Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Data resources for human functional genomics

Kristin G. Ardlie[1] and Roderic Guigó[2]

Abstract

Characterizing and interpreting the function of the millions of genetic variants in the human genome is a pressing need in medical and evolutionary genomics and is fundamental to understanding the molecular, cellular, and organismal consequences of genetic variation. Over recent years, several large-scale efforts have established genome-wide resources that describe and map multiple elements of genome function across many tissue and cell types. Additional population based studies have simultaneously mapped genome-wide regulatory variation across large numbers of individuals. We summarize these resources, review the data types they offer, and insights they provide into human functional variation. We also discuss the challenges and developments needed to integrate both existing and new resources into a detailed map of how genetic differences impact molecular phenotypes and ultimately human health.

Addresses

[1] Broad Institute of MIT and Harvard, Cambridge MA, 02142, USA

[2] Center for Genomic Regulation (CRG), Barcelona, Catalonia, Spain

Corresponding author: Ardlie, Kristin G

Current Opinion in Systems Biology 2017, ■:1–5

This review comes from a themed issue on **Genomics and epigenomics**

Edited by **Tuuli Lappalainen** and **Emmanouil Dermitzakis**

For a complete overview see the Issue and the Editorial

Available online xxx

http://dx.doi.org/10.1016/j.coisb.2016.12.019

One of the most pressing needs in medical and evolutionary genomes is characterizing and interpreting the function of the millions of genetic variants in the human genome, most of which are rare, and many of which are expected to have functional consequences constrained to specific tissue or cell types that may change temporally, or in response to environment. Since over 90% of the genetic variation that is associated to human disease resides in the noncoding portion of the genome, with presumed regulatory impact, it has become imperative to be able to systematically interrogate both the coding and the noncoding genome, to decipher the rules by which genes and gene networks are regulated, and to be able to predict the molecular, cellular, and ultimately organismal phenotypic consequences of genetic variation.

Functional genomics lies at the intersection of the quantitative dissection of genetic variation, and the molecular understanding of the mechanistic function of the genome [1], with a goal of understanding the dynamic properties of an organism from the cellular to the organismal level. Over the last several years, systematic genome wide assays, based mostly on massively parallel sequencing technologies, have been developed to describe and map multiple elements of genome function, ranging from gene transcription and translation to DNA methylation, chromatin accessibility and modifications, Transcription Factor binding, protein−protein interactions, and more. These large-scale efforts have produced concerted genome wide maps of these elements as well as initial insights on the impact of genetic variation on functional variation. Here we describe these key resources, their relationship to one another, their limitations and ongoing needs.

The ENCODE (ENCyclopedia Of DNA Elements) project was the first of the systematic large-scale functional genomics project. It was launched in 2003, after the completion of the Human Genome project, with a goal of identifying all of the elements that confer functionality in the human genome. The initial pilot phase of the project was aimed at identifying the most cost-effective genome analysis methods that were then employed in subsequent phases of the project [2]. In parallel, the modENCODE project [3,4] with goals similar to ENCODE was undertaken on the model organisms *Drosophila melanogaster* and *Caenorhabditis elegans*, as well as on *Mus musculus* (the mouse ENCODE project [5]). Currently the ENCODE Data Coordination Center (https://www.encodeproject.org/matrix) hosts the results of close to 12,000 assays. These are dominated by ChIPSeq monitoring Transcription Factor binding sites and histone modifications, RNAseq monitoring transcription, and with additional assays monitoring DNA accessibility, DNA methylation, RNA binding, replication timing, 3D Chromatin structure and proteomics. Most of the assays in the second phase of ENCODE were performed in immortalized cell lines, many of which were also genotyped. This included a concerted effort to perform as many overlapping assays as possible in two cell lines to enable integrative analyses across the multiple assays − GM12878, a lymphoblastoid cell line (LCL), and K562, a cell line from a chronic myelogenous leukemia (CML) and for which the transcriptome was also assayed in a number of subcellular compartments, including subnuclear ones. In recent ENCODE phases, additional sample types have been surveyed including bulk tissues, primary

cells, stem cells and *in vitro* differentiated cells, using single cell RNAseq analysis in some instances. ENCODE also spawned the GENCODE project [6] which aims to build the reference gene and transcript annotation of the human and mouse genomes. Of note, the ENCODE project demonstrated that the vast majority of the genome exhibits some form of biochemical activity (i.e. regions are transcribed, or bound by transcription factors, methylated, or host histone modifications, interact with other regions proximally or distally and so on). Despite the debate that ensued over equating biological function with biochemical activity [7–9] the resource has been widely impactful with over 1000 published studies of genome analysis and human disease utilizing the data to enhance functional inferences.

