ENCODE 3 Summary

March 6, 2015

The goal of the Encyclopedia of DNA Elements (ENCODE) project is to identify functional elements of the genome and thus providing a framework for interpreting the human genome in the context of human biology and disease. The ENCODE Consortium (Figure 1) consists of Data Production Centers responsible for efficiently generating specific data types, a Data Coordination Center (DCC) for data curation and public release, a Data Analysis Center (DAC) and an Analysis Working Group (AWG) to assist in and lead integrated analyses of the data, and computational analysis and technology development groups who have continuously developed and piloted new technologies and computational approaches. Together, the consortium has generated a product that is far greater than the sum of the parts. Key aspects of this are the development of high and uniform standards for data generation and data quality, uniform, cloud-based data processing pipelines that are transparent, reproducible, and available for use by users outside the consortium, the rapid and restriction-free release of verified datasets as soon as they are generated, and extensive educational and outreach efforts. Below we briefly highlight major accomplishments of the ENCODE project, followed by a more detailed overview of each activity of the ENCODE consortium.

Core accomplishments of the ENCODE project to date can be summarized as follows:

- Creation of vast, accessible catalogs of regulatory DNA, transcription factor occupancy and histone modification patterns, RNA binding protein occupancy, and RNA transcripts, as well as a standard curation of protein-coding and non-coding genes (GENCODE).
- Development and dissemination of standards and experimental methods for producing high-quality, reproducible data in a cost efficient manner from major assay types including ChIP-seq, RNA-seq, and DNase-seq.
- Development and dissemination of algorithms and software for analysis of major regulatory genomic data types, as well as tools and methods for integrating functional genomic data types.

Key biological insights provided by ENCODE:

- The biochemical features, structural and functional diversity, and evolution of regulatory DNA
- The systematic analysis and interpretation of non-coding disease- and trait-associated human variation, e.g., brought to light by GWAS studies
- The organization and architectural principles of transcriptional regulatory networks and circuits
- Relationships between regulatory DNA and higher-order chromatin features and interactions
- The organization, diversity, and nuclear compartmentalization of RNA transcription and its interplay with chromatin and regulatory DNA.

