Published in final edited form as:

Curr Opin Syst Biol. 2017 February ; 1: 75–79. doi:10.1016/j.coisb.2016.12.019.

Data Resources for Human Functional Genomics

Kristin G. Ardlie¹ and Roderic Guigó²

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

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

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]).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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 pluripotent 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 2,805 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 1,500 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 sample sizes, have demonstrated that regulatory variants are widespread in the genome, but these studies are still based on only one or a few readily accessible cells or tissues (LCLs, blood, with some skin and adipose tissue) [19–23]. The Genotype Tissue Expression (GTEx) project expands on these efforts and was specifically established to characterize the role of genetic variation on gene expression variation across a wide range of both primary tissues and a large number of individuals [24,25]. The project, which is still ongoing, has released data on over 7,000 transcriptomes from 449 individuals across 44 tissues (release phs000424.v6.p1). With completion of the project, in late 2017, the total number of transcriptomes is expected to be ~20,000 tissues from over 900 individuals, with

genome sequence data also available on all individuals [24,25]. The project's strengths lie in capturing both dimensions simultaneously, that is, in the large number of tissues collected from each individual, enabling analyses of gene expression differences across a wide variety of tissues, and the relatively large size of the donor population, enabling evaluation of the contribution of individual genetic variation [26–28].

To fully understanding how genetic variants functionally regulate gene expression, through changes in DNA methylation or transcription factor binding affinity etc, and ultimately to protein abundance requires additional molecular phenotypes, however, such as those produced by the functional genomics studies described above. Thus to expand upon GTEx, the enhancing GTEx (eGTEx), and the ENCODE GTEx (EN-TEX) projects were initiated specifically to enable integration between the longitudinal and vertical data resources by adding additional molecular assays, and analyses of intermediate phenotypes, across the same individuals and tissue samples. Planned assays include many of the ENCODE assays, such as DNA methylation, ChIP-sequencing, DNase I hypersensitivity, and others such as allelic and protein expression, with a goal towards building an even higher resolution reference of how genetic differences impact molecular, cellular, and organismal phenotypes. This will allow the same QTL analysis approaches to be applied to other quantitative phenotypes, such as chromatin state [29], methylation [30–32], and protein abundance [33,34] as well as enabling integration of the different QTL types to provide insight in to causal mechanisms from genetic variation through to gene regulation and function [35].

Among the limitations of all the current resources, insufficient population and demographic diversity is a key one. The largest resource in terms of samples, the GTEx project, sampled predominantly from a population of European ancestry with a smaller cohort of African ancestry, and with donor age range spanning 20–70 years (but with an older age bias). Little data is available from earlier childhood and young adult life stages, where age-dependent patterns of gene expression and regulation play a crucial role in development, and where reference data would be especially valuable for the study of pediatric diseases and many common diseases whose development begins early in life [36,37]. Among existing data sets, the Roadmap Epigenomics project generated expression data for a set of fetal tissue sites [10], but no resource spanning childhood through adulthood across multiple sample types has been generated to date. Similarly, the current resources are lacking in both geographic and population diversity, which is of particular importance to enable functional analyses and interpretation of the genetic diversity now well-described in GWAS and other population genetic studies, and to address the growing observations of environmental influences on the epigenetic landscape. The Geuvadis project included individuals from five populations: the CEPH, Finns, British, Toscani and Yoruba, but only LCLs obtained from those individuals have been studied [19]. A recent initiative, the Human Functional Genomics Project (HFGP), with a specific focus on identifying the relationship between genetic variation and immune function in disease [38,39], will enroll several hundred individuals, from different geographic regions, and spanning several disease cohorts. But it too is limited to adults, and will span only a few accessible tissues and cell types.

Another practical limitation of the existing resources is the lack of overlap in samples, assays, analyses and methodologies. This is beginning to be addressed, for example in GTEx

where many of the assays used in ENCODE will be applied to the same individuals and samples as part of the eGTEx and EN-TEX projects. To fully enable cross-study comparisons and integrative analyses of these functional genomics resources as they continue to grow, the implementation of ‘best practices’ for experimental design, data generation and analysis [1], along with mechanisms for easily sharing and disseminating these data will become paramount. Notably the use of compatible genome builds and annotations will be critical for integration of genomic and transcriptomic data with other molecular phenotypes. The development of cloud-based infrastructures will also enable far more rapid integrative analyses of data using common references, and common analytical methods.

The growth in functional genomics data resources that map gene and protein expression, and regulatory elements across a wide range of cells and tissues, and increasingly across individuals, is enabling detailed characterization of the functional effects of genetic variation. Valuable insights into basic biological and disease mechanisms have been gained from these studies, but much remains to be discovered. To fully characterize and interpret gene function and regulation in the human body will require ongoing and future studies to involving larger and more diverse populations and sample types, and to be performed also at the resolution of cell types within individuals.

