Title
Taking Data Science to Heart: Next Scale of Gene Regulation

Journal
Current Cardiology Reports, 23(5)

ISSN
1523-3782

Authors
Chapski, Douglas J
Vondriska, Thomas M

Publication Date
2021-05-01

DOI
10.1007/s11886-021-01467-6

Peer reviewed

HHS Public Access

Author manuscript

Curr Cardiol Rep. Author manuscript; available in PMC 2022 March 15.

Published in final edited form as:

Curr Cardiol Rep. ; 23(5): 46. doi:10.1007/s11886-021-01467-6.

Taking Data Science to Heart: Next Scale of Gene Regulation

Douglas J. Chapski, PhD[1], **Thomas M. Vondriska, PhD**[1,2,3]

[1]Department of Anesthesiology & Perioperative Medicine, David Geffen School of Medicine at University of California, Los Angeles

[2]Department of Medicine, David Geffen School of Medicine at University of California, Los Angeles

[3]Department of Physiology, David Geffen School of Medicine at University of California, Los Angeles

Abstract

Purpose of Review: Technical advances have facilitated high-throughput measurements of the genome in the context of cardiovascular biology. These techniques bring a deluge of gargantuan datasets, which in turn present two fundamentally new opportunities for innovation—data processing and knowledge integration—toward the goal of meaningful basic and translational discoveries.

Recent findings: Big data, integrative analyses, and machine learning have brought cardiac investigations to the cutting edge of chromatin biology, not only to reveal basic principles of gene regulation in the heart, but also to aid in the design of targeted epigenetic therapies.

Summary: Cardiac studies using big data are only beginning to integrate the millions of recorded data points and the tools of machine learning are aiding this process. Future experimental design should take into consideration insights from existing genomic datasets, thereby focusing on heretofore unexplored epigenomic contributions to disease pathology.

Keywords

Genomics; chromatin; bioinformatics; machine learning; transcriptomics; cardiovascular disease

Introduction

Heart failure remains one of the most common causes of death [1]. The syndrome is characterized by a cardiac myocyte transcriptome that becomes deranged, recapitulating some aspects of a developmentally primitive myocyte [2]. To uncover mechanisms

Correspondence: Douglas J. Chapski, PhD, Department of Anesthesiology & Perioperative Medicine, David Geffen School of Medicine at UCLA, CHS 37-100, 650 Charles Young Dr., Los Angeles, CA 90095, dchapski@ucla.edu.

Conflicts of Interest: Douglas J. Chapski and Thomas M. Vondriska report no conflicts of interest.

underpinning the transcriptional disarray observed in heart failure, recent investigations have investigated changes in chromatin organization after cardiac insult. Structure dictates function in biology, so it can be instructive to consider the levels of structural regulation of the epigenome: chemical modifications to the sequence of DNA; posttranslational modification of histone octamers that comprise nucleosomes; accessibility of chromatin fibers composed DNA wrapped around nucleosomes, chromatin structural proteins and modifiers; chromatin compartmentalization; and gene localization with respect to the nuclear periphery (reviewed in [3]). A recurring observation in the study of gene regulation is the genome-wide nature of chromatin structural features—that is, reorganization of chromatin occurs seemingly simultaneously (from a developmental or disease perspective) at multiple specific locations across the genome. Precise measurement of the loci at which these changes occur requires genome-wide high-throughput sequencing based techniques, which are now widely available. However, these technical advances have led to a proliferation in large datasets, the biological meaning of which can be understood through close collaboration with data scientists and computational biologists working as equal partners in the experimental design and data analysis stages of research (Figure 1).

Mechanisms of Gene Regulation Revealed by Early Forays into Cardiac Transcriptomics

Initial work in transcriptomics laid the foundation for how we think about gene regulation in the heart: we now understand that thousands of genes undergo transcriptional changes with pathological perturbation. Each dataset came with challenges and required some creativity to make sense of the big data tables that, until a scientist framed a question and tested a hypothesis, were inert. In 2009, one lab used array-based techniques to reveal an RNA expression paradigm wherein the transcription of differentially expressed microRNAs with heart failure becomes normalized after mechanical unloading, but transcription of differentially expressed mRNAs does not, suggesting a distinct regulatory rubric for non-protein coding RNA species in cardiac disease [4]. These findings were important for spurring the study of cardiac gene regulation forward, however they relied on array-based data. Early open source statistical methods to determine differentially expressed genes were developed for microarrays [5], and were further honed for RNA-seq approaches to examine both known and novel transcripts [6–9], with important considerations made to correct for multiple testing [10, 11]. To this end, the same lab published a 2010 RNA-seq study using these tools to elucidate transcriptional changes in a murine Gαq transgenic model that results in cardiac pathology [12]. Notably, this RNA-seq experiment detected low-abundance transcripts that were not detected in a side-by-side microarray comparison of the same samples [12]. In 2012, another group integrated mRNA and microRNA deep sequencing experiments in mice to reveal that the protective effect of increased PI3Kα signaling during cardiac hypertrophy is mediated by a decrease in TGF-β signaling and reduction miR-21 expression, which results in less fibrosis [13]. In early 2014, the same group then performed deep mRNA and microRNA sequencing experiments in human hearts to understand how noncoding RNA signature differentiate failing vs. nonfailing hearts [14]. This study shines a light on long noncoding RNAs (lncRNAs) as a subset of differentially expressed RNAs during heart failure whose expression *recovers* after mechanical unloading in patients,

suggesting a role for noncoding RNA in shaping the regulatory logic for disease gene expression regimes [14]. A question that persisted during these investigations was *how* these transcripts were regulated during disease, from a chromatin structural perspective, and measurements were already taking place.

Chromatin as a Target in the Heart

Early studies of histone deacetylase (HDAC) inhibition led to FDA approved therapies [15], and their use in the heart looks promising [16]. However, to date little is known about where along the genome individual HDAC isoforms localize, and whether this is altered during the development of disease. On the other hand, histone marks themselves, have been studied by several labs (reviewed in [3] and examples described in the next section). Nevertheless, studies measuring phenotypic outcomes continue to highlight a rationale not only for casting a wider net to discover other chromatin players in heart failure and developing therapeutic modulators of chromatin structure, but also for investigating how chromatin is disrupted in disease. In measuring global chromatin features, big data and computation interrogation of datasets are essential for gaining mechanistic understanding of pathological cardiovascular processes.