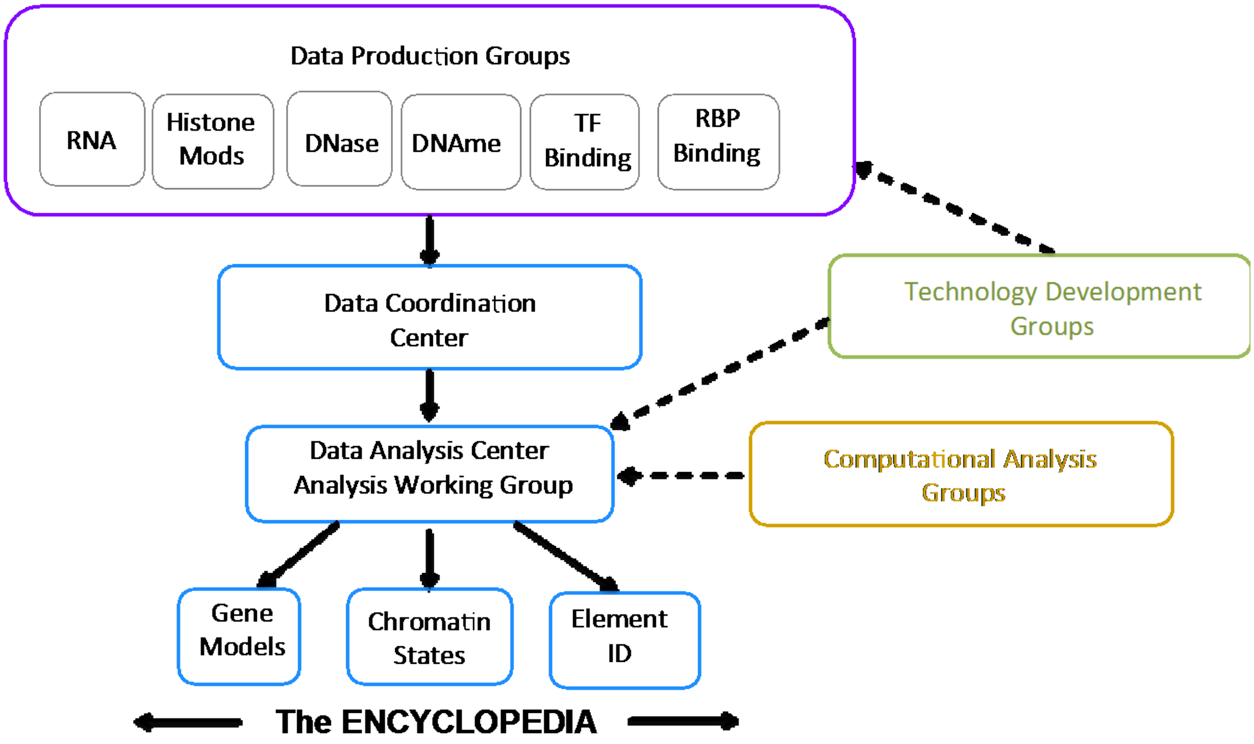

Figure 1. Overview of the organization of the ENCODE consortium.

General Impact

Just as the impact of the Human Genome Project took several years to be realized, the long-term impact of ENCODE has yet to fully mature. There are, however, many components of the project suggesting that ENCODE will have a long-lasting legacy. For example, the ENCODE consortium has a strong track record of publishing high impact papers that have been cited thousands of times each (e.g., the main consortium paper published in *Nature* in 2012 has already been cited 2,379 times according to Google Scholar). To date, ENCODE has published 520 papers, and modENCODE and mouseENCODE have published 162 and 26, respectively. More importantly, however, is the use of ENCODE data beyond the immediate members of the ENCODE consortium. To date, about 750 papers by authors without ENCODE funding have used ENCODE data in their studies of human disease, basic biology, or methods development. Key to the uptake of ENCODE data beyond the consortium is the fact that the datasets are rapidly released to the public, of high quality, are generated and processed in a consistent manner, and have few if any restrictions on community use. Community use should also increase now that the ENCODE data processing pipelines are shared through GitHub and the DNA Nexus cloud. Furthermore, ENCODE is closely working with related projects to increase standardization of data processing, metadata, and APIs across projects. This will facilitate integrated analysis of data between projects further enhancing impact. Accessibility to ENCODE data is easier than ever and available on the new ENCODE portal which allows users to search, download, and display raw data and processed data. To further enhance uptake, ENCODE provides tutorials on use of the resource and will offer a users meeting in the summer of 2015. The ENCODE portal also shares the data standards documents to enhance reproducibility and to allow others to adopt them if they wish. ENCODE has collaborated with many groups including the GENEVA project, CHARGE and eMERGE. A number of projects have been inspired by ENCODE; a zebrafish project (ZENCODE), an agricultural animals project (FAANG), and a psychiatric diseases project (PsychENCODE). Finally, in addition to the individual datasets, technologies, software tools, and publications, the main "product" of ENCODE will be the Encyclopedia of DNA Elements – an annotation of functional elements of the genome. The Encyclopedia will provide the framework for interpreting the human genome and how variation impacts human biology and disease.

Data Production

ENCODE has generated 4,294 genome-wide experiments (nearly all of which consist of two replicates) that have been publicly released or submitted to the DCC for imminent public release. ENCODE projects to complete an additional 6,221 replicated experiments by the end of the project for a total of 10,515, approximately 84% of which are human with the remainder from mouse.

ENCODE has employed a diverse collection of assays to interrogate protein-DNA interactions (TF ChIP-Seq), chromatin structure (histone modification ChIP, DNase-Seq and ATAC-Seq), DNA methylation (whole-genome shotgun bisulfite sequencing, RRBS, and arrays), protein-RNA interactions (iCLIP and RIP-Seq), and RNA transcripts (RNA-Seq for long, short and small RNAs, poly(A)+ and total) to name a few. Collectively, these assays have been used to profile over 500 distinct biological samples, 76% of which are human. Most samples analyzed from mouse are either tissues or primary cells. While presently about half of the completed human samples are from immortalized cell lines, it is projected that by the end of ENCODE3, most human samples will be from tissues or primary cells, in part through a collaboration between ENCODE and GTEx. Over 200 assays have also been conducted in stem cells or induced pluripotent stem cells and over 150 assays in differentiated stem cells or induced pluripotent stem cells. A small number of assays have been conducted in cells treated with various perturbagens (e.g., estradiol, ethanol, Sendai virus, tumor necrosis factor, etc.) though this is a largely unexplored area for ENCODE.

