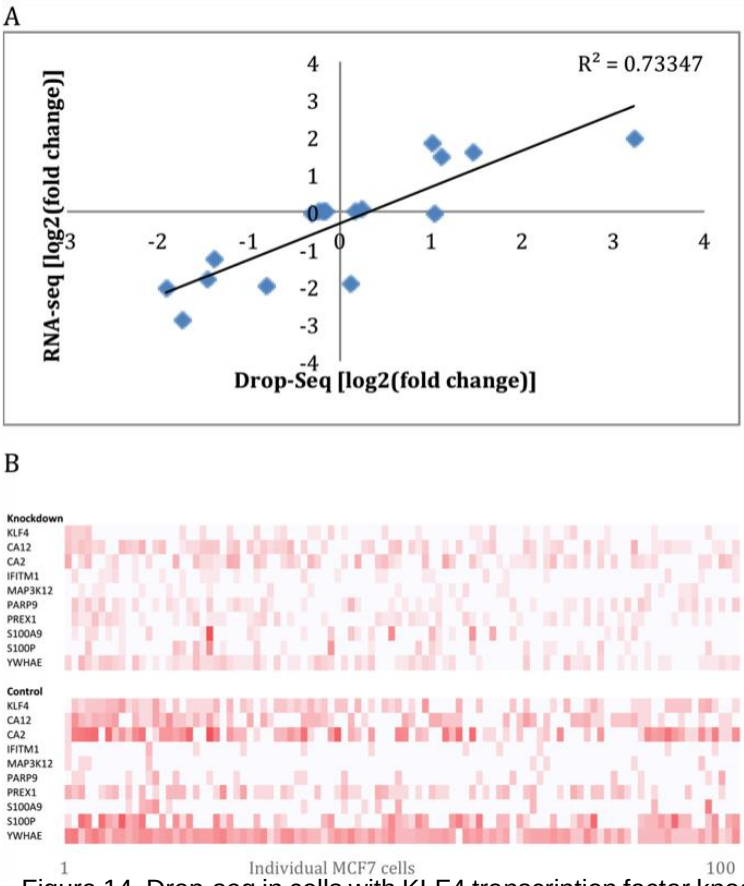


Figure 14. Drop-seq in cells with KLF4 transcription factor knock-down (A) shows the correlation between fold changes of a subset of genes calculated from RNA-seq (y-axis) or Drop-seq (x-axis). (B) is a heatmap showing single cell expression levels of the same genes for 100 individual MCF7 knockdown and control cells.

Timeline: We will assay 5-10 candidate enhancers and 5-10 transcription factors (TFs) in ENCODE cell lines in year 1 to optimize the assay. If results prove to be complementary on a "one at a time" basis, we will continue to assay 5-10 candidate enhancers and TFs per year, switching to our PDAC and CAD models in year 2 and future years. Also in year 1 and 2 we test the gRNA pool approach with the modified bead chemistry. If this proves fruitful we will increase our focus on this assay in year 3-4, and decrease the focus on the high throughput assays in Aim 3.1.

Aim 4: **Testing selected human enhancers *in vivo***

Enhancers are known to specify temporal and cell-specific patterns of gene expression, in developmental, physiological and pathophysiological contexts. Thus, to investigate the authenticity of enhancers identified in the previous Aims in an in vivo setting, it is essential to evaluate their functional activity in the context of such model situations. The primary goals of these studies are to employ in vivo transgenic mouse methodology for validation of enhancer regions identified in the previous Aims, to determine if identified enhancers overlap with known disease variation, and to investigate if risk allelic variation in these enhancers alters the cellular or disease contextual expression. Transgene reporter studies have been widely used to evaluate transcriptional elements for their developmental and cell-specific direction of gene expression, and we anticipate that bona fide enhancers identified through these studies will promote reporter transgene expression in the proposed transgenic models(53, 101). However, in only a few cases have transgenic reporter genes been used to model allelic single base-pair differences in TF binding sites identified through investigation of mechanisms of complex disease associations(56). Nonetheless, this latter type of study is not dissimilar from published studies where mutations in transcriptional elements have been employed to identify precise base-pair sequences that constitute the active transcriptional element(102). The primary approach to investigate activity of identified enhancers will be the use of embryonic development as the expression model, while for enhancers that are identified at disease variation, we will also employ adult mouse models to investigate pancreatic cell-specific expression, and adult vascular cell expression as well as disease cell-specific expression in a moderate throughput atherosclerosis mouse.

Despite the phenomenal progress in identifying genetic loci that harbor disease associated genes, for complex diseases such as cancers and cardiovascular disorders, there has been limited progress toward deciphering the genetic and molecular mechanisms by which causal variation regulates causal gene function. The majority of disease-associated variation has been found to reside outside of structural genes and presumed to regulate gene expression rather than encoded protein structure(81-83). Thus far, single nucleotide polymorphisms appear to be the type of variation that contribute the majority of risk(103), and in a small number of disease loci causal variants have been identified and their mechanism of effect studied. In these cases, the causal SNP has been shown to reside in enhancer regions and alter transcription factor binding and causal gene transcription(55-59, 104-107). Currently, greater than 65 CAD variants have been replicated and another 200 variants associated at an FDR <0.05(108), and pancreatic disease associated variation is currently being identified through genome wide association and whole genome sequencing of patient samples(109-113). Each of these disease variants are expected to reside in enhancer regions that mediate critical disease associated pathways.

4.1 Transgenic mouse developmental model: Embryogenesis is a highly stereotyped and complex cellular process that relies on numerous enhancer regions to activate expression of genes that are downstream of lineage determining signaling pathways. We thus anticipate that many of the enhancers identified through this work will promote cell-specific gene expression in the embryo. Interestingly, developmental and tissue-specific enhancer regions have been noted to harbor disease-associated causal variation, for CAD(47), type 2 diabetes mellitus(114) and various forms of cancer(56, 58).

These mouse models will be generated in the Nobrega lab, which has tested over 300 candidate enhancer sequences using mouse and zebrafish in vivo reporter assays(53, 56, 115, 116). This lab has employed an array of differential applications including (i) testing human candidate enhancer sequences in transient transgenic mice at various embryonic and post-natal stages; (ii) utilizing ZFN, TALEN and, lately, CRISPR/Cas9 technologies to engineer enhancer deletions or enhancer editing, altering specific nucleotides within target enhancers; (iii) utilizing Bacterial Artificial Chromosomes (BACs) to generate humanized transgenic mice for in vivo reporter assays of gene regulatory landscapes, and (iv) engineering enhancer deletions or modifications in human BACs. We propose to capitalize on these applications of in vivo enhancer assaying in mice and test selected regions (20 per year) to validate enhancers identified by mapping approaches in Aims 2 and 3.

For these studies, we will employ a plasmid based "transient transgenic" approach whereby foster mothers will be sacrificed and transgenic embryos resulting from microinjected oocytes harvested at an early (E10.5) and mid-gestational (E13.5) developmental periods. The wholemount Xgal staining approach will be used as well as standard microscopy of sectioned tissues Xgal stained as a wholemount or stained on slides after sectioning. For pancreas development, at E10.5 the focus will be on the developing ventral pancreatic buds, which derive from the ventral foregut endoderm, at E13.5 the focus will be on epithelial cells that undergo a differentiation process and commit to the major pancreatic lineages.

Evaluation of SMC enhancers will be conducted in collaboration with the Quertermous lab, which has extensive experience evaluating transgene reporter expression in the developing vasculature(117-120), and specifically the use of this model as an assay for transcriptional regulatory elements(118, 119). At E10.5, development of the embryonic vascular plexus in the yolk sac and development of major vascular structures will be investigated, and at E13.5 embryonic expression in organ microvascular beds will be evaluated. Reporter gene expression will be scored for temporal and cell-specific patterns of expression. We can combine assessment of transgene expression with evaluation of cell-specific expression as determined by co-staining with antibodies to lineage-specific markers.

4.2. Transgenic mouse atherosclerosis CAD model: One great promise of the work described in the preceding Aims for HCASMC is that enhancers will be identified that are perturbed by allelic variation that is causal for atherosclerotic coronary artery disease. SMC enhancers identified in Aims 2 and 3 that validate in the embryonic model and meet one of the following criteria will be further evaluated in a mouse model of atherosclerosis: (i) identified HCASMC enhancer region colocalizes with CAD causal variation as identified with ATAC-seq, histone modification, and in vitro functional studies, (ii) identified HCASMC enhancer region colocalizes in CAD locus at a mapped TCF21 binding site, and is thus likely part of the TCF21 directed CAD transcriptional network(95), (iii) enhancer function is specific for HCASMC grown under phenotypically modulated, disease-related, conditions. The gold standard for atherosclerosis research in the mouse is the hyperlipidemic model generated by deletion of the *ApoE* or the *LDLR* genes, and this lab has extensive experience employing these models to investigate the genetic role of human genes(121-124). However, for such studies the knockout alleles have to be combined with the reporter gene in the same mouse, requiring prolonged mating. For these studies, we will significantly increase the throughput for enhancer validation by generating adult transgenic animals and then initiating hypercholesterolemia by injecting adeno-associated virus (AAV) encoding the pathological human D374Y hypercholesterolemia gain-of-function mutant form of *PCSK9* (PCSK9(DY). AAV delivery to the liver of this mutant form of *PCSK9* has been shown to produce sustained hypercholesterolemia and atherosclerosis comparable to the *ApoE* and *LDLR* knockout models (125).

