- Margins: 0.5 inches all around
- Normal: Arial 11 pt
- Spacing: 12 pt exactly (except for in-line figures, then automatic 'single spacing')
- **Hyphenate**! (Menu: Page Layout, Page Setup, Hyphenation, Automatic)
- Legends: Arial 8.5 pt, Spacing exactly 8.4 pt. Hyphenated.
- Titles: Cambria 12 pt looks nice, but flexible.
- References: NIH Style EndNote. [1]

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

NIH instructions write: "Use an Arial, Helvetica, Palatino Linotype, or Georgia typeface, a black font color, and a font size of 11 points or larger. (A Symbol font may be used to insert Greek letters or special characters; the font size requirement still applies.). Type density, including characters and spaces, must be no more than 15 characters per inch. Type may be no more than six lines per inch (1 inch = 72 points, six lines/inch = 12 pt exact spacing). Use standard paper size (8 ½" x 11) . Use at least one-half inch margins (top, bottom, left, and right) for all pages. No information should appear in the margins.**Font: 11**

For references, please use: {FirstAuthor, Year, #PMID}
Old text from the DAC3 proposal are in grey.

RFA

Writing Assignments:
Mark

- Organize enhancer prediction challenge - 2pg, 1 figure (Aim3, response to AWG)
- Detection of enhancers - 1pg, 1 figure (Aim 2, Encyclopedia)
- Enhancer-gene linkages - 2pg, 1 figure (Aim 2, Encyclopedia, moved the network portion to Aim 1)
- Personal genome - 2pg, 1 figure (need more connection to ENTEx, Aim 1),
- ENCODE and Cancer (Disease) AWG subgroup - 1pg (Aim 3, AWG subgroup)
- Integrating encode w/ other consortia - gtex,ihec, roadmap, 1000G [1pg]: old text (Aim 1.6)
- Consortium Authorship Network Analysis - 1/2 pg, 1 figure (Aim 3 Consortium paper writing)
- Consortium paper writing: update this old text. (Aim 3, consortium paper writing)
- Get support letter from Mike Snyder, Brent Graveley (both DONE)
- Editing help from native English speaker

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

Zhiping:

- Include RBP pipelines in the next DAC 1/2 pages (Aim 3, pipelines)
- DNAme pipeline and QC (Aim 3, pipelines)
- Encyclopedia (Aim 2)
- Letters from Bing Ren (DONE), Mike Cherry, Rick Myers

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

Manolis:

- Applying ChromHMM to ENCODE3 data, 1 pages, 1 figure. Any future development on ChromHMM? (Aim 2, part of Encyclopedia)
- Write 1/2 page on co-chairing the AWG GWAS sub-group. This is the most active AWG subgroup. Can you also summarize the main results of this subgroup? (Aim 3, AWG subgroup)
- QC and utility. Aim 4. 4 pages, 2 figures.

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

- ChromImpute: Aim 1=main (insert into Section 1.1.1). Aim4=application.
- Support letter from Brad Bernstein, John Stam?
- Skip: HaploReg and its future developments. 1 page, 1 figure. GWAS specifically excluded.
- Skip: Please update this old text on increasing the utility of ENCODE data for the GWAS community. (Aim 3, AWG subgroup)

Manolis todo list:

- ChromImpute (Aim 1.1) - done.
- Human Mouse (Aim 1.4) - done
- QC etc (Aim 4)

Anshul

- Add a paragraph to describe your expertise (In Innovation section, currently page 12)
- ATAC-seq pipeline 1/2 - 1 page, 1 figure (please insert into Aim 3, pipelines)
- ChIP-seq pipeline (TF and histone mark), 2 pages, 1 figure. Please include summary of existing pipelines and plan for future developments. (Aim 3, pipelines)
- I believe that you co-chair the Regulations AWG sub-group. If so, can you write 1 paragraph about your effort of co-chairing and products of the sub-group. (Aim 3, AWG subgroup)
- ~~2 pages plus 1-2 figures on the cool stuff we discussed on the phone, how to visualize encyclopedia on-demand, and deep learning to quantize the fine-scale of regulatory sites. (Aim 2, Encyclopedia)~~

Roderic

ZW comment: tighten on what we have done (in particular the RNA-seq pipeline) and expand on what we plan to do. Novel RNA-seq data-types, such as long-read RNA-seq, single cell RNA-seq, ribosomal profiling, Gro-seq. We should say that we are ready to analyze a large variety of RNA-seq related data types should they be produced in ENCODE4. We should include some general approaches on analyzing these new datatypes so that the reviewers would believe that we have the expertise. Please edit directly in this document. Preferably done (at least the drafty version) by **Saturday 1pm EST**.