Individual configurations of chromatin features contribute to gene regulation differently, depending on how entrenched they are before disease. For example, mutations in a noncoding region near the human *PITX2* locus are associated with atrial fibrillation [17], and a 2019 study showed chromatin interaction between an enhancer from an orthologous region and the mouse *Pitx2c* promoter [18]. Notably, disruption of this interaction by deleting the enhancer—or by independent disruption of a binding site for the chromatin structural protein CTCF within *Pitx2c*—results in increased susceptibility to atrial fibrillation [18]. Thus, we could hypothesize that different levels of chromatin organization likely contribute to cardiac pathology with varying temporal and/or pathological impact. Accordingly, we pursue knowledge of how each level of chromatin organization fits into the context of the global cardiac gene regulation, with the end goal of designing better therapies for heart failure.

New Genomic Tools Advance the Cardiac Chromatin Field

Each level of chromatin organization, when measured with genomics approaches, reveals a huge breadth of biological information. As a result, initial studies examined individual datasets from a single type of epigenomic experiment (e.g. ChIP-seq for a given protein or histone mark) to make conclusions about chromatin in the heart. Early work examining DNA methylation shed light on gene regulation in failing hearts. In 2011, one group used a methylated DNA immunoprecipitation approach followed by deep sequencing and observed DNA methylation dynamics in the promoters of upregulated, but not in downregulated, genes in failing human hearts [19]. In 2016, our lab evaluated DNA methylation dynamics in mouse strains susceptible or resistant to isoproterenol-induced heart failure and revealed strain-specific methylation patterns in the basal, unstressed heart, that presaged the cardiac phenotype following adrenergic stress [20]. These studies distilled big datasets composed of millions of DNA methylation measurements to reveal a role for the DNA methylome (in

addition to genetic sequence) in determining susceptibility to heart failure and subsequently established a rationale for studying the modulation of DNA methylation as a treatment for heart failure. Interestingly, in 2018, a group showed that chemical inhibition of DNA methyltransferases blunts the cardiac hypertrophy observed with pressure overload in mice, suggesting that global chromatin treatment may be a therapy for heart failure, although this specific perturbation of chromatin did not result in widespread changes in DNA methylation [21].

Individual examinations of polymerase, histone mark and transcription factor occupancy in the genome showed modest success in understanding how transcription of diseased hearts becomes dysregulated after pathological stimulus. A 2013 study revealed that RNA Polymerase II undergoes transcriptional pause release at a subset of housekeeping genes modulated by pressure overload hypertrophy or *de novo* recruitment at “specialized genes” [22]. A 2015 follow-up study from the same lab highlighted TFIIB (a member of the preinitiation complex) as a possible therapeutic target in heart failure when they demonstrated that antisense oligo-mediated inhibition of TFIIB transcripts resulted in abolished transcription of cardiac disease genes after pressure overload [23]. As of yet, no one has repeated this strategy to prevent transcription of heart failure genes, but the technique may prove useful in future studies.

A 2013 examination of cardiac histone marks—post-translational modifications that affect accessibility of nucleosomes—demonstrated that enhancers are occupied by H3K27ac (histone 3 lysine 27 acetylation) in health and disease [24]. In 2015, another study showed that pressure overload induced H3K9/K14 acetylation is abrogated when mice are treated with the HDAC inhibitor trichostatin A, suggesting that global treatment at the chromatin level modulates transcriptional readouts to ameliorate cardiac pathology [25]. These studies defined gene regulatory landscapes in both control and pressure overloaded murine hearts and established a rationale for examining regulatory regions in the context of transcription factor binding in cardiac cells, and for examining chromatin accessibility—a direct readout of the possibility for transcription factor binding—in the heart.

Cardiac transcription factors are technically challenging to immunoprecipitate from chromatin, but they may provide clues as to which cardiac enhancers are active at a given developmental or pathological stage. In 2011 a cardiac gene regulation lab overcame this barrier by expressing biotinylated cardiac transcription factor constructs in the HL1 cardiomyocyte cell line and performing streptavidin-based pulldowns followed by deep sequencing [26]. This approach revealed a subset of active enhancers that are occupied by multiple transcription factors independent of the enhancer associated protein p300, a newly identified class of cardiac enhancers [26]. In 2014, the same lab went on to establish an *in vivo* knock-in based method where they used streptavidin to pull down biotinylated GATA4 from mouse hearts [27]. This technique built on their 2011 study and revealed distinct GATA4 binding sites found only in pressure overload hearts—in addition to GATA4 sites that are found in both banded and developing hearts—and may hold clues regarding the logic for transcription factor based pathological gene activation during heart failure [27]. In 2019, the group used this *in vivo* biotinylated factor approach to measure occupancy of seven cardiac transcription factors in murine hearts, and the data suggested that there exist a

subset of enhancers co-occupied by multiple cardiac transcription factors that do not have H3K27ac binding (a marker of active enhancers) [28]. Moreover, this study demonstrated a higher chromatin accessibility in multi- vs. single-factor bound enhancers, suggesting that multi-factor bound regions are likely to be functional in the heart [28]. Taken together, this lab has overcome technical and conceptual challenges to examine pathological gene expression in the heart. Not only did the group develop a creative method to immunoprecipitate cardiac transcription factors, but they also generated *in vivo* models and integrated complex datasets downstream of challenging analytical pipelines to bring the field into a new mindset that embraces the complexity of gene regulatory programs during pathology. Increasingly, chromatin studies are integrating several datasets, similar to the example described above, to make sense of what is going on at distinct levels of chromatin organization in a given disease model.