Among all biosamples, a small number have been deeply sampled with many assays, while a large number have been profiled with a small number of assays. ENCODE is an associate member of the International Human Epigenome Consortium (IHEC), and we project we will complete about 25 human and 50 mouse complete IHEC epigenomes (comprised of at least DNAme WGBS, mRNA-seq, and ChIP-Seq for H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K27ac, and H3K9me3 data). By the end of ENCODE3, it is projected that K562 and HepG2 cells will have been subjected to 1,745 and 1,495 assays, respectively. These assays include complete epigenomes, ChIP-Seq and iCLIP assays for hundreds of transcription factors and RNA binding proteins with DNA and RNA, knockdown experiments followed by RNA-Seq to assess the functional impact of the observed protein-DNA and protein-RNA interactions, as well as other assays providing information on chromatin accessibility, DNA replication timing and 3D chromatin interactions. Together, these incredibly deep and uniformly generated datasets will provide an unprecedented opportunity to model the regulatory networks in these cells. At the other end of the spectrum, approximately 20% of all biosamples have been subjected to only a single genome-wide assay, the majority of which are either RNA-Seq or DNase-Seq, some of which are very widely used by the community.

By the end of ENCODE3, it is projected that over 1,500 different human proteins (or modified proteins) and 69 mouse proteins, will have been assayed by at least one protein-nucleic interaction assay (ChIP-Seq, iCLIP, etc.) in at least one biological sample. Most experiments targeting individual proteins (28%) have focused on major histone modification marks (H3K4me3, H3K27me3, H3K36me3, H3K4me1, H3K27ac, H3K9me3, etc.) as well as the architectural TF CTCF, the large subunit of RNA polymerase II and the histone acetyltransferase p300. The majority of the remaining targets are either transcription factors or RNA binding proteins which have been studied in only a small number of biological samples. Although ENCODE has studied an extremely large number of proteins, hundreds of DNA and RNA binding proteins have not yet been studied.

An extremely important contribution of ENCODE is the generation, identification, and characterization of high-quality reagents used in the project. Using quality criteria that have set a standard in the field, over 500 human and 60 mouse antibodies have been validated for use in immunoprecipitation experiments and it is projected that several hundred additional antibodies will be validated by the end of ENCODE3. In addition, many cell lines expressing epitope tagged proteins have been generated and hundreds of shRNAs that efficiently deplete the target mRNAs have been validated.

ENCODE3 has begun to systematically validate the function of elements identified from the vast array of genome-wide datasets that have been generated. Specifically, several groups are predicting enhancer elements and testing their activity in transgenic mouse enhancer assays and high throughput reporter assays. Such efforts are critical to differentiate elements that have biochemical activity (e.g., can be identified in genome-wide assays) versus those that are biologically functional.

Data Coordination Center (DCC)

During ENCODE3, the Data Coordination Center (DCC), working with production labs, DAC and the AWG, defined a new metadata standard for describing high-throughput sequencing assays and computational analyses, engineered robust uniform processing pipelines that processed these data, and built a new portal for the ENCODE Consortium that allows the scientific community to access these metadata and data as well as serves as a hub for sharing and communication among Consortium members. The implementation of these 3 major deliverables at the ENCODE DCC focused on maintaining interoperability with other genomic resources, capturing data provenance to ensure reproducibility, and providing novel ways to access ENCODE metadata and data.

The new metadata standard defines a data model that can handle 40+ high-throughput assay types in fly, worm, mouse, and human, by capturing key experimental variables, such as biosamples and assay methods. Interoperability between genomic resources is maintained through the use of ontologies also employed by EBI ArrayExpress to capture the experimental variables. In addition, the DCC provides outreach to other projects such as the Nuclear Receptor Signaling Atlas (NURSA), Reference Epigenomics Mapping Consortium (REMC), International Human Epigenome Consortium (IHEC), and GTEx, among others, to promote the use of ontologies for capturing experimental variables for increased interoperability. The metadata standard also promotes data provenance and reproducibility in that the data model supports linking biological samples that originate from a single donor as well as supports the description of software used for a pipeline, parameters used in running that software, and the identification of files that were used as input or output. To provide novel methods to access ENCODE metadata, it is available as a data model in JSON-LD which supports semantic web queries.