References

1. Lappalainen T. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Research*. 2015; 25:1427–1431. [PubMed: 26430152]
2. ENCODE project consortium: An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
3. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*. 2010; 330:1771–1787.
4. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. modENCODE Consortium. Identification of functional elements and regulatory circuits by drosophila modENCODE. *Science*. 2010; 330:1787–1797. [PubMed: 21177974]
5. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*. 2014; 515:355–364. [PubMed: 25409824]
6. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Research*. 2012; 22:1760–1774. [PubMed: 22955987]
7. Eddy SR. The C-value paradox, junk DNA and ENCODE. *Curr Biol*. 2012; 22:R898–899. [PubMed: 23137679]
8. Doolittle FW. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci USA*. 2013; 110:5294–5300. [PubMed: 23479647]
9. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci USA*. 2014; 111:6131–6138. [PubMed: 24753594]
10. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015; 518:317–330. [PubMed: 25693563]

11. Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, Bock C, Boehm B, Campo E, Caricasole A, et al. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol.* 2012; 30:224–226. [PubMed: 22398613]
12. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, et al. FANTOM Consortium and the RIKEN PMI CLST (DGT). A promoter-level mammalian expression atlas. *Nature.* 2014; 507:462–470. [PubMed: 24670764]
13. de Hoon M, Shin JW, Carninci P. Paradigm shifts in genomics through the FANTOM projects. *Mammalian, Genome Springer US.* 2015; 26:391–402.
14. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpour S, Danielsson A, Edlund K, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics.* 2014; 13:397–406. [PubMed: 24309898]
15. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu Å, Sivertsson A, Kampf C, Sjöstedt E, Asplund A, et al. Tissue-based map of the human proteome. *Science.* 2015; 347:1260419. [PubMed: 25613900]
16. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. *Nature.* 2005; 437:1365–1369. [PubMed: 16251966]
17. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, et al. Population genomics of human gene expression. *Nat Genet.* 2007; 39:1217–1224. [PubMed: 17873874]
18. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet.* 2015; 16:197–212. [PubMed: 25707927]
19. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, González-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.* 2013; 501:506–511. [PubMed: 24037378]
20. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* 2011; 7:e1002003. [PubMed: 21304890]
21. Buil A, Brown AA, Lappalainen T, Viñuela A, Davies MN, Zheng H-F, Richards JB, Glass D, Small KS, Durbin R, et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet.* 2015; 47:88–91. [PubMed: 25436857]
22. Brown, AA., Buil, A., Viñuela, A., Lappalainen, T., Zheng, HF., Richards, JB., Small, KS., Spector, TD., Dermitzakis, ET., Durbin, R. *eLife*. Vol. 3. eLife Sciences Publications Ltd; 2014. Genetic interactions affecting human gene expression identified by variance association mapping; p. e01381-16.
23. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 2014; 24:14–24. [PubMed: 24092820]
24. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genetics.* 2013; 45:580–585. [PubMed: 23715323]
25. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science.* 2015; 348:648–660. [PubMed: 25954001]
26. Aguet F, Brown AA, Castel S, Davis JR, Mohammadi P, Segrè AV, Zappala Z, Abell NS, Fresard L, Gamazon ER, et al. Local genetic effects on gene expression across 44 human tissues. *BiorXiv.* 2016; doi: 10.1101/074450
27. Jo B, He Y, Strober BJ, Parsana P, Aguet F, Brown AA, Castel SE, Gamazon ER, Gewirtz A, Gliner G, et al. Distant regulatory effects of genetic variation in multiple human tissues. *BioRxiv.* 2016; doi: 10.1101/074419
28. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, Zappala Z, Strober BJ, Scott AJ, Ganna A, et al. The impact of rare variation on gene expression across tissues. *BioRxiv.* 2016; doi: 10.1101/074443

29. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*. 2012; 482:390–394. [PubMed: 22307276]
30. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet*. 2010; 6:e1000952. [PubMed: 20485568]
31. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol*. 2011; 12:R10. [PubMed: 21251332]
32. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife*. 2013; 2:e00523. [PubMed: 23755361]
33. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. Genomic variation. Impact of regulatory variation from RNA to protein. *Science*. 2015; 347:664–667. [PubMed: 25657249]
34. Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, Alsallakh B, Tilgner H, Araya CL, Tang H, et al. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res*. 2015; 25:1610–1621. [PubMed: 26297486]
35. Pai AA, Pritchard JK, Gilad Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet*. 2015; 8:e1004857.
36. Sekar A, Biala AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, Tooley K, Presumey J, Baum M, Van Doren V, et al. Schizophrenia risk from complex variation of complement component 4. *Nature*. 2016; 530:177–183. [PubMed: 26814963]
37. Anand A, Koller DL, Lawson WB, Gershon ES, Nurnberger JI. BiGS Collaborative. Genetic and childhood trauma interaction effect on age of onset in bipolar disorder: An exploratory analysis. *J Affect Disord*. 2015; 179:1–5. [PubMed: 25837715]
38. Netea MG, Joosten LA, Li Y, Kumar V, Oosting M, Smeekens S, Jaeger M, Ter Horst R, Schirmer M, Vlamakis H, et al. Understanding human immune function using the resources from the Human Functional Genomics Project. *Nature Med*. 2016; 22:831–833. [PubMed: 27490433]
39. Li Y, Oosting M, Deelen P, Ricaño-Ponce I, Smeekens S, Jaeger M, Matzaraki V, Swertz MA, Xavier RJ, Franke L, et al. Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. *Nature Med*. 2016; 22:952–960. [PubMed: 27376574]