Integrating Datasets: Data Gets Bigger and Biological Insights More Impactful

A piecemeal approach works for understanding individual components of chromatin organization in cardiac cells, however more recently the field has appreciated integration of a variety of massive datasets to more powerfully illustrate the global picture of chromatin organization. A challenge with epigenomics is not how many data points exist within a dataset, but the statistical determination of which ones are important for biological inference. For example, a ChIP-seq experiment requires straightforward processing of millions of reads to determine occupancy of an immunoprecipitated factor at thousands of loci. Although this is straightforward for a computer to calculate, a statistical determination of significant changes in occupancy requires biological replicates and clever methods to process the thousands of data points and to perform statistical tests within a reasonable timeframe. As multiple factors can be immunoprecipitated during a given investigation, including at several time points, the complexity of an experiment requires increasing computational sophistication to integrate datasets. In addition, data analysis becomes even more complex as orthogonal experiments are integrated into the workflow.

In the context of heart disease, measuring chromatin architecture along with other features of chromatin structure is as a more comprehensive method to examine the dynamic nucleus during heart failure, especially in tandem with a pathological stimulus that disrupts global chromatin architecture as a positive control for genomic disorganization. In 2017, we showed that chromatin structure, as measured by high throughput chromatin conformation capture (Hi-C), is globally disrupted with pressure overload in mice, as well as in a cardiac-specific CTCF depletion model [29]. The chromatin structural protein CTCF is important for mediating stable interaction landscapes in the healthy cardiac myocyte: our study revealed that depletion of CTCF had a minor overall effect on genome structure, but was sufficient to induce pathologic gene activation and heart failure [29]. This investigation required generation of large chromatin interaction matrices comprised of millions of data points, followed by use of statistical techniques to determine which represented bona fide structural features. The results enabled determination of which regions of the genome physically interact with each other in the three-dimensional context of the nucleus, a key insight for

understanding gene regulation *in vivo*. This information was then integrated with gene expression and protein binding data, allowing us to distill billions of measurements into a comprehensible set of biologically meaningful observations. This heart failure phenotype, along with several of the conclusions about the changes to chromatin structure during pressure overload, were reproduced using similar molecular approaches but with a distinct CTCF depletion strategy [30].

In 2017, a Hi-C and DNA methylation study was reported in which it was shown that compartmentalization of the nucleus into active and inactive structural compartments occurs during differentiation while DNA methylation patterns take longer and are established slowly during development [31]. Although not chromatin structure based in nature, a 2017 study identified two classes of enhancers in three fetal human cardiac cells types and their induced pluripotent cell counterparts: one class with H3K4me1 and H3K4me3 deposition in all cell types, and another with cell-specific histone mark deposition [32]. This study showed a utility in studying chromatin interactions between distal regulatory regions and gene promoters in the future, because most genetic variation associated with disease in humans occurs in non-coding regions and thus may be regulatory in nature. To understand single nucleotide polymorphisms (SNPs) associated with cardiovascular disease, which are typically localized to non-coding regions, one lab published a 2018 promoter capture Hi-C experiment that demonstrates physical contact of ~2000 disease associated SNPs to hundreds of genes along the genome of human induced pluripotent stem cell derived cardiac myocytes [33]. The integration of chromatin structural data with SNP data was a powerful approach to reveal that the majority of SNP-gene interactions (greater than 90%) do not occur with the closest gene to a given SNP [33].

An innovative approach to recover endogenous structural information from epigenomic experiments involves 3D chromatin modeling using Hi-C contacts, which employs statistical learning techniques to optimize 3D positioning of pairwise contacts [34]. This technique contextualizes pairwise interactions into a 3D reconstruction of genomic structure, which can provide clues about chromatin structural regulation beyond the 2D maps generated from traditional Hi-C analyses. In 2018, we collaborated with the Alber lab to generate 3D models from cardiac myocytes and liver cells and demonstrated distinct chromatin structural strategies underlying organ-specific gene expression patterns [35]. This class of modeling is powerful because it takes pairwise contact data from a population of cells, uses statistical methods to predict thousands of structures that would likely occur within single cells, and then allows for prediction of how chromatin is organized in a cell population. In our study, we investigated relative radial positioning of cardiac and liver specific genes within the nucleus, as well as interchromosomal chromatin interactions, to understand the chromatin structural logic underlying organ-specific transcription [35]. Such an approach helps drive hypothesis generation within cardiac biology and can easily be applied to other organ systems. For example, during pathology down-regulated genes might be pushed towards the nuclear periphery or heterochromatin centers (as has been shown in microscopy experiments [36]), regions associated with inactivation. Access to a 3D model of the entire genome allows for direct testing of this phenomenon across all genes—an exercise that would be presently impossible with microscopy approaches.

Machine Learning in Cardiac Chromatin

The field of machine learning is vast, with innumerable discoveries and algorithmic improvements being made each year. A major utility of machine learning over manual statistical methods is the leveraging of computational mathematics to perform three types of tasks with minimal interference from humans: supervised, unsupervised, and reinforcement learning [37]. During supervised learning, input vectors and their outcomes from a training dataset are used to perform classification or regression techniques on new input data [37]. For example, a clinical data vector can be used to predict whether a new patient has heart failure using a model built from clinical data and outcomes from thousands of previous patients. Unsupervised learning allows for pattern recognition via clustering, density estimation, and/or dimensionality reduction [37]. For instance, patterns indicative of a pathological state could be mathematically untangled from a series of epigenomic or other large datasets, after using unsupervised clustering techniques on the datasets from healthy and sick patients. Reinforcement learning makes use of a positive and negative reward system to optimize an outcome without knowing what it should look like *a priori* [37]. An example of this would be an algorithm learning to beat an expert player at chess.

Some of the most cutting-edge cardiac machine learning work is being done with a clinical measurement used daily by the cardiologist: the electrocardiogram (ECG). A 2019 study used a deep learning approach to classify 12 subsets of cardiac rhythm using single-lead ECGs from over 53,000 patients [38]. The astounding conclusion was that the F1 score, a readout of the precision and recall, for their classifier was better than the one calculated using annotations from board certified cardiologists [38]. Notably, this study had high patient enrollment, which allowed for an adequate training set and a separate cohort as a test set. Altogether, this machine learning approach could impact the clinic by helping healthcare providers prioritize certain classes arrhythmias over others or by detecting potentially life-threatening ones faster than currently available technologies.