These studies will be performed in the Nobrega lab, in collaboration with Snyder and Quertermous labs at Stanford. The transgenic constructs will be human BAC clones that have been recombineered to represent the risk and protective alleles at the at the disease associated variant(s) under study. This approach has been employed in the Nobrega lab to investigate the rs6801957 as a candidate causal SNP associated with ventricular arrhythmias as identified by GWAS. This SNP is within SCN10A, which had been postulated to be a gene mediating functional abnormalities in cardiac conduction. Utilizing 3D genomic interactions, epigenetic marks, and human BAC engineering, we demonstrated that (i) rs6801957 is indeed within a cardiac conduction enhancer, (ii) that this SNP leads to strong allele-specific enhancer activity, (iii) this enhancer regulates SCN5A, not SCN10A expression, and (iv) a humanized transgenic mouse harboring a 200KB human BAC recapitulates the allele-specific expression properties of this enhancer, underscoring the power of mouse in vivo experimentation for validation of genomic predictions (Fig. 15). Further, with a similar approach the Quertermous lab has validated that the protective and risk CAD enhancers at the SMAD3 locus(49, 50) have different

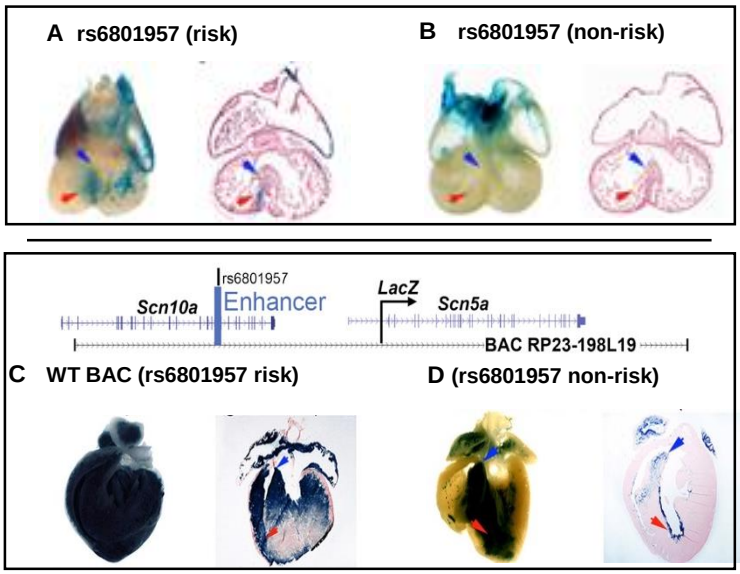


Fig. 15. Mouse reporter assays identify allele-specific enhancer properties of rs6801957, a GWAS SNP associated with cardiac conduction system defects. Transient reporter assays show that the risk allele of rs6801957 defines a strong ventricular conduction system of SCN5A (A), and that the non-risk allele lacks enhancer activity (B). (C) A human BAC containing the SCN5A locus and regulatory landscape, including rs6801957 is converted into an enhancer-trapping system, resulting in myocardial and conduction system expression mimicking SCN5A endogenous expression. (D) Conversion of rs6801957 risk allele into non-risk by recombineering results in loss of conduction system enhancer activity, demonstrating, *in vivo*, the allele-specific properties of this SNP in regulating SCN5A expression.

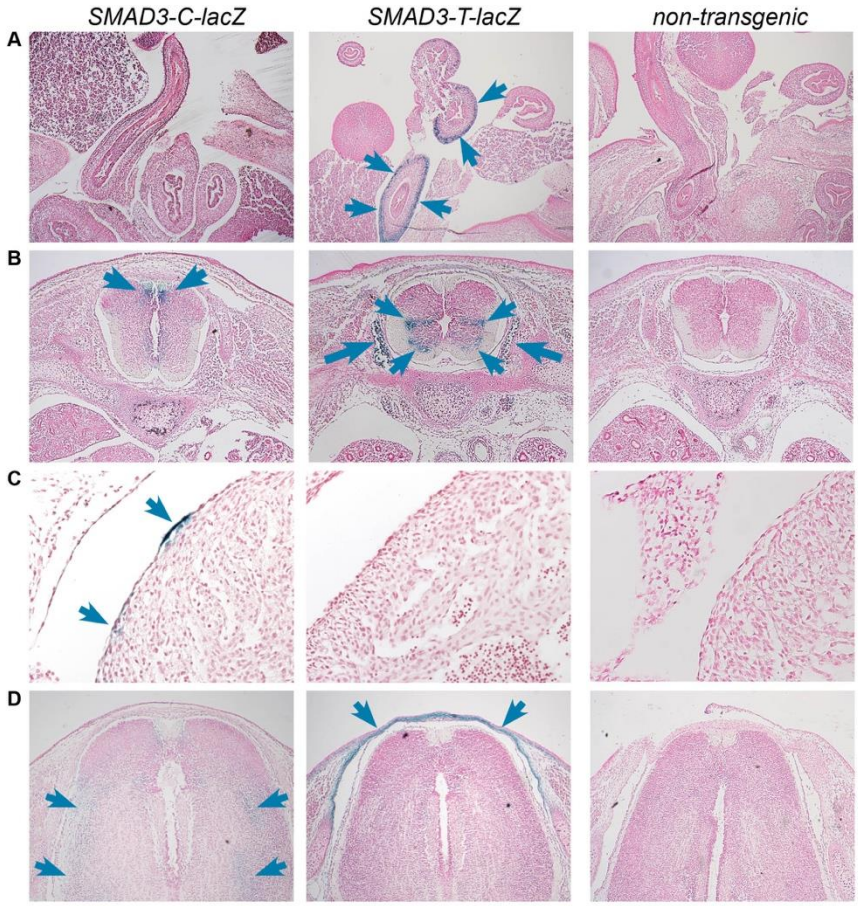


Fig. 16. In vivo assessment of risk C vs. protective T enhancer alleles at the SMAD3 locus for SNP rs17293632 in E15.5 mouse embryos. Blue arrows depict LacZ positive X-gal staining in (A) midgut loops for the T allele, (B) cells of neural tube roof plate for the C-allele, and cells of the ventricular zone of neural tube and dorsal root ganglion for the T-allele, (C) vascular cells of coronary vein in pericardium for the C-allele, and (D) neural cells of mantle layer for the C-allele and mesenchymal cells around medulla oblongata for the T-allele. Non-transgenic littermate control embryos show negative X-gal staining in matched regions (right).

embryonic expression patterns (Fig. 16). With this model, we will simultaneously utilize risk and protective human enhancer sequences to verify that the protective and risk alleles promote different cell-type specific and temporal patterns of expression. Transgenic mice will be developed as below, and will evaluate at least 4 enhancer allele pairs, with 5 lines generated with recombineered BAC constructs for each of the two alleles, as described below. We will evaluate and compare cell-specific expression of the lacZ reporter by Xgal assay done as wholemount and followed by tissue sectioning or performed on slides, as well as immunohistochemistry with the β-gal antibody to allow comparison to other cellular markers. As above, we can combine assessment of transgene expression with evaluation of cell-specific expression as determined by co-staining with antibodies to lineage-specific markers. Also, quantitation of reporter gene expression will be evaluated with qRT-PCR methods as we have done previously(122).

4.3. Transgenic mouse pancreatic enhancers

For pancreatic cancer-related enhancers, we will be primarily interested in the cell-specific expression in adult pancreatic tissues. These studies will be non-overlapping and complementary to the organoid model described in Aim 2, and they will provide *in vivo* correlation to the organoid culture model. We will test at least 4 enhancers per year, primarily focusing on those enhancers validated in Aims 2 and 3 and that also co-localize with pancreatic cancer mutational variation from whole genome sequencing (see Aim 1 Preliminary results). Both normal and mutated alleles will be modeled with BAC constructs, and at least 5 lines evaluated per construct, testing at least 4 enhancer pairs per year.

The pancreas as an organ is less well organized in mouse compared to humans, but pancreatic tissues can be readily identified in histology sections. The primary type of human pancreatic cancer is ductal adenocancarcinoma, and there is not a good model system for generating such cancers in mouse that would allow reasonable throughput. We will thus focus on identifying enhancers that drive reporter gene expression in ductal epithelial cells, and we will look for differences in cell-specific patterns of expression for the presumptive disease mutant alleles. With these studies we will specifically address whether STARR-seq and CRISPR-seq in organoids for the same enhancer predict relevant in vivo expression in disease-related cell types. Also, we anticipate validating pancreatic enhancers that are associated with disease mutational variation, testing to determine whether expression affected by normal and mutant alleles of enhancer candidates tested in Aims 2 and 3 will also show differences *in vivo* using the mouse transgenic model.

4.4. Transgenic model generation: Mouse models will be generated in the Nobrega lab, employing methodology developed to study enhancers, including those enhancers that regulate embryonic development and complex human disease phenotypes (53, 56, 101, 116).