ENCODE3 has engineered and distributed uniform processing pipelines in order to promote data provenance & reproducibility as well as allow interoperability between genomic resources. The pipelines have defined metadata that promotes data provenance & reproducibility. All data files, reference genome versions, software versions, and parameters used by pipeline are captured and available via the ENCODE Portal. Pipelines are publicly available so a diverse range of biomedical researchers have access to and can run the exact pipeline that are used to generate ENCODE results. ENCODE pipelines maintained and used by the DCC are freely available for public use at DNAnexus.com via web browser, or linux command line via the ENCODE DCC github, so that researchers can process their data with “the” ENCODE pipelines, or create modified versions of the pipelines to suit their needs. All ENCODE primary and processed data are available and distributed without charge via the Amazon Web Services (AWS). This allows traditional download but also provides access to the complete data warehouse from an account at AWS. Access to the pipelines and data via the cloud allows even small labs the ability to use the data or software without access to institutional compute clusters. Additionally, the DCC is providing outreach to IHEC, California Institute for Regenerative Medicine (CIRM), and bioinformatics core groups to encourage use of common processing pipelines.

A new ENCODE Portal (encodeproject.org) has been created to promote use of ENCODE data and results by the scientific community and serve as a hub for sharing and communication among Consortium members. The portal pages list the antibody characterization and data standards defined by the Consortium, providing transparency about the methods and standards supporting the assays and analysis performed by the Consortium. The portal also displays the relationship between donors and biosamples as well as graphical displays of pipelines, software used for pipelines, and files generated by the pipeline encourage data reuse. In addition, the Portal provides an integration point for other significant consortium data, for example the REMC and the Genomics of Gene Regulation (GGR) results and metadata integrated with ENCODE products. Development of the portal has included novel ways to access ENCODE data. The ENCODE Portal is based on a REST API which is an industry standard for interacting with websites and other databases. The development of an ENCODE REST API provides programmatic data submission by members of the Consortium as well as programmatic data retrieval for the scientific community. The portal provides innovative search features utilizing the structured metadata as well as the ability to search processed ENCODE data by a genomic coordinate (eg, a variant), a genomic region (eg, a coordinate range), or a gene name plus flanking region (eg, a gene name +/- 10 kb) in order to retrieve data files that contain ENCODE-defined elements in that region. The ENCODE Encyclopedia of Elements will also be integrated within this view of the data. ENCODE assays can be searched using faceted browsing to narrow down the list of assays interesting to the user. The Portal also includes alternatives to visualize ENCODE data, including visualization of files from any search or arbitrary set of files at the UCSC Genome Browser via a track hub and graphical summaries of the ENCODE encyclopedia from search results. This allows peak files from only a subset of ChIP-seq assays to be displayed instead of the whole set of transcription factor ChIP-seq assays. To promote usage of the data, results, pipelines and files are stored in the cloud. This provides transparent and open access to ENCODE results, achievements and deliverables allowing the data to be reused, analyses queried and integrated, software to be reused, and pipelines to be effortlessly applied to biomedical research.

Data Analysis Center (DAC) and Analysis Working Group (AWG)

ENCODE 3 maintains a Data Analysis Center (DAC), as a component of the ENCODE Data Coordination and Analysis Center (EDCAC). DAC members are tasked with coordinating and assisting in the integrative analysis of data produced by the ENCODE Consortium, developing pipelines for data processing, and working with the DCC component of the EDCAC to integrate them for automatic application on all datasets. In addition, DAC members have developed independent software tools for general use (Appendix 1). The DAC is responsible for coordinating the analyses required to generate the Encyclopedia of DNA elements - the main product of ENCODE.

The Analysis Working Group (AWG) is composed of individuals from several of the data production and analysis projects and was responsible for the integrated analysis of multiple datatypes. The AWG led many of the analyses that resulted in the high impact consortium papers.

Computational Analysis Groups

One major difference between ENCODE3 and the previous phases of the project was the inclusion of six groups who are funded to generate innovative computational tools and approaches to analyze ENCODE data. These projects have developed new statistical and computational approaches to reduce the complexity of ENCODE data, allow comparisons involving many ENCODE datasets at once, identify regulatory elements in the human genome, including in repetitive elements, to determine how regulatory elements work together, integrate ENCODE data with GWAS data, determine how changes in DNA sequence lead to changes in gene expression, and identify genetic differences that alter RNA processing. To date, the computational analysis groups have generated at least 25 software tools that are currently being used to analyze ENCODE data. The software tools used and developed by the ENCODE Consortium are listed in Appendix 1 and on the ENCODE portal (https://www.encodeproject.org/software).