Computational investigators tend to deposit their work on preprint servers—internet databases of unpublished work—*before* sending manuscripts to journals so that the world can evaluate studies in tandem with reviewers. Platforms such as arXiv, bioRxiv, medRxiv, and ChemRxiv are useful for dissemination of scientific work and for nucleation of scientific discussion in their respective fields (Table 1). Despite the utility of preprint servers, uploaded manuscripts must be taken with a grain of salt since preprints are not yet evaluated by expert reviewers (discussed in the cardiovascular context in [39]), although online discussion of cutting-edge work has brough a novel speed to incorporation of new datasets into experimental workflows and has fostered more open communication between investigators.

One example of the utility of machine learning in chromatin is a 2013 single author publication by Steve Horvath that describes a statistical model to predict biological age using DNA methylation status of cytosines across the genome [40]. This investigation revealed that “DNA methylation age”, or biological age as predicted by epigenomic features from the model, becomes accelerated in cancer [40]. In other words, predicted DNA methylation age is significantly higher than actual chronological age in cancer samples [40].

Although the original model building did not test cardiac hypotheses, a follow up study from 2016 showed no relationship between epigenetic age and coronary heart disease incidence [41], suggesting that coronary heart disease may have other epigenetic markers—independent of DNA methylation-based epigenetic age—that predict disease incidence. In 2019, Horvath used data from the Framingham Heart Study [42] to refine his epigenetic aging models and predict lifespan (a unit called GrimAge) and age acceleration (AgeAccelGrim), in addition to seven DNA methylation based surrogate biomarkers such as a smoking pack years [43]. Strikingly, in a validation dataset made up of diverse patient cohorts, predicted age acceleration and several surrogate biomarkers including smoking pack years are predictive of time-to-coronary heart disease [43], which illustrates the power of statistical learning in evaluating cardiovascular health. Another study that shed light on the diagnostic potential of machine learning in the cardiac field was a 2018 examination of genotype and cytosine methylation to predict five year incidence of coronary heart disease [44]. In this investigation, authors used a random forest approach (a task that makes use of many decision trees to optimize a set of candidate predictors that do not necessarily need to have a linear relationship) [45] to discover 4 genetic loci and 4 cytosines whose respective genotype and methylation status together predict coronary heart disease [44]. The training and test datasets for this retrospective study consisted of 1180 and 524 individuals, respectively, and while the model performs well on a large sample size, it relies on published data [44]. A more powerful future approach would be to perform a prospective study, enrolling thousands of patients and measuring their genotype and DNA methylation status to determine whether this strategy can work in a predictive manner.

An elephant in the room of computational biology is that machine learning techniques require sufficient sample size to perform model training and subsequent testing. Low n results in poor model performance on new datasets, typically due to overfitting. The above-mentioned studies used thousands of patients in their experimental workflow, however sometimes this is not a possibility in the basic research realm. Despite this, in the cardiac chromatin field, millions of measurements can be made in thousands of cells during a given sequencing experiment, and these cells could comprise a large enough n to produce training and test datasets for machine learning applications.

Cellular diversity in the heart is now being quantitatively deciphered during cardiac genomics investigations. In 2017, one group performed RNA-seq on four FACS-sorted cell types from healthy and infarcted hearts and revealed that adult cardiac myocytes and endothelial cells do not revert to a “fetal-like” transcriptional program after myocardial infarction, but that fibroblasts and leukocytes do [46]. This investigation addressed cell type heterogeneity in the heart by examining a sorted, pre-defined subset of cell types, yet other cell types were missing in addition to any assessment of cell-to-cell variability. To address this, single-cell sequencing approaches—although not machine learning *per se*—use advanced dimensionality reduction and statistical classification techniques to garner insights about cellular diversity in the heart. In 2017, a single-cell RNA-seq experiment from mouse and human left ventricles showed that distinct subpopulations of cardiac myocytes undergo different transcriptional regulation with heart failure, and that long intergenic noncoding RNAs may play a role in the regulatory landscape engendered by pathological processes [47]. This approach relied on weighted gene correlation network analysis [48] and

traditional principal component analysis for dimensionality reduction [47]. A single-cell strategy provides advantages over bulk sequencing techniques because it reveals the extent of differential gene expression across a cell type within an organ. To help with this determination, a relatively new machine learning technique that is gaining popularity in the genomics realm is t-SNE, or t-distributed stochastic neighbor embedding [49]. This nonlinear dimensionality reduction technique condenses high dimensional data into two dimensions and works well for data interpretation across many disciplines. The utility of the technique is that it provides two-dimensional visualization of sample clusters, meaning one could use this strategy to tease out a transcriptional behavior within a subpopulation of cells in a visually comprehensible manner [50]. In 2018, a single-cell RNA-seq study in mice used t-SNE as part of a bioinformatics toolkit to reveal cytoskeleton associated protein 4 as a previously unknown activated fibroblast marker [51]. UMAP, or Uniform Manifold Approximation and Projection, is a more powerful dimensionality reduction technique for single cell experiments because it outperforms t-SNE with larger datasets while preserving global dataset structure, thereby allowing visual comparison between clusters on a lower dimensional space [52]. A 2019 study used UMAP to reduce the dimensionality of a single cell RNA-seq experiment measuring dissected cardiogenic regions from three stages of mouse embryonic development with the end goal of better understanding the subpopulation structures of mesodermal and neural crest cells over time [53]. Notably, t-SNE and UMAP are easily deployable on transcriptomic data via a variety of software resources such as updated versions of Monocle [54], CellRanger [55] and Seurat [56, 57] which all make single-cell data analysis accessible to dry lab novices and experts alike. These and other useful tools for computational biology are outlined in Table 2.