We will employ a plasmid based transgenic approach for the developmental model. Each candidate enhancer sequence will be cloned into an hsp68LacZ vector and transgenic mice generated and assayed for lacZ expression in E10.5 and E13.5 embryos . For each transgenic embryonic reporter construct, we will test 20 enhancers per year, transfer 200 injected fertilized oocytes to 8-10 female mice, which will be used to assay for lacZ expression in transient transgenics. We require at least 3 consistent expression patterns in transient transgenic embryos per candidate sequence tested to call it an enhancer. We routinely get 10-15 transgenics per round of injections at the Nobrega lab. For detailed studies of disease disease-associated enhancer variation, we will employ a BAC-based transient transgenic assay. We will utilize a modified RED/ET homologous recombination in E. coli protocol to recombineer sequences within BACs, including LacZ reporter genes and/or modifications of enahncers within the BACs, such as enhancer deletions and allelic substitutions (Fig. 15, 16). For this approach, modified BAC DNA is extracted using Nucleobond AX Kit (Macherey-Nagel) and used for pronuclear injections of CD1 embryos in accordance with standard protocols approved by the University of Chicago.

Reporter gene expression patterns for both the developmental and atherosclerosis CAD models will be evaluated jointly by the Nobrega and Snyder / Quertermous labs at Stanford.

D. Bibliography

1. Benoist C, Chambon P. In vivo sequence requirements of the SV40 early promotor region. Nature. 1981;290(5804):304-10. PubMed PMID: 6259538.
2. Khoury G, Gruss P. Enhancer elements. Cell. 1983;33(2):313-4. PubMed PMID: 6305503.
3. Banerji J, Olson L, Schaffner W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell. 1983;33(3):729-40. PubMed PMID: 6409418.
4. Gillies SD, Morrison SL, Oi VT, Tonegawa S. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell. 1983;33(3):717-28. PubMed PMID: 6409417.
5. Queen C, Baltimore D. Immunoglobulin gene transcription is activated by downstream sequence elements. Cell. 1983;33(3):741-8. PubMed PMID: 6409419.
6. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013;14(4):288-95. Epub 2013/03/19. doi: nrg3458 [pii]
10.1038/nrg3458. PubMed PMID: 23503198.
7. Levine M. Transcriptional enhancers in animal development and evolution. Current biology : CB. 2010;20(17):R754-63. doi: 10.1016/j.cub.2010.06.070. PubMed PMID: 20833320.
8. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15(4):272-86. doi: 10.1038/nrg3682. PubMed PMID: 24614317.
9. Jeziorska DM, Jordan KW, Vance KW. A systems biology approach to understanding cis-regulatory module function. Seminars in cell & developmental biology. 2009;20(7):856-62. doi: 10.1016/j.semcdb.2009.07.007. PubMed PMID: 19660565.
10. Yuh CH, Bolouri H, Davidson EH. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science. 1998;279(5358):1896-902. PubMed PMID: 9506933.
11. mod EC, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A, Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova A, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, van Baren M, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SC, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B, Russell S, Cherbas P, Graveley BR, Lewis S, Micklem G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010;330(6012):1787-97. doi: 10.1126/science.1198374. PubMed PMID: 21177974; PMCID: PMC3192495.
12. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, Li Z, Ishii H, Spokony RF, Chen J, Hwang L, Cheng C, Auburn RP, Davis MB, Domanus M, Shah PK, Morrison CA, Zieba J, Suchy S, Senderowicz L, Victorsen A, Bild NA, Grundstad AJ, Hanley D, MacAlpine DM, Mannervik M, Venken K, Bellen H, White R, Gerstein M, Russell S, Grossman RL, Ren B, Posakony JW, Kellis M, White KP. A cis-regulatory map of the Drosophila genome. Nature. 2011;471(7339):527-31. doi: 10.1038/nature09990. PubMed PMID: 21430782; PMCID: PMC3179250.
13. Negre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, Feng X, Ahmad K, Russell S, White RA, Stein L, Henikoff S, Kellis M, White KP. A comprehensive map of insulator elements for the Drosophila genome. PLoS Genet. 2010;6(1):e1000814. doi: 10.1371/journal.pgen.1000814. PubMed PMID: 20084099; PMCID: PMC2797089.
14. Slattery M, Ma L, Spokony RF, Arthur RK, Kheradpour P, Kundaje A, Negre N, Crofts A, Ptashkin R, Zieba J, Ostapenko A, Suchy S, Victorsen A, Jameel N, Grundstad AJ, Gao W, Moran JR, Rehm EJ, Grossman RL, Kellis M, White KP. Diverse patterns of genomic targeting by transcriptional regulators in Drosophila melanogaster. Genome research. 2014;24(7):1224-35. doi: 10.1101/gr.168807.113. PubMed PMID: 24985916; PMCID: PMC4079976.
15. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS,

Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dose AC, Du J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller DM, 3rd, Muroyama A, Murray JI, Ooi SL, Pham H, Phippen T, Preston EA, Rajewsky N, Ratsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan KK, Zeller G, Zha Z, Zhong M, Zhou X, mod EC, Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330(6012):1775-87. doi: 10.1126/science.1196914. PubMed PMID: 21177976; PMCID: PMC3142569.

16. Niu W, Lu ZJ, Zhong M, Sarov M, Murray JI, Brdlik CM, Janette J, Chen C, Alves P, Preston E, Slightham C, Jiang L, Hyman AA, Kim SK, Waterston RH, Gerstein M, Snyder M, Reinke V. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. Genome research. 2011;21(2):245-54. doi: 10.1101/gr.114587.110. PubMed PMID: 21177963; PMCID: PMC3032928.

17. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. doi: 10.1038/nature11247. PubMed PMID: 22955616; PMCID: PMC3439153.

18. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science. 1998;281(5373):60-3. PubMed PMID: 9679020.

19. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, Lu Y, Wu Y, Jia Z, Li W, Zhang MQ, Ren B, Krainer AR, Maniatis T, Wu Q. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell. 2015;162(4):900-10. doi: 10.1016/j.cell.2015.07.038. PubMed PMID: 26276636; PMCID: PMC4642453.

20. Swamynathan SK, Piatigorsky J. Orientation-dependent influence of an intergenic enhancer on the promoter activity of the divergently transcribed mouse Shsp/alpha B-crystallin and Mkbp/HspB2 genes. J Biol Chem. 2002;277(51):49700-6. doi: 10.1074/jbc.M209700200. PubMed PMID: 12403771.

21. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012;488(7409):116-20. doi: 10.1038/nature11243. PubMed PMID: 22763441.

22. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. doi: 10.1038/nature11247. PubMed PMID: 22955616; PMCID: 3439153.

23. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006;444(7118):499-502. doi: 10.1038/nature05295. PubMed PMID: 17086198.

24. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;457(7231):854-8. doi: 10.1038/nature07730. PubMed PMID: 19212405; PMCID: 2745234.

25. Dickel DE, Visel A, Pennacchio LA. Functional anatomy of distant-acting mammalian enhancers. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2013;368(1620):20120359. doi: 10.1098/rstb.2012.0359. PubMed PMID: 23650633; PMCID: 3682724.

26. Morcillo P, Rosen C, Dorsett D. Genes regulating the remote wing margin enhancer in the Drosophila cut locus. Genetics. 1996;144(3):1143-54. PubMed PMID: 8913756; PMCID: 1207607.

27. Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, Myeroff L, Lutterbaugh J, Jarrar A, Kalady MF, Willis J, Moore JH, Tesar PJ, Laframboise T, Markowitz S, Lupien M, Scacheri PC. Epigenomic enhancer profiling defines a signature of colon cancer. Science. 2012;336(6082):736-9. doi: 10.1126/science.1217277. PubMed PMID: 22499810; PMCID: PMC3711120.

28. Fraser P. Transcriptional control thrown for a loop. Current opinion in genetics & development. 2006;16(5):490-5. doi: 10.1016/j.gde.2006.08.002. PubMed PMID: 16904310.