Technology Development Groups

In 2012, 11 groups were funded to develop revolutionary technologies to help identify genomic elements that play a role in determining what genes are expressed and at what levels in different cells. These technology development areas were focused in three areas – the discovery of functional elements, the characterization of functional elements, and computational analyses. Together these projects led to the development of new technologies that facilitated the identification of branchpoints involved in pre-mRNA splicing, measuring mRNA degradation and splicing kinetics, improving the power of ChIP-Seq, high throughput assays to validate a variety of functional elements, and computational approaches to model cell-specific gene expression programs and chromatin structure. The technologies developed in these projects have not only had an important impact on ENCODE, but have also been broadly adapted by researchers outside the consortium. The ENCODE3 technology development groups have published 27 papers, and 141 throughout the entire project.

Appendix 1. Analysis tools generated by ENCODE

ACT: The aggregation and correlation toolbox (ACT) is an efficient, multifaceted toolbox for analyzing continuous signal and discrete region tracks from high-throughput genomic experiments, such as RNA-seq or ChIP-chip signal profiles from the ENCODE and modENCODE projects, or lists of single nucleotide polymorphisms from the 1000 genomes project. It is able to generate aggregate profiles of a given track around a set of specified anchor points, such as transcription start sites. It is also able to correlate related tracks and analyze them for saturation--i.e. how much of a certain feature is covered with each new succeeding experiment. The ACT site contains downloadable code in a variety of formats, interactive web servers (for use on small quantities of data), example datasets, documentation and a gallery of outputs. http://act.gersteinlab.org/ PMID: 21349863

AlleleSeq: A computational pipeline that is used to study allele-specific expression (ASE) and allele specific binding (ASB). The pipeline first constructs a diploid personal genome sequence, then maps RNA-seq and ChIP-seq functional genomic data onto this personal genome. Consequently, locations in which there are differences in the number of mapped reads between maternally- and paternally-derived sequences can be identified, thereby providing evidence for allele-specific events. http://alleleseq.gersteinlab.org/home.html PMID: 21811232

ASARP: Allele-Specific Alternative mRNA Processing; This is a method to identify SNPs that influence mRNA processing. An updated version of this software is released; https://github.com/cyruschan/ASARP PMID: 22467206 PMCID: PMC3401465

atSNP: a fast importance sampling method for evaluating binding affinity changing potential of SNPs. https://github.com/chandlerzuo/atSNP Manuscript in preparation.

BETA: integrating ChIP-seq binding and differential expression (i) to predict whether the factor has activating or repressive function; (ii) to infer the factor's target genes; and (iii) to identify the motif of the factor and its collaborators, which might modulate the factor's activating or repressive function. http://cistrome.org/BETA/ PMID: 24263090

CCM: Cooperative Chromatin Model (CCM) - predicts chromatin accessibility (DNase-seq data) from novel DNA sequence. Evaluates every base of the genome for its estimated importance for controlling accessibility (manuscript under review)

Census: Tool to estimate sequencing library complexity from test sequencing runs (released, manuscript in preparation)

cnvCSEM: Copy number variation (CNV) guided multi-read allocation. Software available from http://www.stat.wisc.edu/~qizhang/ PMCID: PMC4184254

curveHist: a functional curve testing approach for identifying differential histone modifications. Manuscript and software in preparation.

dPeak: a model-based tool for identifying closely located binding events from ChIP-seq and ChIP-exo data.https://github.com/dongjunchung/dpeak PMCID: PMC3798280

Fit-Hi-C: Assigning statistical confidence estimates to Hi-C data. PMCID: PMC4032863

FixSeq: Method to adjust read counts so they are not overdispersed. An alternative to simple de-duplication of reads. Improves the performance of many high-throughput analysis packages (released, https://bitbucket.org/thashim/fixseq) PMCID: PMC3945112

FusionSeq: FusionSeq may be used to identify fusion transcripts from paired-end RNA-sequencing. FusionSeq includes filters to remove spurious candidate fusions with artifacts, such as misalignment or random pairing of transcript fragments, and it ranks candidates according to several statistics. It also includes a module to identify exact sequences at breakpoint junctions. http://archive.gersteinlab.org/proj/rnaseq/fusionseq/ PMID: 20964841

GCAP: Data QC and analysis pipeline for DNase-seq analysis: https://github.com/qinqian/GCAP