In 2019, one lab combined single-cell RNA-seq, ATAC-seq (a measure of chromatin accessibility in the heart using high-throughput sequencing), and ChIP-seq of three transcription factors to build a machine learning model that predicts which transcription factors are important for early fibroblast reprogramming into cardiac myocytes [58]. Also in 2019, another group published their single-cell RNA-seq on $CD45^+$ cells from sham and pressure overloaded mouse hearts and garnered a new appreciation for the diversity of immune cell types that become activated during heart disease—it was far more cell types than previously thought [59]. In 2020, a single-nucleus RNA-seq study of almost 300,000 nuclei demonstrated between-chamber and sex differences of transcriptional programs in the human heart [60]. Another examination of single-cell and -nucleus RNA-seq data from adult human heart revealed 5 respective subpopulations of ventricular and atrial cardiac myocytes, in addition to subpopulations of fibroblasts, immune, and vascular cells, among other cell types [61]. A single-nucleus and -cell RNA-seq study of fetal gene expression revealed 77 cell types across human fetal tissues (including the heart), 54 of which exist in only one organ [62]. In addition to common cell type markers, this investigation used unbiased statistical approaches to identify novel cell type markers, which will enable biological roles of these cells to be interrogated in future [62]. Importantly, these datasets serve as resources for understanding the landscape of druggable targets given the cell type diversity in human cardiac biology. Taken together, single-cell sequencing approaches have provided ample data for deployment of machine learning approaches, and investigation of single cells after

pathological stimulus will continue revealing the extent to which gene regulatory environments in distinct cell types work together or antagonistically during disease.

Conclusions and Future Directions

Making use of established datasets to integrate published information with new measurements will help investigators to paint a more comprehensive picture of chromatin in disease. A goal for computational biology is near real-time integration of epigenomic datasets, independent of the lab in which they were acquired—similar to a blood pressure measurement, ECG or troponin test. Furthermore, modeling of epigenomes needs to become dynamic, taking into account cell-to-cell variability and changes over time due to normal physiological development or pathologic stimuli. Probabilistic modeling and machine learning can facilitate such model building, while also identifying (and quantifying) previously unrecognized emergent properties in chromatin that correspond to changes in heart health. For example, a 3D representation of the genome may reveal a structural or accessibility feature corresponding to health or disease that cannot be identified by any single epigenomic measurement alone. Such approaches can advance basic biology as well as our understanding of disease.

We advocate for incorporation of wet and dry lab training components to training regimens to foster the cultivation of more diverse technical repertoires. Data mining and new dataset generation will shape how we answer chromatin questions in the coming years. Importantly, understanding how computers solve problems (differently from the way humans do), and how to frame questions computationally, will foment a shared vocabulary that brings projects to completion. Team members need not have all the big data skills, but a collaborative environment is key to success in large-scale epigenomic investigations. The QCBio Collaboratory at UCLA stands out as an exemplary resource that provides training to non-programmers and offers collaboration to answer biological questions [63]. Moreover, the Collaboratory advocates for use of open source tools, which makes genomics datasets accessible to the lay scientist. Many bioinformatics tools already exist—and others will be developed to usher in new understanding—but fundamental knowledge of how computers work and how to answer questions with big data will continue to empower scientists to test the most meaningful hypotheses with the appropriate tools to reveal novel insights about cardiac biology.

Acknowledgments:

We regret the space considerations have not allowed us to cite many excellent papers in the field. Research in the Vondriska laboratory is supported by the NIH, the David Geffen School of Medicine and the Department of Anesthesiology and Perioperative Medicine at UCLA.

References

Papers of particular interest, published recently, have been highlighted as:

• Of importance

•• Of major importance

1. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP et al. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. Circulation. 2020;141(9):e139–e596. doi:10.1161/CIR.0000000000000757. [PubMed: 31992061]
2. Rajabi M, Kassiotis C, Razeghi P, Taegtmeyer H. Return to the fetal gene program protects the stressed heart: a strong hypothesis. Heart Fail Rev. 2007;12(3–4):331–43. doi:10.1007/s10741-007-9034-1. [PubMed: 17516164]
3. Rosa-Garrido M, Chapski DJ, Vondriska TM. Epigenomes in Cardiovascular Disease. Circ Res. 2018;122(11):1586–607. doi:10.1161/CIRCRESAHA.118.311597. [PubMed: 29798902]
4. Matkovich SJ, Van Booven DJ, Youker KA, Torre-Amione G, Diwan A, Eschenbacher WH et al. Reciprocal regulation of myocardial microRNAs and messenger RNA in human cardiomyopathy and reversal of the microRNA signature by biomechanical support. Circulation. 2009;119(9):1263–71. doi:10.1161/CIRCULATIONAHA.108.813576. [PubMed: 19237659]
5. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98(9):5116–21. doi:10.1073/pnas.091062498. [PubMed: 11309499]
6. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5. doi:10.1038/nbt.1621. [PubMed: 20436464]
7. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31(1):46–53. doi:10.1038/nbt.2450. [PubMed: 23222703]
8. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106. [PubMed: 20979621]
9. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi:10.1186/s13059-014-0550-8. [PubMed: 25516281]
10. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440–5. doi:10.1073/pnas.1530509100. [PubMed: 12883005]
11. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.• Original statistical paper describing the Benjamini-Hochberg method for multiple testing correction
12. Matkovich SJ, Zhang Y, Van Booven DJ, Dorn GW 2nd. Deep mRNA sequencing for in vivo functional analysis of cardiac transcriptional regulators: application to Galphaq. Circ Res. 2010;106(9):1459–67. doi:10.1161/CIRCRESAHA.110.217513. [PubMed: 20360248]
13. Yang KC, Ku YC, Lovett M, Nerbonne JM. Combined deep microRNA and mRNA sequencing identifies protective transcriptomal signature of enhanced PI3Kalpha signaling in cardiac hypertrophy. J Mol Cell Cardiol. 2012;53(1):101–12. doi:10.1016/j.yjmcc.2012.04.012. [PubMed: 22580345]
14. Yang KC, Yamada KA, Patel AY, Topkara VK, George I, Cheema FH et al. Deep RNA sequencing reveals dynamic regulation of myocardial noncoding RNAs in failing human heart and remodeling with mechanical circulatory support. Circulation. 2014;129(9):1009–21. doi:10.1161/CIRCULATIONAHA.113.003863. [PubMed: 24429688]
15. McKinsey TA, Olson EN. Toward transcriptional therapies for the failing heart: chemical screens to modulate genes. J Clin Invest. 2005;115(3):538–46. doi:10.1172/JCI24144. [PubMed: 15765135]
16. McKinsey TA. Targeting inflammation in heart failure with histone deacetylase inhibitors. Mol Med. 2011;17(5–6):434–41. doi:10.2119/molmed.2011.00022. [PubMed: 21267510]
17. Roselli C, Chaffin MD, Weng LC, Aeschbacher S, Ahlberg G, Albert CM et al. Multi-ethnic genome-wide association study for atrial fibrillation. Nat Genet. 2018;50(9):1225–33. doi:10.1038/s41588-018-0133-9. [PubMed: 29892015]
18. Zhang M, Hill MC, Kadow ZA, Suh JH, Tucker NR, Hall AW et al. Long-range Pitx2c enhancer-promoter interactions prevent predisposition to atrial fibrillation. Proc Natl Acad Sci U S A. 2019;116(45):22692–8. doi:10.1073/pnas.1907418116. [PubMed: 31636200]