29. Vilar JM, Saiz L. DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise. Current opinion in genetics & development. 2005;15(2):136-44. doi: 10.1016/j.gde.2005.02.005. PubMed PMID: 15797196.
30. Song SH, Hou C, Dean A. A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. Molecular cell. 2007;28(5):810-22. doi: 10.1016/j.molcel.2007.09.025. PubMed PMID: 18082606; PMCID: 2195932.
31. Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, Dean A, Blobel GA. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell. 2012;149(6):1233-44. doi: 10.1016/j.cell.2012.03.051. PubMed PMID: 22682246; PMCID: 3372860.
32. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jorgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Muller F, Consortium F, Forrest AR, Carninci P, Rehli M, Sandelin A, Kawaji H, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jorgensen M, Dimont E, Arner E, Schmid C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple CA, Ishizu Y, Young RS, Francescatto M, Alam I, Albanese D, Altschuler GM, Arakawa T, Archer JA, Arner P, Babina M, Rennie S, Balwierz PJ, Beckhouse AG, Pradhan-Bhatt S, Blake JA, Blumenthal A, Bodega B, Bonetti A, Briggs J, Brombacher F, Burroughs AM, Califano A, Cannistraci CV, Carbajo D, Chen Y, Chierici M, Ciani Y, Clevers HC, Dalla E, Davis CA, Detmar M, Diehl AD, Dohi T, Drablos F, Edge AS, Edinger M, Ekwall K, Endoh M, Enomoto H, Fagiolini M, Fairbairn L, Fang H, Farach-Carson MC, Faulkner GJ, Favorov AV, Fisher ME, Frith MC, Fujita R, Fukuda S, Furlanello C, Furuno M, Furusawa J, Geijtenbeek TB, Gibson AP, Gingeras T, Goldowitz D, Gough J, Guhl S, Guler R, Gustincich S, Ha TJ, Hamaguchi M, Hara M, Harbers M, Harshbarger J, Hasegawa A, Hasegawa Y, Hashimoto T, Herlyn M, Hitchens KJ, Ho Sui SJ, Hofmann OM, Hoof I, Hori F, Huminiecki L, Iida K, Ikawa T, Jankovic BR, Jia H, Joshi A, Jurman G, Kaczkowski B, Kai C, Kaida K, Kaiho A, Kajiyama K, Kanamori-Katayama M, Kasianov AS, Kasukawa T, Katayama S, Kato S, Kawaguchi S, Kawamoto H, Kawamura YI, Kawashima T, Kempfle JS, Kenna TJ, Kere J, Khachigian LM, Kitamura T, Klinken SP, Knox AJ, Kojima M, Kojima S, Kondo N, Koseki H, Koyasu S, Krampitz S, Kubosaki A, Kwon AT, Laros JF, Lee W, Lennartsson A, Li K, Lilje B, Lipovich L, Mackay-sim A, Manabe R, Mar JC, Marchand B, Mathelier A, Mejhert N, Meynert A, Mizuno Y, de Lima Morais DA, Morikawa H, Morimoto M, Moro K, Motakis E, Motohashi H, Mummery CL, Murata M, Nagao-Sato S, Nakachi Y, Nakahara F, Nakamura T, Nakamura Y, Nakazato K, van Nimwegen E, Ninomiya N, Nishiyori H, Noma S, Nozaki T, Ogishima S, Ohkura N, Ohmiya H, Ohno H, Ohshima M, Okada-Hatakeyama M, Okazaki Y, Orlando V, Ovchinnikov DA, Pain A, Passier R, Patrikakis M, Persson H, Piazza S, Prendergast JG, Rackham OJ, Ramilowski JA, Rashid M, Ravasi T, Rizzu P, Roncador M, Roy S, Rye MB, Saijyo E, Sajantila A, Saka A, Sakaguchi S, Sakai M, Sato H, Satoh H, Savvi S, Saxena A, Schneider C, Schultes EA, Schulze-Tanzil GG, Schwegmann A, Sengstag T, Sheng G, Shimoji H, Shimoni Y, Shin JW, Simon C, Sugiyama D, Sugiyama T, Suzuki M, Suzuki N, Swoboda RK, t Hoen PA, Tagami M, Takahashi N, Takai J, Tanaka H, Tatsukawa H, Tatum Z, Thompson M, Toyoda H, Toyoda T, Valen E, van de Wetering M, van den Berg LM, Verardo R, Vijayan D, Vorontsov IE, Wasserman WW, Watanabe S, Wells CA, Winteringham LN, Wolvetang E, Wood EJ, Yamaguchi Y, Yamamoto M, Yoneda M, Yonekura Y, Yoshida S, Zabierowski SE, Zhang PG, Zhao X, Zucchelli S, Summers KM, Suzuki H, Daub CO, Kawai J, Heutink P, Hide W, Freeman TC, Lenhard B, Bajic VB, Taylor MS, Makeev VJ, Hume DA, Hayashizaki Y. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455-61. doi: 10.1038/nature12787. PubMed PMID: 24670763.
33. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ,

Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30. doi: 10.1038/nature14248. PubMed PMID: 25693563; PMCID: PMC4530010.
34. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75-82. doi: 10.1038/nature11232. PubMed PMID: 22955617; PMCID: 3721348.
35. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome research. 2007;17(6):877-85. doi: 10.1101/gr.5533506. PubMed PMID: 17179217; PMCID: 1891346.
36. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43-9. doi: 10.1038/nature09906. PubMed PMID: 21441907; PMCID: 3088773.
37. May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Afzal V, Simpson PC, Rubin EM, Black BL, Bristow J, Pennacchio LA, Visel A. Large-scale discovery of enhancers from human heart tissue. Nature genetics. 2012;44(1):89-93. doi: 10.1038/ng.1006. PubMed PMID: 22138689; PMCID: 3246570.
38. Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, Glass CK, Rosenfeld MG, Fu XD. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature. 2011;474(7351):390-4. doi: 10.1038/nature10006. PubMed PMID: 21572438; PMCID: 3117022.
39. Moorman C, Sun LV, Wang J, de Wit E, Talhout W, Ward LD, Greil F, Lu XJ, White KP, Bussemaker HJ, van Steensel B. Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(32):12027-32. doi: 10.1073/pnas.0605003103. PubMed PMID: 16880385; PMCID: PMC1567692.
40. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009;459(7243):108-12. doi: 10.1038/nature07829. PubMed PMID: 19295514; PMCID: 2910248.
41. Nord AS, Blow MJ, Attanasio C, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry M, Afzal V, Rubenstein JL, Rubin EM, Pennacchio LA, Visel A. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell. 2013;155(7):1521-31. doi: 10.1016/j.cell.2013.11.033. PubMed PMID: 24360275; PMCID: 3989111.
42. Sakabe NJ, Savic D, Nobrega MA. Transcriptional enhancers in development and disease. Genome biology. 2012;13(1):238. doi: 10.1186/gb-2012-13-1-238. PubMed PMID: 22269347; PMCID: 3334578.
43. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24(R1):R102-10. doi: 10.1093/hmg/ddv259. PubMed PMID: 26152199; PMCID: PMC4572001.
44. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190-5. Epub 2012/09/08. doi: 10.1126/science.1222794. PubMed PMID: 22955828; PMCID: 3771521.
45. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, Raychaudhuri S. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature genetics. 2013;45(2):124-30. doi: 10.1038/ng.2504. PubMed PMID: 23263488; PMCID: PMC3826950.

46. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012;482(7385):390-4. doi: 10.1038/nature10808. PubMed PMID: 22307276; PMCID: PMC3501342.
47. Miller CL, Anderson DR, Kundu RK, Raiesdana A, Nurnberg ST, Diaz R, Cheng K, Leeper NJ, Chen CH, Chang IS, Schadt EE, Hsiung CA, Assimes TL, Quertermous T. Disease-Related Growth Factor and Embryonic Signaling Pathways Modulate an Enhancer of TCF21 Expression at the 6q23.2 Coronary Heart Disease Locus. PLoS Genet. 2013;9(7):e1003652. Epub 2013/07/23. doi: 10.1371/journal.pgen.1003652. PubMed PMID: 23874238; PMCID: 3715442.
48. Miller CL, Haas U, Diaz R, Leeper NJ, Kundu RK, Patlolla B, Assimes TL, Kaiser FJ, Perisic L, Hedin U, Maegdefessel L, Schunkert H, Erdmann J, Quertermous T, Sczakiel G. Coronary Heart Disease-Associated Variation in TCF21 Disrupts a miR-224 Binding Site and miRNA-Mediated Regulation. PLoS Genet. 2014;10(3):e1004263. Epub 2014/03/29. doi: 10.1371/journal.pgen.1004263. PubMed PMID: 24676100.
49. Miller CL, Pjanic M, Wang T, Nguyen T, Cohain A, Perisic L, Hedin U, Betsholtz C, Ruusalepp A, Franzen O, Assimes TL, Montgomery SB, Schadt EE, Bjorkegren JLM, Quertermous T. Integrative fine-mapping of regulatory variants and mechanisms at coronary heart disease loci. in review. 2015.
50. Turner AW, Martinuk A, Silva A, Lau P, Nikpay M, Eriksson P, Folkersen L, Perisic L, Hedin U, Soubeyrand S, McPherson R. Functional Analysis of a Novel Genome-Wide Association Study Signal in SMAD3 That Confers Protection From Coronary Artery Disease. Arteriosclerosis, thrombosis, and vascular biology. 2016. doi: 10.1161/ATVBAHA.116.307294. PubMed PMID: 26966274.
51. Turner AW, Nikpay M, Silva A, Lau P, Martinuk A, Linseman TA, Soubeyrand S, McPherson R. Functional interaction between COL4A1/COL4A2 and SMAD3 risk loci for coronary artery disease. Atherosclerosis. 2015;242(2):543-52. doi: 10.1016/j.atherosclerosis.2015.08.008. PubMed PMID: 26310581.
52. Beaudoin M, Gupta RM, Won HH, Lo KS, Do R, Henderson CA, Lavoie-St-Amour C, Langlois S, Rivas D, Lehoux S, Kathiresan S, Tardif JC, Musunuru K, Lettre G. Myocardial Infarction-Associated SNP at 6p24 Interferes With MEF2 Binding and Associates With PHACTR1 Expression Levels in Human Coronary Arteries. Arteriosclerosis, thrombosis, and vascular biology. 2015;35(6):1472-9. doi: 10.1161/ATVBAHA.115.305534. PubMed PMID: 25838425; PMCID: PMC4441556.
53. Smemo S, Campos LC, Moskowitz IP, Krieger JE, Pereira AC, Nobrega MA. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. Hum Mol Genet. 2012;21(14):3255-63. doi: 10.1093/hmg/dds165. PubMed PMID: 22543974; PMCID: PMC3384386.
54. Savic D, Park SY, Bailey KA, Bell GI, Nobrega MA. In vitro scan for enhancers at the TCF7L2 locus. Diabetologia. 2013;56(1):121-5. doi: 10.1007/s00125-012-2730-y. PubMed PMID: 23011354; PMCID: PMC3525810.
55. Savic D, Ye H, Aneas I, Park SY, Bell GI, Nobrega MA. Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. Genome research. 2011;21(9):1417-25. Epub 2011/06/16. doi: 10.1101/gr.123745.111. PubMed PMID: 21673050; PMCID: 3166827.
56. Wasserman NF, Aneas I, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. Genome research. 2010;20(9):1191-7. Epub 2010/07/16. doi: 10.1101/gr.105361.110. PubMed PMID: 20627891; PMCID: 2928497.
57. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M, Yao K, Kehoe SM, Lenz HJ, Haiman CA, Yan C, Henderson BE, Frenkel B, Barretina J, Bass A, Tabernero J, Baselga J, Regan MM, Manak JR, Shivdasani R, Coetzee GA, Freedman ML. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature genetics. 2009;41(8):882-4. Epub 2009/06/30. doi: ng.403 [pii]
10.1038/ng.403. PubMed PMID: 19561607.
58. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Bjorklund M, Wei G, Yan J, Niittymaki I, Mecklin JP, Jarvinen H, Ristimaki A, Di-Bernardo M, East P, Carvajal-Carmona L, Houlston RS, Tomlinson I, Palin K, Ukkonen E, Karhu A, Taipale J, Aaltonen LA. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nature genetics. 2009;41(8):885-90. Epub 2009/06/30. doi: ng.406 [pii]
10.1038/ng.406. PubMed PMID: 19561604.
59. Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, Liao H, Stephens RM, Harris TJ, Munroe DJ, Wu X. Long-range enhancers on 8q24 regulate c-Myc. Proceedings of the National Academy of Sciences