GEM: High resolution ChIP seq event caller. (released, in use by DCC) http://sysbio.mit.edu/gem/ High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. Yuchun Guo, Shaun Mahony & David K Gifford, (2012) PLoS Computational Biology 8(8): e1002638. DOI: 10.1371/journal.pcbi.1002638 PMCID: PMC3415389

GIREMI: Genome-independent Identification of RNA Editing by Mutual Information; This is a new method to predict RNA editing sites using RNA-Seq data alone without the need of genome sequencing data. doi:10.1038/nmeth.3314 PMID: 25730491

GoShifter: Software for testing if a set of SNPs are enriched in particular functional annotations of the genome through peak shifting the functional annotations. https://www.broadinstitute.org/mpg/goshifter/ Preprint at http://biorxiv.org/content/early/2014/09/18/009258

GRIT: (http://grit-bio.org/), a tool for the integrative analysis of RNA-seq, CAGE, PAS-seq, RAMPAGE, and other RNA datatypes. PMID: 24633242.

GSC: Genome Structure Correction, (http://projecteuclid.org/euclid.aoas/1294167794), a tool for assessing when two or more features defined on genomes are more associated than expected by chance alone.

HAYSTACK: chromatin state variation and cell-type specific regulators. PMCID: PMC3903219

IDR: irreproducible Discovery Rate, (https://www.encodeproject.org/software/idr/), a tool for the assessment of the reproducibility of results in high-throughput studies, analygous to the False Discovery Rate.

iGRAS: intronic tag SNPs for Genetic Regulation of Alternative Splicing; This is a method to identify intronic SNPs that are involved in causing splicing alteration using RNA-Seq data obtained from cellular fractions (nucleus vs. cytosol). (manuscript under review)

ILM: Introspective Learning Machines, a tool for feature detection, classification, and regression. Includes Unconstrained Surface Mapping (**USM**). In production; publication pending.

IQseq: A tool for isoform quantification with RNA-seq data. Given isoform annotation and alignment of RNA-seq reads, it will use an EM algorithm to infer the most probable expression level for each isoform of a gene. http://archive.gersteinlab.org/proj/rnaseq/IQSeq/ PMID: 22238592

jMOSAiCS: Joint peak calling and inferring patterns of binding/modification across many ChIP-seq datasets. Zeng X (2013). jmosaics: Joint analysis of multiple ChIP-Seq data sets. R package version 1.6.0. http://www.bioconductor.org/packages/release/bioc/html/jmosaics.html

KMAC: De novo motif discovery method - Discover enriched motifs (in KSM and PWM representation) from a set of sequences (typically TF ChIP-seq data), or enriched in a set of sequences comparing to another set of sequences. (manuscript in preparation)

Loregic: a computational method integrating gene expression and regulatory network data to characterize the logical cooperativity of regulatory factors. Loregic uses all 16 possible two-input-one-output logic gates (e.g. AND or XOR) to describe triplets of two factors regulating a common target, and finds the gate that best matches each triplet's observed gene expression pattern across many conditions. Using human ENCODE ChIP-Seq and TCGA RNA-Seq data, we are able to demonstrate how Loregic characterizes complex circuits involving both proximally and distally regulating transcription factors (TFs) and also miRNAs. github.com/gersteinlab/loregic PMID: *(in press)*

Mango: A complete ChIA-PET data analysis pipeline that provides statistical confidence estimates for interactions and corrects for major sources of bias including differential peak enrichment and genomic proximity. (Manuscript under review)

MBASIC: a generative model for analyzing and grouping multiple loci based on ENCODE ChIP-seq data. These loci can be a collection of motif locations, a liberal set of SNPs, or peaks from an ENCODE or non-ENCODE experiment. (Manuscript under review)
Software: https://github.com/chandlerzuo/mbasic

MOSAiCS-HMM: Peak caller specific for Histone ChIP-seq data (builds on and extends our peak caller MOSAiCS). Software: https://github.com/dongjunchung/mosaics

MultiGPS: Multi-condition ChIP seq event caller that aligns events across conditions and permitting the detection of differential events. http://mahonylab.org/software/multigps/ PMCID: PMC3967921

MUSIC: An algorithm for identification of enriched regions at multiple scales in the read depth signals from ChIP-Seq experiments. MUSIC first filters the ChIP-Seq read-depth signal for systematic noise from non-uniform mappability, which fragments enriched regions. It then performs a multiscale decomposition, using median filtering, identifying enriched regions at multiple length scales. https://github.com/gersteinlab/MUSIC PMID: 25292436