19. Movassagh M, Choy MK, Knowles DA, Cordeddu L, Haider S, Down T et al. Distinct epigenomic features in end-stage failing human hearts. Circulation. 2011;124(22):2411–22. doi:10.1161/CIRCULATIONAHA.111.040071. [PubMed: 22025602]

20. Chen H, Orozco LD, Wang J, Rau CD, Rubbi L, Ren S et al. DNA Methylation Indicates Susceptibility to Isoproterenol-Induced Cardiac Pathology and Is Associated With Chromatin States. Circ Res. 2016;118(5):786–97. doi:10.1161/CIRCRESAHA.115.305298. [PubMed: 26838786]

21. Stenzig J, Schneeberger Y, Loser A, Peters BS, Schaefer A, Zhao RR et al. Pharmacological inhibition of DNA methylation attenuates pressure overload-induced cardiac hypertrophy in rats. J Mol Cell Cardiol. 2018;120:53–63. doi:10.1016/j.yjmcc.2018.05.012. [PubMed: 29792884]

22. Sayed D, He M, Yang Z, Lin L, Abdellatif M. Transcriptional regulation patterns revealed by high resolution chromatin immunoprecipitation during cardiac hypertrophy. J Biol Chem. 2013;288(4):2546–58. doi:10.1074/jbc.M112.429449. [PubMed: 23229551]

23. Sayed D, Yang Z, He M, Pfleger JM, Abdellatif M. Acute targeting of general transcription factor IIB restricts cardiac hypertrophy via selective inhibition of gene transcription. Circ Heart Fail. 2015;8(1):138–48. doi:10.1161/CIRCHEARTFAILURE.114.001660. [PubMed: 25398966]

24. Papait R, Cattaneo P, Kunderfranco P, Greco C, Carullo P, Guffanti A et al. Genome-wide analysis of histone marks identifying an epigenetic signature of promoters and enhancers underlying cardiac hypertrophy. Proc Natl Acad Sci U S A. 2013;110(50):20164–9. doi:10.1073/pnas.1315155110. [PubMed: 24284169]

25. Ooi JY, Tuano NK, Rafehi H, Gao XM, Ziemann M, Du XJ et al. HDAC inhibition attenuates cardiac hypertrophy by acetylation and deacetylation of target genes. Epigenetics. 2015;10(5):418–30. doi:10.1080/15592294.2015.1024406. [PubMed: 25941940]

26. Hse A, Kong SW, Ma Q, Pu WT. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc Natl Acad Sci U S A. 2011;108(14):5632–7. doi:10.1073/pnas.1016959108. [PubMed: 21415370]

27. He A, Gu F, Hu Y, Ma Q, Ye LY, Akiyama JA et al. Dynamic GATA4 enhancers shape the chromatin landscape central to heart development and disease. Nat Commun. 2014;5:4907. doi:10.1038/ncomms5907. [PubMed: 25249388]

28. Akerberg BN, Gu F, VanDusen NJ, Zhang X, Dong R, Li K et al. A reference map of murine cardiac transcription factor chromatin occupancy identifies dynamic and conserved enhancers. Nat Commun. 2019;10(1):4907. doi:10.1038/s41467-019-12812-3. [PubMed: 31659164]

29. Rosa-Garrido M, Chapski DJ, Schmitt AD, Kimball TH, Karbassi E, Monte E et al. High-Resolution Mapping of Chromatin Conformation in Cardiac Myocytes Reveals Structural Remodeling of the Epigenome in Heart Failure. Circulation. 2017;136(17):1613–25. doi:10.1161/CIRCULATIONAHA.117.029430. [PubMed: 28802249] • Investigation revealing deranged chromatin structure in heart failure and in a cardiac-specific model of chromatin disruption that results in pathological cardiac phenotype

30. Lee DP, Tan WLW, Anene-Nzelu CG, Lee CJM, Li PY, Luu TDA et al. Robust CTCF-Based Chromatin Architecture Underpins Epigenetic Changes in the Heart Failure Stress-Gene Response. Circulation. 2019;139(16):1937–56. doi:10.1161/CIRCULATIONAHA.118.036726. [PubMed: 30717603]

31. Nothjunge S, Nuhrenberg TG, Gruning BA, Doppler SA, Preissl S, Schwaderer M et al. DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes. Nat Commun. 2017;8(1):1667. doi:10.1038/s41467-017-01724-9. [PubMed: 29162810]

32. Zhao MT, Shao NY, Hu S, Ma N, Srinivasan R, Jahanbani F et al. Cell Type-Specific Chromatin Signatures Underline Regulatory DNA Elements in Human Induced Pluripotent Stem Cells and Somatic Cells. Circ Res. 2017;121(11):1237–50. doi:10.1161/CIRCRESAHA.117.311367. [PubMed: 29030344]

33. Montefiori LE, Sobreira DR, Sakabe NJ, Aneas I, Joslin AC, Hansen GT et al. A promoter interaction map for cardiovascular disease genetics. Elife. 2018;7. doi:10.7554/eLife.35788.