Let us Skype on Tuesday 11am EDT, March 15.

- RNA-seq pipeline. 2-3 pages, 1 figure. Please include summary of existing pipelines and plan for future developments. (Aim 3)
- Can you write 2 paragraphs on your contribution to the RNA-calls and working with the RNA group? (Aim 3)
- Do you want to include 1 page, 1 figure on chromatin vs. splicing? (Aim 1)
- Do you want to include 1 page, 1 figure on your splicing method, or other things that your group is working on and plan to work on? (Aim 1)
- Get support letter from Tom Gingeras

Noble (link to pdf)

- Hi-C pipeline, 1 page, 1 figure (Aim 3)
- Applying Segway to ENCODE3 data, 2 pages, 1 figure. Any future development on Segway? (Aim 2, part of Encyclopedia)
- Assay selection, 1 page, 1 figure. Please include future development. (Aim 4)
- Please write 2 paragraphs on your contribution to the 3D nucleome AWG sub group. Please describe the main products of the nucleome group. (Aim 3)

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

- Can you write 2 paragraphs on how we can integrate ENCODE data with data generated by the 4D nucleome consortium? The goal is to increase the utility of ENCODE data. (Aim 1, integration with other consortia)
- 2 paragraphs on co-chairing 4D nucleome with Bing Ren (Aim 3)

Shirley

- target genes of transcription factors 1 page, 1 figure (Aim 2, Encyclopedia)
- expand the coverage of ENCODE data annotation by taking advantage of public DNase and H3K27ac datasets. For example, we have determined that the combination of DNase and H3K27ac is the best approach for predicting enhancers. However there are many cell types that have only DNase but not H3K27ac, also vice versa. Can we use public data to supplement? Talk about the collection of DNase and H3K27ac data in Cistrome. Need to be able to match meta data. 1 page, 1 figure (Aim 2, Encyclopedia)
- Prediction of silencers? 1/2-1 page, figure? (Aim 2, Encyclopedia)
- CRISPR DOI 10.1186/s13059-015-0843-6, which can help the Characterization centers, 1 page 1 figure (Aim 3)

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

Rafa

- QC for RNA-seq, DNase, histone makrs, etc..., 1 pg, 1 figure (Aim 3 or Aim 4, pipelines and QC)
- Identification of DMRs, 1 page 1 figure (Aim 2, Encyclopedia)
- Normalization of gene expression matrix, DNase signal matrix, histone mark level matrix etc. across the entire panel of ENCODE cell types. 1-2 pages, 1 figure. (Aim 2 or Aim 4?)
- How do we guard the Consortium against batch effects? 1 page, 1 figure (Aim 1 or Aim 4?)

Author 3/20/2016 4:55 PM
Formatted: Indent: Hanging: 0.25", Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.75" + Indent at: 0.5"

Support letters:
Cover letter that precedes all the support letters
Mark: Mike Snyder, Brent Graveley
Manolis: John Stam?
Roderic: Tom Gingeras
Zhiping: Bing Ren, Rick Myers (emailed but did not respond), Brad Bernstein, Nadav Ahituv, Uwe Ohler, Job Dekker, Len Pennacchio, Ross Hardison, Ben Brown and Peter Bickel, Sunduz Keles, Mike Cherry, Steven Brenner

Author 3/20/2016 4:55 PM
Deleted: Ben Brown and Peter Bickle

DAC3 old letters

Third DAC Annual Progress Report

Aim 1:
Current: 7K words.
Propose cuts: we are aiming to cut it to at least 4.5K by Thu.

Aim 2:
Current: 9.4K words.
Propose cuts: Zhiping's group to cut it to 6.5K

Aim 3.1
Current: 270 words.
Propose: Leave as it is.

Aim 3.2
Current: 8K words.
Propose cuts: Zhiping's group to cut it to 4K

Aim 3.3
Current: 3.3K words.
Propose cuts: we could cut it to 2.25K after Thursday. Unless we hear otherwise by Wed. at 7pm, we are going to commence on cuts on aim 3.3

Aim 4
Current: 5.3K words.
Propose cuts: Manolis's group to cut it to 3.5K. Please coordinate this with Manolis.

Note that even these cuts seem harsh, this plan is still 3 - 4K more words than the 20K targets. We are working on the specific aims to 1 page. We will also do the SWAT duty on Sat. and/or Sun., and Anurag will be the contacting person for that. We also assume your group will deal with the figures and references of the grant.