Pastis: Inferring 3D structure from Hi-C data. PMCID: PMC4229903

PeakSeq: A tool for calling peaks corresponding to transcription factor binding sites from ChIP-Seq data scored against a matched control such as input DNA. PeakSeq employs a two-pass strategy in which putative binding sites are first identified in order to compensate for genomic variation in the 'mappability' of sequences, before a second pass filters out sites not significantly enriched relative to the normalized control, computing precise enrichments and significances. http://info.gersteinlab.org/PeakSeq PMID: 19122651

Perm-seq: a probabilistic read mapping method that can supervise multi-read allocation in TF ChIP -seq ChIP and related -seq experiments data.
Software: http://www.stat.wisc.edu/~keles/Software/perm-seq/ (Manuscript under review)

PIQ: Method to resolve protein-DNA binding from DNase-seq data. Can produce binding calls for hundreds of different factors from a single DNase-seq experiment http://piq.csail.mit.edu (released)

pRSEM: Prior-enhanced version of RNA-seq quantification method RSEM (RSEM is in use by DCC). Manuscript and software in preparation.

RABIT: Regression analysis with background integration. Fast feature selection and regression method integrating ENCODE ChIP-seq, TF motifs, RBP motifs and TCGA expression, CNV, and DNA methylome data to identify key TFs and RBPs that regulate gene expression changes in different tumors. Paper under revision at PNAS, website http://rabit.dfci.harvard.edu will be available by ENCODE 2015 consortium meeting date.

RASER: Reads Aligner for SNPs and Editing sites of RNA; This is a new read aligner customized for accurate mapping of reads harboring SNPs or RNA editing sites. (manuscript in preparation)

RSEQtools: A suite of tools that use Mapped Read Format (MRF) for the analysis of RNA-Seq experiments. MRF is a compact data format that enables anonymization of confidential sequence information while maintaining the ability to conduct subsequent functional genomics studies. RSEQtools provides a suite of modules that convert to/from MRF data and perform common tasks such as calculating gene expression values, generating signal tracks of mapped reads, and segmenting that signal into actively transcribed regions. http://archive.gersteinlab.org/proj/rnaseq/rseqtools/ PMID: 21134889

Segway: Performing semi-automated genome annotation on the basis of heterogeneous collections of genomic data, including histone modification, DNase sensitivity, TF binding, RNA expression, Hi-C, etc. PMCID: PMC3340533

SeqGL: Software to learn multiple sequence signals from ChIP-, DNase-, and ATAC-seq data. Manuscript under review. https://bitbucket.org/leslielab/seqgl/wiki/Home

spliceVAR: Method to identify regulatory networks (SNPs and proteins) that underlie the variation of alternative splicing events. (manuscript in preparation)

Sprout: ChIA-PET interaction analysis tool with improved detection capability (released)

Statmap: (http://statmap-bio.org/) a tool for aligning short reads to repetitive genomes. PMID: 21177961.

Sushi: An R/Bioconductor package that allows flexible integration of genomic visualizations into highly customizable, publication-ready, multi-panel figures from common genomic data formats including Browser Extensible Data (BED), bedGraph and Browser Extensible Data Paired-End (BEDPE). http://bioconductor.org/packages/release/bioc/html/Sushi.html PMID: 24903420

UES (Uncovering Enrichment through Simulation): Software for testing if a set of SNPs are enriched in particular functional annotations of the genome through selection of random sets of SNPs. Manuscript in preparation.

WASP: WASP is a software package for two related tasks: (1) correcting allelic bias in mapped sequencing reads and, (2) identifying molecular quantitative trait loci (QTLs) using next-generation sequencing data (e.g. gene expression QTLs or histone mark QTLs). The WASP mapper works with any read mapping pipeline that outputs BAM or SAM format. WASP identifies molecular QTLs using a statistical test that combines information about the total depth and allelic imbalance of mapped reads. WASP can call QTLs with very small sample sizes (as few as 10) compared to traditional QTL mapping approaches. https://www.encodeproject.org/software/wasp/ van de Geijn B, McVicker G, Gilad Y, Pritchard J.. WASP: allele-specific software for robust discovery of molecular quantitative trait loci bioRxiv. 2014 Nov 7; doi:10.1101/011221