Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Applications of ENCODE data to systematic analyses via data integration

Yanding Zhao[1,2], Evelien Schaafsma[1,2] and Chao Cheng[1,2,3]

Abstract

Large-scale genomic data have been utilized to generate unprecedented biological findings and new hypotheses. To delineate functional elements in the human genome, the Encyclopedia of DNA Elements (ENCODE) project has generated an enormous amount of genomic data, yielding around 7000 data profiles in different cell and tissue types. In this article, we reviewed the systematic analyses that have integrated ENCODE data with other data sources to reveal new biological insights, ranging from human genome annotation to the identification of new candidate drugs. These analyses demonstrate the critical impact of ENCODE data on basic biology and translational research.

Addresses

[1] Department of Biomedical Data Science, The Geisel School of Medicine at Dartmouth College, One Medical Center Dr., Dartmouth-Hitchcock Medical Center, Lebanon, NH, 03756, United States

[2] Department of Molecular and Systems Biology, The Geisel School of Medicine at Dartmouth College, One Medical Center Dr., Dartmouth-Hitchcock Medical Center, Lebanon, NH, 03756, United States

[3] Norris Cotton Cancer Center, The Geisel School of Medicine at Dartmouth College, One Medical Center Dr., Dartmouth-Hitchcock Medical Center, Lebanon, NH, 03756, United States

Corresponding author: Cheng, Chao (chao.cheng@dartmouth.edu)

Current Opinion in Systems Biology 2018, ■:1–8

This review comes from a themed issue on **Big data acquisition and analysis**

Edited by **Mark Gerstein** and **Haiyuan Yu**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.coisb.2018.08.010

Introduction

The development of high-throughput technologies has generated enormous amounts of data which allow biologists to examine numerous biological hypotheses. In this review, we discuss how the Encyclopedia of DNA Elements (ENCODE) project has contributed to our understanding of many aspects of biology. The ENCODE project is an international collaborative project funded by the National Human Genome Institute (NHGRI). It aims to identify and characterize functional regions in the human genome by utilizing a variety of high-throughput approaches [1].

The pilot phase (2003–2007) of the ENCODE project was launched to explore 1% of the human genome to establish protocols for scaling up analyses to the entire genome [2–4]. During the pilot phase, a variety of experimental and computational methods were compared and refined for large-scale analyses. Then, the ENCODE project was expanded to the entire human genome during the production phase (2007–2012). In alignment with ENCODE, the modENCODE project [5] and the mouseENCODE project [6] were launched to systematically identify the DNA elements in the genomes of three model organisms including fly, worm and mouse. To date, ENCODE has generated abundant genomic data of different types as well as various tools and software. In total, 7694 profiles of different assay categories have been released to the public so far as summarized in Table 1 [7].

Based on these data, many important biological analyses have been performed, including genome annotation, chromatin state classification, and the identification of regulatory regions in the genome [2,3]. Moreover, ENCODE data has been integrated with other data sources such as cancer data to gain new insights into cancer development and drug discovery. Up to May, 2018, 2563 articles have been published by the ENCODE project and its community, and 6683 articles that cited the ENCODE landmark paper have been published (Figure 1) [2]. In this review, we focus on systematic analyses that have integrated ENCODE with other data sources to better understand complex biological processes and human diseases.

Improving human genome annotation

The Human Genome Project has completed the sequence of the human genome in 2003, however, at that time the annotation of the genome was far from accurate [8]. Based on ENCODE data, protein-coding and noncoding transcripts, long non-coding RNAs and pseudogenes have been carefully annotated through a combination of computational analyses, manual annotation, and experiment validation [9,10]. According to the refined annotation, it is now estimated that a total of 74.7% of the human genome is covered by primary

Table 1

Summary of ENCODE data [8]. The number of datasets for each assay category across different Tiers, cell types and tissues are shown until May, 2018 (ENCODE data portal: https://www.encodeproject.org). The Tiers refer to the ENCODE classification system of cell lines according to priority of performing the assays. Tier 1 cell lines were considered of the highest priority in ENCODE project and therefore the included cell lines (K562, GM12878 and H1-hESC) are presented individually. The assay category refers to the type of genomic features described by the assay. Q4

Assay category		DNA binding	Transcription	DNA-accessibility	RNA-binding	DNA-methylation	Replication timing	Genotyping	3D-structure	Proteomics
Assay name		ChIP-seq	shRNA RNA-seq, total RNA-seq, RNA microarray, small RNA-seq, RAMPAGE, polyA RNA-seq, CAGE, CRISPRi RNA-seq, siRNA RNA-seq, CRISPR RNA-seq, single cell RNA-seq, microRNA counts, microRNA-seq, potyA depleted RNA-seq, RNA-PFT	DNase-seq, ATAC-seq, genetic modification DNase-seq, FAIRE-seq, MNase-seq	eCLIP, RIP-seq, RIP-chip, iCLIP, Switchgear	DNAme array, RRBS, WGBS, MRE-seq, MeDIP-seq	Repli-seq. Repli-chip	Genotyping array, DNA-PET, genotyping HTS	ChIA-PET, Hi-C, 5C	MS−MS
Tier 1	K562	697	506	52	237	8	6	5	11	6
	GM12878	249	60	5	25	10	6	5	4	5
	H1-hESC	106	21	3	3	6	1	1	1	2
Tier 2		978	481	38	172	34	40	13	16	1
Tier 3		1065	400	255	6	252	61	86	40	0
other		926	382	190	0	116	35	32	34	0