of the United States of America. 2010;107(7):3001-5. Epub 2010/02/06. doi: 10.1073/pnas.0906067107. PubMed PMID: 20133699; PMCID: 2840341.
60. Sazonova O, Zhao Y, Nurnberg S, Miller C, Pjanic M, Castano VG, Kim JB, Salfati EL, Kundaje AB, Bejerano G, Assimes T, Yang X, Quertermous T. Characterization of TCF21 Downstream Target Regions Identifies a Transcriptional Network Linking Multiple Independent Coronary Artery Disease Loci. PLoS Genet. 2015;11(5):e1005202. doi: 10.1371/journal.pgen.1005202. PubMed PMID: 26020271; PMCID: PMC4447360.
61. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. Science. 2013;339(6122):957-9. doi: 10.1126/science.1229259. PubMed PMID: 23348506; PMCID: PMC4423787.
62. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gumus ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GR, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, Genomes Project C, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C, Gerstein M. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013;342(6154):1235587. doi: 10.1126/science.1235587. PubMed PMID: 24092746; PMCID: PMC3947637.
63. Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, Etchin J, Lawton L, Sallan SE, Silverman LB, Loh ML, Hunger SP, Sanda T, Young RA, Look AT. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science. 2014;346(6215):1373-7. doi: 10.1126/science.1259037. PubMed PMID: 25394790; PMCID: PMC4720521.
64. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nature genetics. 2014;46(11):1160-5. doi: 10.1038/ng.3101. PubMed PMID: 25261935; PMCID: PMC4217527.
65. Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nature genetics. 2014;46(12):1258-63. doi: 10.1038/ng.3141. PubMed PMID: 25383969.
66. Melton C, Reuter JA, Spacek DV, Snyder M. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nature genetics. 2015;47(7):710-6. doi: 10.1038/ng.3332. PubMed PMID: 26053494; PMCID: PMC4485503.
67. Puente XS, Bea S, Valdes-Mas R, Villamor N, Gutierrez-Abril J, Martin-Subero JI, Munar M, Rubio-Perez C, Jares P, Aymerich M, Baumann T, Beekman R, Belver L, Carrio A, Castellano G, Clot G, Colado E, Colomer D, Costa D, Delgado J, Enjuanes A, Estivill X, Ferrando AA, Gelpi JL, Gonzalez B, Gonzalez S, Gonzalez M, Gut M, Hernandez-Rivas JM, Lopez-Guerra M, Martin-Garcia D, Navarro A, Nicolas P, Orozco M, Payer AR, Pinyol M, Pisano DG, Puente DA, Queiros AC, Quesada V, Romeo-Casabona CM, Royo C, Royo R, Rozman M, Russinol N, Salaverria I, Stamatopoulos K, Stunnenberg HG, Tamborero D, Terol MJ, Valencia A, Lopez-Bigas N, Torrents D, Gut I, Lopez-Guillermo A, Lopez-Otin C, Campo E. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature. 2015;526(7574):519-24. doi: 10.1038/nature14666. PubMed PMID: 26200345.
68. Kron KJ, Bailey SD, Lupien M. Enhancer alterations in cancer: a source for a cell identity crisis. Genome Med. 2014;6(9):77. doi: 10.1186/s13073-014-0077-3. PubMed PMID: 25473436; PMCID: PMC4254433.
69. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009;27(1):66-75. doi: 10.1038/nbt.1518. PubMed PMID: 19122651; PMCID: PMC2924752.
70. Harmanci A, Rozowsky J, Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. Genome biology. 2014;15(10):474. doi: 10.1186/s13059-014-0474-3. PubMed PMID: 25292436; PMCID: PMC4234855.
71. Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput Biol. 2013;9(3):e1002968. doi: 10.1371/journal.pcbi.1002968. PubMed PMID: 23526891; PMCID: PMC3597546.
72. Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn KA, Minoda A, Tolstorukov MY, Appert A, Parker SC, Gu T, Kundaje A, Riddle NC, Bishop E, Egelhofer TA, Hu SS, Alekseyenko AA, Rechtsteiner A, Asker D, Belsky JA, Bowman SK, Chen QB, Chen RA, Day DS, Dong Y, Dose AC, Duan X, Epstein CB, Ercan

S, Feingold EA, Ferrari F, Garrigues JM, Gehlenborg N, Good PJ, Haseley P, He D, Herrmann M, Hoffman MM, Jeffers TE, Kharchenko PV, Kolasinska-Zwierz P, Kotwaliwale CV, Kumar N, Langley SA, Larschan EN, Latorre I, Libbrecht MW, Lin X, Park R, Pazin MJ, Pham HN, Plachetka A, Qin B, Schwartz YB, Shoresh N, Stempor P, Vielle A, Wang C, Whittle CM, Xue H, Kingston RE, Kim JH, Bernstein BE, Dernburg AF, Pirrotta V, Kuroda MI, Noble WS, Tullius TD, Kellis M, MacAlpine DM, Strome S, Elgin SC, Liu XS, Lieb JD, Ahringer J, Karpen GH, Park PJ. Comparative analysis of metazoan chromatin organization. Nature. 2014;512(7515):449-52. doi: 10.1038/nature13415. PubMed PMID: 25164756; PMCID: PMC4227084.

73. Yip KY, Alexander RP, Yan KK, Gerstein M. Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. PLoS One. 2010;5(1):e8121. doi: 10.1371/journal.pone.0008121. PubMed PMID: 20126643; PMCID: PMC2811182.

74. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome biology. 2012;13(9):R48. doi: 10.1186/gb-2012-13-9-r48. PubMed PMID: 22950945; PMCID: PMC3491392.

75. Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res. 2011;39(16):7058-76. doi: 10.1093/nar/gkr342. PubMed PMID: 21596777; PMCID: PMC3167619.

76. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339(6123):1074-7. Epub 2013/01/19. doi: science.1232542 [pii]
10.1126/science.1232542. PubMed PMID: 23328393.

77. Koo BK, Stange DE, Sato T, Karthaus W, Farin HF, Huch M, van Es JH, Clevers H. Controlled gene expression in primary Lgr5 organoid cultures. Nat Methods. 2012;9(1):81-3. doi: 10.1038/nmeth.1802. PubMed PMID: 22138822.

78. Andersson-Rolf A, Fink J, Mustata RC, Koo BK. A video protocol of retroviral infection in primary intestinal organoid culture. J Vis Exp. 2014(90):e51765. doi: 10.3791/51765. PubMed PMID: 25146755; PMCID: PMC4758749.

79. Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, Miller DK, Christ AN, Bruxner TJ, Quinn MC, Nourse C, Murtaugh LC, Harliwong I, Idrisoglu S, Manning S, Nourbakhsh E, Wani S, Fink L, Holmes O, Chin V, Anderson MJ, Kazakoff S, Leonard C, Newell F, Waddell N, Wood S, Xu Q, Wilson PJ, Cloonan N, Kassahn KS, Taylor D, Quek K, Robertson A, Pantano L, Mincarelli L, Sanchez LN, Evers L, Wu J, Pinese M, Cowley MJ, Jones MD, Colvin EK, Nagrial AM, Humphrey ES, Chantrill LA, Mawson A, Humphris J, Chou A, Pajic M, Scarlett CJ, Pinho AV, Giry-Laterriere M, Rooman I, Samra JS, Kench JG, Lovell JA, Merrett ND, Toon CW, Epari K, Nguyen NQ, Barbour A, Zeps N, Moran-Jones K, Jamieson NB, Graham JS, Duthie F, Oien K, Hair J, Grutzmann R, Maitra A, Iacobuzio-Donahue CA, Wolfgang CL, Morgan RA, Lawlor RT, Corbo V, Bassi C, Rusev B, Capelli P, Salvia R, Tortora G, Mukhopadhyay D, Petersen GM, Australian Pancreatic Cancer Genome I, Munzy DM, Fisher WE, Karim SA, Eshleman JR, Hruban RH, Pilarsky C, Morton JP, Sansom OJ, Scarpa A, Musgrove EA, Bailey UM, Hofmann O, Sutherland RL, Wheeler DA, Gill AJ, Gibbs RA, Pearson JV, Waddell N, Biankin AV, Grimmond SM. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016;531(7592):47-52. doi: 10.1038/nature16965. PubMed PMID: 26909576.

80. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013;93(5):779-97. doi: 10.1016/j.ajhg.2013.10.012. PubMed PMID: 24210251; PMCID: PMC3824120.

81. Gaffney DJ. Global properties and functional complexity of human gene regulatory variation. PLoS Genet. 2013;9(5):e1003501. Epub 2013/06/06. doi: 10.1371/journal.pgen.1003501. PubMed PMID: 23737752; PMCID: 3667745.

82. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome research. 2012;22(9):1748-59. Epub 2012/09/08. doi: 10.1101/gr.136127.111. PubMed PMID: 22955986; PMCID: 3431491.

83. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(23):9362-7. Epub 2009/05/29. doi: 10.1073/pnas.0903103106. PubMed PMID: 19474294; PMCID: 2687147.

84. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA, Hatan M, Carrasco-Alfonso MJ, Mayer D, Luckey CJ, Patsopoulos NA, De Jager PL, Kuchroo VK, Epstein CB, Daly MJ, Hafler DA, Bernstein BE. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2014. doi: 10.1038/nature13835. PubMed PMID: 25363779.
85. Braenne I, Civelek M, Vilne B, Di Narzo A, Johnson AD, Zhao Y, Reiz B, Codoni V, Webb TR, Foroughi Asl H, Hamby SE, Zeng L, Tregouet DA, Hao K, Topol EJ, Schadt EE, Yang X, Samani NJ, Bjorkegren JL, Erdmann J, Schunkert H, Lusis AJ, Leducq Consortium CADGd. Prediction of Causal Candidate Genes in Coronary Artery Disease Loci. Arteriosclerosis, thrombosis, and vascular biology. 2015;35(10):2207-17. doi: 10.1161/ATVBAHA.115.306108. PubMed PMID: 26293461; PMCID: PMC4583353.
86. Gomez D, Owens GK. Smooth muscle cell phenotypic switching in atherosclerosis. Cardiovascular research. 2012;95(2):156-64. Epub 2012/03/13. doi: 10.1093/cvr/cvs115. PubMed PMID: 22406749; PMCID: 3388816.
87. Shah PK. Mechanisms of plaque vulnerability and rupture. Journal of the American College of Cardiology. 2003;41(4 Suppl S):15S-22S. PubMed PMID: 12644336.
88. Sakakura K, Nakano M, Otsuka F, Ladich E, Kolodgie FD, Virmani R. Pathophysiology of atherosclerosis plaque progression. Heart, lung & circulation. 2013;22(6):399-411. doi: 10.1016/j.hlc.2013.03.001. PubMed PMID: 23541627.
89. Falk E, Nakano M, Bentzon JF, Finn AV, Virmani R. Update on acute coronary syndromes: the pathologists' view. European heart journal. 2013;34(10):719-28. doi: 10.1093/eurheartj/ehs411. PubMed PMID: 23242196.
90. Clarke M, Bennett M. The emerging role of vascular smooth muscle cell apoptosis in atherosclerosis and plaque stability. American journal of nephrology. 2006;26(6):531-5. doi: 10.1159/000097815. PubMed PMID: 17159340.
91. Shankman LS, Gomez D, Cherepanova OA, Salmon M, Alencar GF, Haskins RM, Swiatlowska P, Newman AA, Greene ES, Straub AC, Isakson B, Randolph GJ, Owens GK. KLF4-dependent phenotypic modulation of smooth muscle cells has a key role in atherosclerotic plaque pathogenesis. Nature medicine. 2015;21(6):628-37. doi: 10.1038/nm.3866. PubMed PMID: 25985364.
92. Nurnberg ST, Cheng K, Raiesdana A, Kundu R, Miller CL, Kim JB, Arora K, Carcamo-Orive I, Xiong Y, Tellakula N, Nanda V, Murthy N, Boisvert WA, Hedin U, Perisic L, Aldi S, Maegdefessel L, Pjanic M, Owens GK, Tallquist MD, Quertermous T. Coronary artery disease associated transcription factor TCF21 regulates smooth muscle precursor cells that contribute to the fibrous cap. PLoS Genet. 2015;11(5); PMCID: in process.
93. Leeper NJ, Raiesdana A, Kojima Y, Kundu RK, Cheng H, Maegdefessel L, Toh R, Ahn GO, Ali ZA, Anderson DR, Miller CL, Roberts SC, Spin JM, de Almeida PE, Wu JC, Xu B, Cheng K, Quertermous M, Kundu S, Kortekaas KE, Berzin E, Downing KP, Dalman RL, Tsao PS, Schadt EE, Owens GK, Quertermous T. Loss of CDKN2B Promotes p53-Dependent Smooth Muscle Cell Apoptosis and Aneurysm Formation. Arteriosclerosis, thrombosis, and vascular biology. 2013;33(1):e1-e10. Epub 2012/11/20. doi: 10.1161/ATVBAHA.112.300399. PubMed PMID: 23162013.
94. Kojima Y, Downing K, Kundu R, Miller C, Dewey F, Lancero H, Raaz U, Perisic L, Hedin U, Schadt E, Maegdefessel L, Quertermous T, Leeper NJ. Cyclin-dependent kinase inhibitor 2B regulates efferocytosis and atherosclerosis. The Journal of clinical investigation. 2014;124(3):1083-97. Epub 2014/02/18. doi: 10.1172/JCI70391. PubMed PMID: 24531546; PMCID: 3938254.
95. Sazonova O, Zhao Y, Nurnberg S, Miller C, Pjanic M, Castano VG, Kim JB, Salfati EL, Kundaje AB, Bejerano G, Assimes TL, Yang X, Quertermous T. Characterization of TCF21 downstream target regions identifies a transcriptional network linking multiple independent coronary artery disease loci. PLoS Genet. 2015;11(5); PMCID: in process.
96. Owens GK, Kumar MS, Wamhoff BR. Molecular regulation of vascular smooth muscle cell differentiation in development and disease. Physiol Rev. 2004;84(3):767-801. Epub 2004/07/23. doi: 10.1152/physrev.00041.2003
84/3/767 [pii]. PubMed PMID: 15269336.
97. Cheung C, Bernardo AS, Pedersen RA, Sinha S. Directed differentiation of embryonic origin-specific vascular smooth muscle subtypes from human pluripotent stem cells. Nat Protoc. 2014;9(4):929-38. doi: 10.1038/nprot.2014.059. PubMed PMID: 24675733.
98. Patsch C, Challet-Meylan L, Thoma EC, Urich E, Heckel T, O'Sullivan JF, Grainger SJ, Kapp FG, Sun L, Christensen K, Xia Y, Florido MH, He W, Pan W, Prummer M, Warren CR, Jakob-Roetne R, Certa U,