34. Hua N, Tjong H, Shin H, Gong K, Zhou XJ, Alber F. Producing genome structure populations with the dynamic and automated PGS software. Nat Protoc. 2018;13(5):915–26. doi:10.1038/nprot.2018.008. [PubMed: 29622804] • Software paper highlighting mathematical concepts underlying the authors' method to build 3D genome models from pairwise Hi-C contacts

35. Chapski DJ, Rosa-Garrido M, Hua N, Alber F, Vondriska TM. Spatial Principles of Chromatin Architecture Associated With Organ-Specific Gene Regulation. Front Cardiovasc Med. 2018;5:186. doi:10.3389/fcvm.2018.00186. [PubMed: 30697540]
36. Karbassi E, Rosa-Garrido M, Chapski DJ, Wu Y, Ren S, Wang Y et al. Direct visualization of cardiac transcription factories reveals regulatory principles of nuclear architecture during pathological remodeling. J Mol Cell Cardiol. 2019;128:198–211. doi:10.1016/j.yjmcc.2019.02.003. [PubMed: 30742811]
37. Bishop CM. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag; 2006.
38. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65–9. doi:10.1038/s41591-018-0268-3. [PubMed: 30617320] • Innovative application of machine learning to clinical data to predict arrhythmia
39. Nallamothu BK, Hill JA. Preprints and Cardiovascular Science: Prescient or Premature? Circ Cardiovasc Qual Outcomes. 2017;10(9). doi:10.1161/HCQ.0000000000000033.
40. Horvath S DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115. doi:10.1186/gb-2013-14-10-r115. [PubMed: 24138928]
41. Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. 2016;17(1):171. doi:10.1186/s13059-016-1030-0. [PubMed: 27511193]
42. Dawber TR, Meadors GF, Moore FE Jr. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health. 1951;41(3):279–81. doi:10.2105/ajph.41.3.279. [PubMed: 14819398]
43. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019;11(2):303–27. doi:10.18632/aging.101684. [PubMed: 30669119]
44. Dogan MV, Beach SRH, Simons RL, Lendasse A, Penaluna B, Philibert RA. Blood-Based Biomarkers for Predicting the Risk for Five-Year Incident Coronary Heart Disease in the Framingham Heart Study via Machine Learning. Genes (Basel). 2018;9(12). doi:10.3390/genes9120641.
45. Breiman L Random Forests. Machine Learning. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
46. Quaife-Ryan GA, Sim CB, Ziemann M, Kaspi A, Rafehi H, Ramialison M et al. Multicellular Transcriptional Analysis of Mammalian Heart Regeneration. Circulation. 2017;136(12):1123–39. doi:10.1161/CIRCULATIONAHA.117.028252. [PubMed: 28733351]
47. See K, Tan WLW, Lim EH, Tiang Z, Lee LT, Li PYQ et al. Single cardiomyocyte nuclear transcriptomes reveal a lincRNA-regulated de-differentiation and cell cycle stress-response in vivo. Nat Commun. 2017;8(1):225. doi:10.1038/s41467-017-00319-8. [PubMed: 28790305]
48. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. doi:10.1186/1471-2105-9-559. [PubMed: 19114008]
49. van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9.
50. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10(1):5416. doi:10.1038/s41467-019-13056-x. [PubMed: 31780648] • Useful technical guide on how to perform t-SNE on single-cell RNA-seq datasets
51. Gladka MM, Molenaar B, de Ruiter H, van der Elst S, Tsui H, Versteeg D et al. Single-Cell Sequencing of the Healthy and Diseased Heart Reveals Cytoskeleton-Associated Protein 4 as a New Modulator of Fibroblasts Activation. Circulation. 2018;138(2):166–80. doi:10.1161/CIRCULATIONAHA.117.030742. [PubMed: 29386203]
52. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018.
53. de Soysa TY, Ranade SS, Okawa S, Ravichandran S, Huang Y, Salunga HT et al. Single-cell analysis of cardiogenesis reveals basis for organ-level developmental defects. Nature. 2019;572(7767):120–4. doi:10.1038/s41586-019-1414-x. [PubMed: 31341279]

54. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381–6. doi:10.1038/nbt.2859. [PubMed: 24658644]
55. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049. doi:10.1038/ncomms14049. [PubMed: 28091601]
56. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20. doi:10.1038/nbt.4096. [PubMed: 29608179]
57. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888–902 e21. doi:10.1016/j.cell.2019.05.031. [PubMed: 31178118]
58. Stone NR, Gifford CA, Thomas R, Pratt KJB, Samse-Knapp K, Mohamed TMA et al. Context-Specific Transcription Factor Functions Regulate Epigenomic and Transcriptional Dynamics during Cardiac Reprogramming. Cell Stem Cell. 2019;25(1):87–102 e9. doi:10.1016/j.stem.2019.06.012. [PubMed: 31271750]
59. Martini E, Kunderfranco P, Peano C, Carullo P, Cremonesi M, Schorn T et al. Single-Cell Sequencing of Mouse Heart Immune Infiltrate in Pressure Overload-Driven Heart Failure Reveals Extent of Immune Activation. Circulation. 2019;140(25):2089–107. doi:10.1161/CIRCULATIONAHA.119.041694. [PubMed: 31661975]
60. Tucker NR, Chaffin M, Fleming SJ, Hall AW, Parsons VA, Bedi KC Jr. et al. Transcriptional and Cellular Diversity of the Human Heart. Circulation. 2020. doi:10.1161/CIRCULATIONAHA.119.045401.
61. Litvinukova M, Talavera-Lopez C, Maatz H, Reichart D, Worth CL, Lindberg EL et al. Cells of the adult human heart. Nature. 2020. doi:10.1038/s41586-020-2797-4.
62. Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM et al. A human cell atlas of fetal gene expression. Science. 2020;370(6518). doi:10.1126/science.aba7721.
63. Mangul S, Martin LS, Hoffmann A, Pellegrini M, Eskin E. Addressing the Digital Divide in Contemporary Biology: Lessons from Teaching UNIX. Trends Biotechnol. 2017;35(10):901–3. doi:10.1016/j.tibtech.2017.06.007. [PubMed: 28720283]
64. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. doi:10.1093/bioinformatics/btq033. [PubMed: 20110278]
65. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature. 2012;481(7381):389–93. doi:10.1038/nature10730. [PubMed: 22217937]
66. Wolff J, Bhardwaj V, Nothjunge S, Richard G, Renschler G, Gilsbach R et al. Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. Nucleic Acids Res. 2018;46(W1):W11–w6. doi:10.1093/nar/gky504. [PubMed: 29901812] • This software is particularly suited for the novice bioinformatician because it contains a graphical user interface
67. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259. doi:10.1186/s13059-015-0831-x. [PubMed: 26619908]
68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.• This well maintained Python library contains many of the tools needed to perform machine learning tasks