Project Summary and Relevance

SPECIFIC AIMS

This proposal aims to continue the Data Analysis Center (DAC) for the ENCODE project, towards completing the catalog of functional elements in the human genome using high-throughput experiments and computational methods. The proposed DAC will respond to the Analysis Working Group (AWG) and help process, integrate, analyze, and interpret data from all groups in the ENCODE Consortium. By leveraging expertise spanning multiple data types, the DAC aims to fully utilize data from ENCODE as well as from external consortia. The analyses we propose will substantially augment the value of ENCODE. The proposed DAC members (Z Weng, M Gerstein, M Kellis, R Guigo, R Irizarry, S Liu, A Kundaje, and W Noble) are leaders in their respective subfields of computational genomics. They have extensive experience working together, many having worked together since Pilot ENCODE (>12 years ago). Our proposed activities are grouped into four aims:

Aim 1. Analyzing & integrating data and metadata from a broad range of functional genomics projects.

The DAC will perform integration of data generated by the Mapping and Characterization Centers and the computational groups of the ENCODE consortium, as well as other public data, with the goal of annotating different classes of functional elements. In particular, the DAC will carry out initial exploratory data analyses using state-of-the-art machine learning techniques to seek novel insights and correlations between classes of functional elements. These analyses will make use of various concepts in network analysis (e.g., hubs and motifs). To achieve this goal, the DAC will integrate data from large-scale genomic consortia (e.g., Roadmap, 1000 Genomes, GTEx, TCGA, and IHEC), normalizing, calibrating and harmonizing these with the ENCODE annotations. As a key aspect of this integration, the DAC will use human variation data to construct personal genomes as a platform for individualized annotation. Finally, the DAC will perform comparative analyses between human and mouse to understand the degree of conservation of regulatory mechanisms.

Aim 2. Serving as an informatics resource by supporting the activities of the ENCODE AWG.

The DAC will work with the AWG to prioritize analyses, administering AWG conference calls to receive requests for analysis and reports on progress. The DAC will help define shared computational guidelines and infrastructure for data processing, common analytic tasks, and data exchange. In particular, the DAC will coordinate with the ENCODE Data Collection Center (DCC) to establish robust uniform analysis pipelines. Building on the pipelines developed by the current DAC in collaboration with other ENCODE members, the DAC aims to refine these and complete the development of uniform analyses for all major ENCODE data types. The DAC has currently built pipelines for ChIP-seq, RNA-seq, and methylation sequencing, which will be refined. We plan to further implement pipelines for Hi-C and ATAC-seq. For analyses that require more extensive interaction, the DAC will bring together data producers and analysis experts in focused groups, subgroups of the overall AWG, and develop specific, context-focused annotations, such as the integration of ENCODE data with GWAS data. As novel biological insights emerge from these analyses, the DAC will facilitate writing manuscripts in coordination with the AWG.

Aim 3. Creating high-quality Encyclopedias of DNA elements in the human and mouse genomes.

Based on the integrative work in Aim 1 and the construction of analysis pipelines in Aim 2, the DAC will generate catalogs of functional elements in a cell-type specific manner, collectively called the Encyclopedia. The Encyclopedia comprises three levels. The ground level includes genomic regions that are enriched in various biochemical signals such as DNase accessibility, histone modifications, transcription factor (TF) occupancy, DNA methylation, RNA transcription, RNA-binding protein (RBP) occupancy, and chromatin-driven

Author 3/20/2016 4:55 PM
Deleted: data

Author 3/20/2016 4:55 PM
Deleted: in order

Author 3/20/2016 4:55 PM
Deleted: the

Author 3/20/2016 4:55 PM
Deleted: GTex

Author 3/20/2016 4:55 PM
Deleted: iHEC

Author 3/20/2016 4:55 PM
Deleted: individual-specific

Author 3/20/2016 4:55 PM
Deleted: report

Author 3/20/2016 4:55 PM
Deleted: analysis

Author 3/20/2016 4:55 PM
Deleted: analysis

Author 3/20/2016 4:55 PM
Deleted: annotation (e.g., to best integrate

Author 3/20/2016 4:55 PM
Deleted:).

topologically associating domains (TADs). By leveraging the work of the Characterization Centers, the ground-level catalog will be integrated to predict the locations of functional elements such as enhancers, insulators, and silencers, forming the second level of the Encyclopedia. The third level of the catalog includes higher-order interactions among the predicted elements—e.g., enhancer-target interactions and system-wide regulatory networks in a specific cell type. To make the Encyclopedia accessible to a broad range of biomedical researchers, we will develop and implement approaches to present it in an intuitive and detail-on-demand manner.

Aim 4. Assessing quality and utility of the ENCODE data and providing feedback to the Consortium.