With aims closely related to those of the ENCODE project, the Roadmap Epigenomics Project was launched in 2007 by the NIH [10]. The project produced profiles of histone modification patterns, DNA accessibility, DNA methylation and RNA expression, producing 111 reference epigenomes for a range of primary tissues and cell types representative of all major lineages in the human body, including multiple brain, heart, muscle, gastrointestinal tract, adipose, skin and reproductive samples, as well as immune lineages, embryonic stem (ES) cells and induced pleuripotent stem (iPS) cells, as well as several differentiated lineages derived from ES cells. The Reference Epigenome Mapping Centers (REMCs) within the project generated a total of 2805 genome-wide data sets (http://www.roadmapepigenomics.org).

Following on the heels of the Roadmap Epigenome project, Blueprint was launched in 2011, as a European-based companion to the Roadmap, but focusing on distinct types of haematopoietic cells from healthy individuals and their malignant leukaemic counterparts [11]. To date, the project has produced more than 1500 blood-based genome wide maps of histone modifications, DNA accessibility, DNA methylation and RNA expression on a variety of cell types, including monocytes, granulocyte neutrophils, eosinophils, macrophages (M0, M1 and M2), naive CD4+ and naive CD8+ cells as well as several cell line samples (http://www.blueprint-epigenome.eu/). With the Roadmap Epigenomics and Blueprint projects as funding partners, and including ENCODE and other related projects, the International Human Genomics Consortium (IHEC, http://ihec-epigenomes.org/) was founded in 2010 to further the activities of these projects, and to coordinate overall global efforts in the field of epigenomics. A key goal of IHEC is to set quality standards and provide recommendations for epigenomic data generation and analysis.

Another long running functional data resource project is the FANTOM project, initiated by RIKEN in the year 2000 and funded by the Japanese Government. The initial aim of the project was to assign functional annotations to full-length mouse cDNAs, but FANTOM has evolved towards a broader functional genomics project with a specific focus on transcriptome analysis. FANTOM pioneered the high throughput use of Cap Analysis of Gene Expression (CAGE) analysis. CAGE is a tagging technology in which short sequence tags are extracted from the 5′ends of capped RNA molecules, and sequenced. It thus provides quantitative information on the usage of all transcription start sites (TSS), both known and novel, in the cellular condition assayed. It is uniquely able to measure the genome wide transcriptional activity of promoters, although conversely cannot provide information about exon and transcript abundances. During the fifth and current phase of the project (FANTOM5), CAGE analysis has been performed across a total of 975 individual, or pooled, human and 399 mouse samples, including a range of primary cells, cancer cell lines, and some primary tissues (http://fantom.gsc.riken.jp) [12], making FANTOM5 one of the most diverse transcriptomic surveys to date [13].

Combining both transcriptome and detailed protein analysis across different tissues and organs of the human body, the more recent Human Protein Atlas (HPA) [14,15] has produced a resource map of the human tissue proteome by integrating quantitative transcriptomics at the tissue and organ level, combined with tissue microarray-based immunohistochemistry (http://www.proteinatlas.org/) across a range of primary human tissues. The resource provides a detailed map of the spatial location, even to single cell level, of ~90% of all putative protein-coding genes across 32 different tissues and organs.

A major challenge to integrating these functional genomics data in human disease studies is that most of their multi-omic assays were conducted in cell lines instead of primary tissues, or where tissue/primary cell-based, represent relatively few, or different individuals and in some cases samples pooled across individuals. While cumulatively producing rich catalogs of functional elements across the genome, these studies have not captured the variation in genome function among individuals that would enable direct measurement of the genetic contribution (as assessed by disease association studies) to human phenotypic diversity at the cellular level.

Orthogonal to these approaches are the population based studies that map genome-wide regulatory variation by quantitative trait locus analysis. Expression quantitative trait locus (eQTL) analysis is now a standard approach to map regulatory genetic variants that are associated with changes in gene expression levels [16–18]. Recent eQTL studies, with large individual