Jagasia R, Freskgard PO, Adatto I, Kling D, Huang P, Zon LI, Chaikof EL, Gerszten RE, Graf M, Iacone R, Cowan CA. Generation of vascular endothelial and smooth muscle cells from human pluripotent stem cells. Nat Cell Biol. 2015;17(8):994-1003. doi: 10.1038/ncb3205. PubMed PMID: 26214132; PMCID: PMC4566857.
99. Wei X, Das J, Fragoza R, Liang J, Bastos de Oliveira FM, Lee HR, Wang X, Mort M, Stenson PD, Cooper DN, Lipkin SM, Smolka MB, Yu H. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet. 2014;10(12):e1004819. doi: 10.1371/journal.pgen.1004819. PubMed PMID: 25502805; PMCID: PMC4263371.
100. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-guided human genome engineering via Cas9. Science. 2013;339(6121):823-6. doi: 10.1126/science.1232033. PubMed PMID: 23287722; PMCID: PMC3712628.
101. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, Lee JH, Puviindran V, Tam D, Shen M, Son JE, Vakili NA, Sung HK, Naranjo S, Acemel RD, Manzanares M, Nagy A, Cox NJ, Hui CC, Gomez-Skarmeta JL, Nobrega MA. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014;507(7492):371-5. doi: 10.1038/nature13138. PubMed PMID: 24646999; PMCID: PMC4113484.
102. Iwahori A, Fraidenraich D, Basilico C. A conserved enhancer element that drives FGF4 gene expression in the embryonic myotomes is synergistically activated by GATA and bHLH proteins. Developmental biology. 2004;270(2):525-37. Epub 2004/06/09. doi: 10.1016/j.ydbio.2004.03.012. PubMed PMID: 15183731.
103. Myocardial Infarction Genetics C, Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, Erdmann J, Reilly MP, Rader DJ, Morgan T, Spertus JA, Stoll M, Girelli D, McKeown PP, Patterson CC, Siscovick DS, O'Donnell CJ, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Melander O, Altshuler D, Ardissino D, Merlini PA, Berzuini C, Bernardinelli L, Peyvandi F, Tubaro M, Celli P, Ferrario M, Fetiveau R, Marziliano N, Casari G, Galli M, Ribichini F, Rossi M, Bernardi F, Zonzin P, Piazza A, Mannucci PM, Schwartz SM, Siscovick DS, Yee J, Friedlander Y, Elosua R, Marrugat J, Lucas G, Subirana I, Sala J, Ramos R, Kathiresan S, Meigs JB, Williams G, Nathan DM, MacRae CA, O'Donnell CJ, Salomaa V, Havulinna AS, Peltonen L, Melander O, Berglund G, Voight BF, Kathiresan S, Hirschhorn JN, Asselta R, Duga S, Spreafico M, Musunuru K, Daly MJ, Purcell S, Voight BF, Purcell S, Nemesh J, Korn JM, McCarroll SA, Schwartz SM, Yee J, Kathiresan S, Lucas G, Subirana I, Elosua R, Surti A, Guiducci C, Gianniny L, Mirel D, Parkin M, Burtt N, Gabriel SB, Samani NJ, Thompson JR, Braund PS, Wright BJ, Balmforth AJ, Ball SG, Hall A, Wellcome Trust Case Control C, Schunkert H, Erdmann J, Linsel-Nitschke P, Lieb W, Ziegler A, Konig I, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Schunkert H, Samani NJ, Erdmann J, Ouwehand W, Hengstenberg C, Deloukas P, Scholz M, Cambien F, Reilly MP, Li M, Chen Z, Wilensky R, Matthai W, Qasim A, Hakonarson HH, Devaney J, Burnett MS, Pichard AD, Kent KM, Satler L, Lindsay JM, Waksman R, Knouff CW, Waterworth DM, Walker MC, Mooser V, Epstein SE, Rader DJ, Scheffold T, Berger K, Stoll M, Huge A, Girelli D, Martinelli N, Olivieri O, Corrocher R, Morgan T, Spertus JA, McKeown P, Patterson CC, Schunkert H, Erdmann E, Linsel-Nitschke P, Lieb W, Ziegler A, Konig IR, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Holm H, Thorleifsson G, Thorsteinsdottir U, Stefansson K, Engert JC, Do R, Xie C, Anand S, Kathiresan S, Ardissino D, Mannucci PM, Siscovick D, O'Donnell CJ, Samani NJ, Melander O, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Altshuler D. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nature genetics. 2009;41(3):334-41. doi: 10.1038/ng.327. PubMed PMID: 19198609; PMCID: PMC2681011.
104. Stitzel ML, Sethupathy P, Pearson DS, Chines PS, Song L, Erdos MR, Welch R, Parker SC, Boyle AP, Scott LJ, Margulies EH, Boehnke M, Furey TS, Crawford GE, Collins FS. Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. Cell metabolism. 2010;12(5):443-55. Epub 2010/11/03. doi: 10.1016/j.cmet.2010.09.012. PubMed PMID: 21035756; PMCID: 3026436.
105. Pittman AM, Naranjo S, Webb E, Broderick P, Lips EH, van Wezel T, Morreau H, Sullivan K, Fielding S, Twiss P, Vijayakrishnan J, Casares F, Qureshi M, Gomez-Skarmeta JL, Houlston RS. The colorectal cancer risk at 18q21 is caused by a novel variant altering SMAD7 expression. Genome research. 2009;19(6):987-93. Epub 2009/04/28. doi: 10.1101/gr.092668.109. PubMed PMID: 19395656; PMCID: 2694486.
106. Pittman AM, Naranjo S, Jalava SE, Twiss P, Ma Y, Olver B, Lloyd A, Vijayakrishnan J, Qureshi M, Broderick P, van Wezel T, Morreau H, Tuupanen S, Aaltonen LA, Alonso ME, Manzanares M, Gavilan A,

Visakorpi T, Gomez-Skarmeta JL, Houlston RS. Allelic variation at the 8q23.3 colorectal cancer risk locus functions as a cis-acting regulator of EIF3H. PLoS Genet. 2010;6(9). Epub 2010/09/24. doi: 10.1371/journal.pgen.1001126. PubMed PMID: 20862326; PMCID: 2940760.
107. Karczewski KJ, Dudley JT, Kukurba KR, Chen R, Butte AJ, Montgomery SB, Snyder M. Systematic functional regulatory assessment of disease-associated variants. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(23):9607-12. Epub 2013/05/22. doi: 10.1073/pnas.1219099110. PubMed PMID: 23690573; PMCID: 3677437.
108. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, Webb TR, Zeng L, Dehghan A, Alver M, Armasu SM, Auro K, Bjonnes A, Chasman DI, Chen S, Ford I, Franceschini N, Gieger C, Grace C, Gustafsson S, Huang J, Hwang SJ, Kim YK, Kleber ME, Lau KW, Lu X, Lu Y, Lyytikainen LP, Mihailov E, Morrison AC, Pervjakova N, Qu L, Rose LM, Salfati E, Saxena R, Scholz M, Smith AV, Tikkanen E, Uitterlinden A, Yang X, Zhang W, Zhao W, de Andrade M, de Vries PS, van Zuydam NR, Anand SS, Bertram L, Beutner F, Dedoussis G, Frossard P, Gauguier D, Goodall AH, Gottesman O, Haber M, Han BG, Huang J, Jalilzadeh S, Kessler T, Konig IR, Lannfelt L, Lieb W, Lind L, Lindgren CM, Lokki ML, Magnusson PK, Mallick NH, Mehra N, Meitinger T, Memon FU, Morris AP, Nieminen MS, Pedersen NL, Peters A, Rallidis LS, Rasheed A, Samuel M, Shah SH, Sinisalo J, Stirrups KE, Trompet S, Wang L, Zaman KS, Ardissino D, Boerwinkle E, Borecki IB, Bottinger EP, Buring JE, Chambers JC, Collins R, Cupples LA, Danesh J, Demuth I, Elosua R, Epstein SE, Esko T, Feitosa MF, Franco OH, Franzosi MG, Granger CB, Gu D, Gudnason V, Hall AS, Hamsten A, Harris TB, Hazen SL, Hengstenberg C, Hofman A, Ingelsson E, Iribarren C, Jukema JW, Karhunen PJ, Kim BJ, Kooner JS, Kullo IJ, Lehtimaki T, Loos RJ, Melander O, Metspalu A, Marz W, Palmer CN, Perola M, Quertermous T, Rader DJ, Ridker PM, Ripatti S, Roberts R, Salomaa V, Sanghera DK, Schwartz SM, Seedorf U, Stewart AF, Stott DJ, Thiery J, Zalloua PA, O'Donnell CJ, Reilly MP, Assimes TL, Thompson JR, Erdmann J, Clarke R, Watkins H, Kathiresan S, McPherson R, Deloukas P, Schunkert H, Samani NJ, Farrall M, Consortium CAD. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nature genetics. 2015;47(10):1121-30. doi: 10.1038/ng.3396. PubMed PMID: 26343387; PMCID: PMC4589895.
109. Childs EJ, Mocci E, Campa D, Bracci PM, Gallinger S, Goggins M, Li D, Neale RE, Olson SH, Scelo G, Amundadottir LT, Bamlet WR, Bijlsma MF, Blackford A, Borges M, Brennan P, Brenner H, Bueno-de-Mesquita HB, Canzian F, Capurso G, Cavestro GM, Chaffee KG, Chanock SJ, Cleary SP, Cotterchio M, Foretova L, Fuchs C, Funel N, Gazouli M, Hassan M, Herman JM, Holcatova I, Holly EA, Hoover RN, Hung RJ, Janout V, Key TJ, Kupcinskas J, Kurtz RC, Landi S, Lu L, Malecka-Panas E, Mambrini A, Mohelnikova-Duchonova B, Neoptolemos JP, Oberg AL, Orlow I, Pasquali C, Pezzilli R, Rizzato C, Saldia A, Scarpa A, Stolzenberg-Solomon RZ, Strobel O, Tavano F, Vashist YK, Vodicka P, Wolpin BM, Yu H, Petersen GM, Risch HA, Klein AP. Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. Nature genetics. 2015;47(8):911-6. doi: 10.1038/ng.3341. PubMed PMID: 26098869; PMCID: PMC4520746.
110. Wolpin BM, Rizzato C, Kraft P, Kooperberg C, Petersen GM, Wang Z, Arslan AA, Beane-Freeman L, Bracci PM, Buring J, Canzian F, Duell EJ, Gallinger S, Giles GG, Goodman GE, Goodman PJ, Jacobs EJ, Kamineni A, Klein AP, Kolonel LN, Kulke MH, Li D, Malats N, Olson SH, Risch HA, Sesso HD, Visvanathan K, White E, Zheng W, Abnet CC, Albanes D, Andreotti G, Austin MA, Barfield R, Basso D, Berndt SI, Boutron-Ruault MC, Brotzman M, Buchler MW, Bueno-de-Mesquita HB, Bugert P, Burdette L, Campa D, Caporaso NE, Capurso G, Chung C, Cotterchio M, Costello E, Elena J, Funel N, Gaziano JM, Giese NA, Giovannucci EL, Goggins M, Gorman MJ, Gross M, Haiman CA, Hassan M, Helzlsouer KJ, Henderson BE, Holly EA, Hu N, Hunter DJ, Innocenti F, Jenab M, Kaaks R, Key TJ, Khaw KT, Klein EA, Kogevinas M, Krogh V, Kupcinskas J, Kurtz RC, LaCroix A, Landi MT, Landi S, Le Marchand L, Mambrini A, Mannisto S, Milne RL, Nakamura Y, Oberg AL, Owzar K, Patel AV, Peeters PH, Peters U, Pezzilli R, Piepoli A, Porta M, Real FX, Riboli E, Rothman N, Scarpa A, Shu XO, Silverman DT, Soucek P, Sund M, Talar-Wojnarowska R, Taylor PR, Theodoropoulos GE, Thornquist M, Tjonneland A, Tobias GS, Trichopoulos D, Vodicka P, Wactawski-Wende J, Wentzensen N, Wu C, Yu H, Yu K, Zeleniuch-Jacquotte A, Hoover R, Hartge P, Fuchs C, Chanock SJ, Stolzenberg-Solomon RS, Amundadottir LT. Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. Nature genetics. 2014;46(9):994-1000. doi: 10.1038/ng.3052. PubMed PMID: 25086665; PMCID: PMC4191666.
111. Liu C, Wang Y, Huang H, Wang C, Zhang H, Kong Y, Zhang H. Association between CLPTM1L-TERT rs401681 polymorphism and pancreatic cancer risk among Chinese Han population. Tumour Biol. 2014;35(6):5453-7. doi: 10.1007/s13277-014-1711-9. PubMed PMID: 24577890.