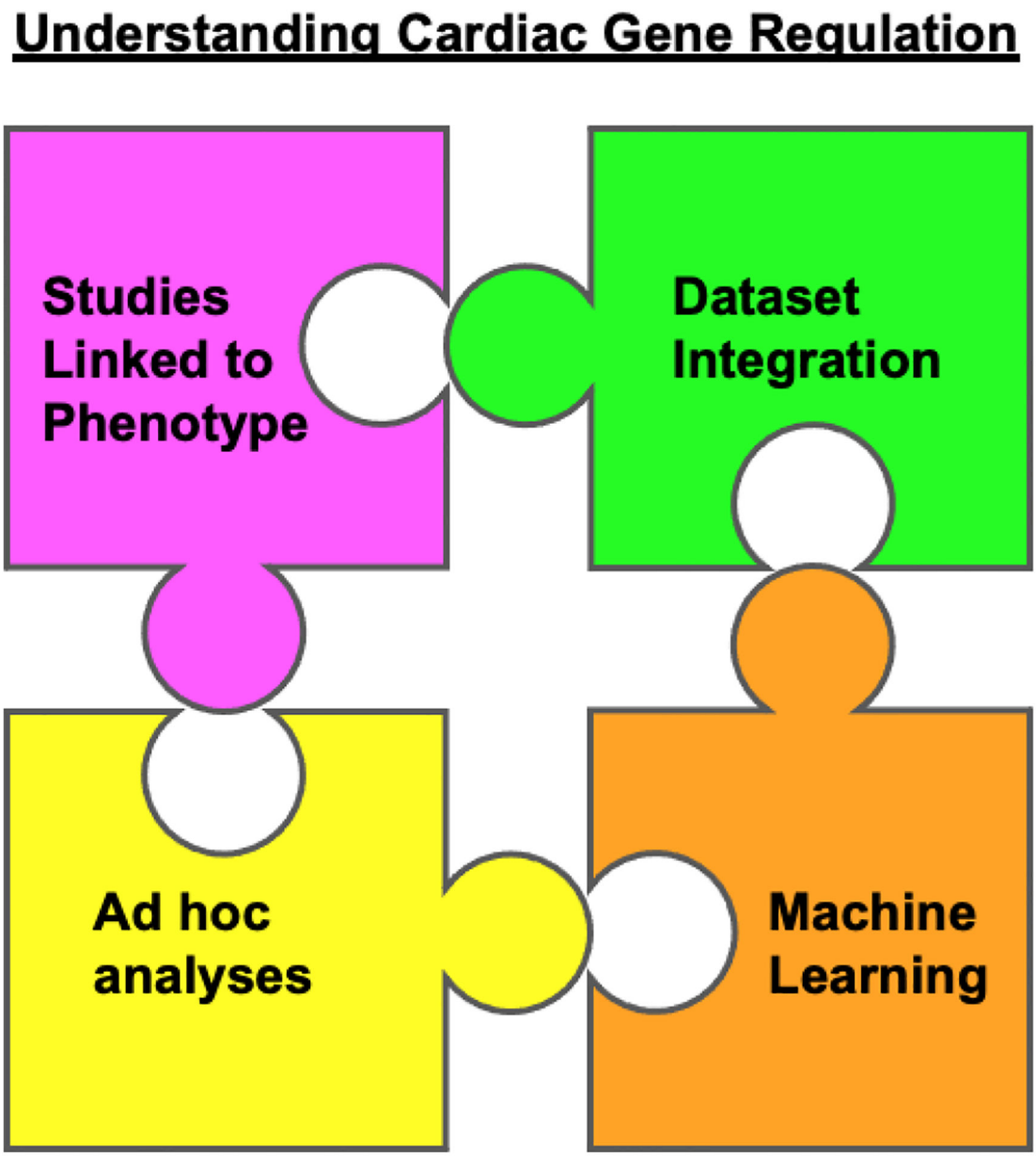

Figure 1. Strategies to examine gene regulation in the heart.

The utility of bioinformatics approaches to understand cardiac gene regulation comes from incorporation of data from: epigenomic measurements linked to phenotype (*pink*) to shed light on one aspect of the nucleus; data integration and modeling tools to paint a more comprehensive picture of distinct gene regulatory levels (*green*); incorporation of published data and non-genomics techniques (*yellow*); and machine learning approaches (*orange*) to understand which genomic features predict cardiac pathology.

Table 1.

Overview of common preprint servers in biomedical sciences

Server	Subfield	Website
arXiv	Physics, Math, Computer Science, Economics	arxiv.org
bioRxiv	Biology	biorxiv.org
medRxiv	Health Sciences	medrxiv.org
ChemRxiv	Chemistry	chemrxiv.org

Dissemination of scientific discoveries occurs at a fast pace. Preprint servers are increasingly popular because they offer an opportunity to share manuscripts before acceptance into peer reviewed journals. A caveat of preprints is their lack of peer review, and a layperson may not possess the skillset to evaluate them. Contrastingly, preprint servers provide new data to researchers to move fields forward without the delay of peer review. Corresponding authors individually decide whether to upload unpublished work to these servers.

Table 2.

Selected software packages for epigenomics research

Software	Utility	Reference
bedtools	Operations (e.g., overlap analysis) on genomic features	[64]
CellRanger	Single-cell RNA-seq tool maintained by 10X Genomics	[55]
Cufflinks	Classic RNA-seq tool for quantifying transcripts	[6]
DESeq2	Robust differential expression analysis software	[9]
DiffBind	Differential occupancy analysis for ChIP-seq, ATAC-seq	[65]
HiCExplorer	Galaxy-based web server for analyzing Hi-C data	[66]
HiC-Pro	Strong pipeline for analyzing Hi-C data	[67]
Monocle	Single-cell RNA-seq analysis package	[54]
PGS	3D genome modeling from Hi-C data	[34]
scikit-learn	Machine learning library for Python users	[68]
Seurat	Single-cell RNA-seq analysis package	[56, 57]