112. Parikh H, Jia J, Zhang X, Chung CC, Jacobs KB, Yeager M, Boland J, Hutchinson A, Burdett L, Hoskins J, Risch HA, Stolzenberg-Solomon RZ, Chanock SJ, Wolpin BM, Petersen GM, Fuchs CS, Hartge P, Amundadottir L. A resequence analysis of genomic loci on chromosomes 1q32.1, 5p15.33, and 13q22.1 associated with pancreatic cancer risk. Pancreas. 2013;42(2):209-15. doi: 10.1097/MPA.0b013e318264cea5. PubMed PMID: 23295781; PMCID: PMC3618611.
113. Roberts NJ, Klein AP. Genome-wide sequencing to identify the cause of hereditary cancer syndromes: with examples from familial pancreatic cancer. Cancer Lett. 2013;340(2):227-33. doi: 10.1016/j.canlet.2012.11.008. PubMed PMID: 23196058; PMCID: PMC3652916.
114. Ragvin A, Moro E, Fredman D, Navratilova P, Drivenes O, Engstrom PG, Alonso ME, de la Calle Mustienes E, Gomez Skarmeta JL, Tavares MJ, Casares F, Manzanares M, van Heyningen V, Molven A, Njolstad PR, Argenton F, Lenhard B, Becker TS. Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(2):775-80. doi: 10.1073/pnas.0911591107. PubMed PMID: 20080751; PMCID: PMC2818943.
115. Shen T, Aneas I, Sakabe N, Dirschinger RJ, Wang G, Smemo S, Westlund JM, Cheng H, Dalton N, Gu Y, Boogerd CJ, Cai CL, Peterson K, Chen J, Nobrega MA, Evans SM. Tbx20 regulates a genetic program essential to adult mouse cardiomyocyte function. The Journal of clinical investigation. 2011;121(12):4640-54. doi: 10.1172/JCI59472. PubMed PMID: 22080862; PMCID: PMC3223071.
116. van den Boogaard M, Smemo S, Burnicka-Turek O, Arnolds DE, van de Werken HJ, Klous P, McKean D, Muehlschlegel JD, Moosmann J, Toka O, Yang XH, Koopmann TT, Adriaens ME, Bezzina CR, de Laat W, Seidman C, Seidman JG, Christoffels VM, Nobrega MA, Barnett P, Moskowitz IP. A common genetic variant within SCN10A modulates cardiac SCN5A expression. The Journal of clinical investigation. 2014;124(4):1844-52. doi: 10.1172/JCI73140. PubMed PMID: 24642470; PMCID: PMC3973109.
117. Hidai H, Bardales R, Goodwin R, Quertermous T, Quertermous EE. Cloning of capsulin, a basic helix-loop-helix factor expressed in progenitor cells of the pericardium and the coronary arteries. Mech Dev. 1998;73(1):33-43. Epub 1998/05/28. doi: S0925-4773(98)00031-8 [pii]. PubMed PMID: 9545526.
118. Boutet SC, Quertermous T, Fadel BM. Identification of an octamer element required for in vivo expression of the TIE1 gene in endothelial cells. Biochem J. 2001;360(Pt 1):23-9. PubMed PMID: 11695988.
119. Fadel BM, Boutet SC, Quertermous T. Octamer-dependent in vivo expression of the endothelial cell-specific TIE2 gene. J Biol Chem. 1999;274(29):20376-83. PubMed PMID: 10400661.
120. Hidai C, Zupancic T, Penta K, Mikhail A, Kawana M, Quertermous EE, Aoka Y, Fukagawa M, Matsui Y, Platika D, Auerbach R, Hogan BL, Snodgrass R, Quertermous T. Cloning and characterization of developmental endothelial locus-1: an embryonic endothelial cell protein that binds the alphavbeta3 integrin receptor. Genes Dev. 1998;12(1):21-33. PubMed PMID: 9420328.
121. Chun HJ, Ali ZA, Kojima Y, Kundu RK, Sheikh AY, Agrawal R, Zheng L, Leeper NJ, Pearl NE, Patterson AJ, Anderson JP, Tsao PS, Lenardo MJ, Ashley EA, Quertermous T. Apelin signaling antagonizes Ang II effects in mouse models of atherosclerosis. The Journal of clinical investigation. 2008;118(10):3343-54. Epub 2008/09/05. doi: 10.1172/JCI34871. PubMed PMID: 18769630; PMCID: 2525695.
122. Kojima Y, Kundu RK, Cox CM, Leeper NJ, Anderson JA, Chun HJ, Ali ZA, Ashley EA, Krieg PA, Quertermous T. Upregulation of the apelin-APJ pathway promotes neointima formation in the carotid ligation model in mouse. Cardiovascular research. 2010;87(1):156-65. Epub 2010/02/24. doi: cvq052 [pii]
10.1093/cvr/cvq052. PubMed PMID: 20176814; PMCID: 2883899.
123. Tabibiazar R, Wagner RA, Ashley EA, King JY, Ferrara R, Spin JM, Sanan DA, Narasimhan B, Tibshirani R, Tsao PS, Efron B, Quertermous T. Signature patterns of gene expression in mouse atherosclerosis and their correlation to human coronary disease. Physiol Genomics. 2005;22(2):213-26. Epub 2005/05/05. doi: 00001.2005 [pii]
10.1152/physiolgenomics.00001.2005. PubMed PMID: 15870398.
124. Ishida T, Choi SY, Kundu RK, Spin J, Yamashita T, Hirata K, Kojima Y, Yokoyama M, Cooper AD, Quertermous T. Endothelial lipase modulates susceptibility to atherosclerosis in apolipoprotein-E-deficient mice. J Biol Chem. 2004;279(43):45085-92. Epub 2004/08/12. doi: 10.1074/jbc.M406360200
M406360200 [pii]. PubMed PMID: 15304490.
125. Roche-Molina M, Sanz-Rosa D, Cruz FM, Garcia-Prieto J, Lopez S, Abia R, Muriana FJ, Fuster V, Ibanez B, Bernal JA. Induction of sustained hypercholesterolemia by single adeno-associated virus-mediated

gene transfer of mutant hPCSK9. Arteriosclerosis, thrombosis, and vascular biology. 2015;35(1):50-9. doi: 10.1161/ATVBAHA.114.303617. PubMed PMID: 25341796.