The DAC will develop computational methods to assess the quality and utility of ENCODE data in a systematic and unbiased way. We will work with the consortium to standardize data type specific metrics of dataset quality. We will further develop measures for assessing experimental quality based on systematic signal imputation. We will develop metrics and methods for evaluating the progress and completeness of the entire ENCODE corpus, towards the goal of covering all active functional elements in all cellular states. We will develop and apply methods for evaluating the per-nucleotide information content of each data type based on genomic coverage, resolution, and reproducibility. We will evaluate the predictive utility of an assay (or cell type) based on its ability to predict data from another assay (or cell type). We will then combine utility scores to rank datasets and prioritize the experiments that will be of greatest overall utility. Specifically in the context of disease, we will provide a ranking based on predictive ability for disease-associated genetic variants (e.g., from GWAS).

Author 3/20/2016 4:55 PM
Deleted: taking advantage of

Author 3/20/2016 4:55 PM
Deleted: -

Author 3/20/2016 4:55 PM
Deleted:

Author 3/20/2016 4:55 PM
Deleted: experiment

SIGNIFICANCE

The ENCODE Project is one of a number of ambitious projects building on the foundation of the Human Genome Project. The goal of ENCODE is to apply high-throughput, cost-efficient approaches to generate a comprehensive catalog of functional elements in the human genome, including transcribed regions, chromatin features, transcriptional control regions, and post-transcriptional control regions. The impact of this project is tremendous, broadly affecting biomedical research and personalized medicine, because functional genomic elements are the basis of all biological processes. As a coordinated effort, ENCODE members prioritize common biological samples, enforce quality standards, and implement rapid data release policies. The value of ENCODE data is greatly enhanced by efforts to increase the breadth, depth, quality, and utility of the data. Although the individual datasets produced by the ENCODE project are highly effective in the study of any individual region, their true potential is achieved by integrative genome-wide analyses {TheENCODEProjectConsortium:2004db}{ENCODEProjectConsortium:2007fu}{ENCODEProjectConsortium:2012gc}.

We propose an ENCODE Data Analysis Center (EDAC, DAC in short) to support, facilitate, and enhance integrative analyses of the ENCODE Consortium data on human and mouse. We will work closely with Consortium members to identify and prioritize integrative analyses that should be carried out, identify the best groups and methods to accomplish them, coordinate all necessary data transformations, and undertake these analyses with the other Consortium members. Our ultimate goal is to ensure a successful final product of high-quality annotation in human and mouse, called the ENCODE Encyclopedia, and gain new insights into the biology and genetic regulation of animal genomes.

We envision that the DAC will minimally perform the following roles, organized into four aims of the proposal in direct response to the RFA: (Aim 1). Analyzing & integrating data and metadata from a broad range of functional genomics projects. (Aim 2). Serving as an informatics resource by supporting the activities of the ENCODE AWG. (Aim 3). Creating high quality Encyclopedias of DNA elements in the human and mouse genomes. (Aim 4) Assessing quality & utility of the ENCODE data & providing feedback to the Consortium. To achieve these four aims, the proposed DAC will work closely with members of the Consortium and in particular two entities within it: firstly, the Analysis Working Group (AWG), consisting of all Principal Investigators (PIs) of the production centers, PIs of functional characterization centers, informatics PIs, and personnel from each of the groups; and, secondly, the Data Coordination Center (DCC), responsible for all data and metadata submission, data formatting and uniform processing, and data sharing with the larger scientific community.

The four aims and the interactions needed to achieve them are greatly facilitated by the composition of the proposed DAC. The DAC members' expertise spans human and mouse. We have experts in specific functional elements like promoters, enhancers, silencers, protein-coding and non-coding genes, micro-RNAs, sequence motifs, splicing, RNA binding proteins, and 3'-untranslated regions, as well as experts in analyses and various machine learning techniques such as principal component analysis, clustering, support vector machines, hidden Markov models, dynamic Bayesian networks, and deep learning. DAC members are collaborating directly with every production group in ENCODE and thus will have intrinsic knowledge of the experimental intricacies of each data type. The team has the unique experts to take on any aspect of the integrative analysis, but is also highly integrative as all members have longstanding experience of working together in ENCODE and related consortia. We have established a strong leadership and organizational structure for the DAC, as described in the PI Leadership Plan. A three-person committee composed of Weng, Gerstein, and Kellis will jointly make decisions on all important matters. The three labs are located in physical proximity (~2 hours' drive), in the same time zone, and thus can collaborate extensively and respond quickly to any analysis needs, which is particularly important during the final pushes of putting together the Encyclopedia and

Author 3/20/2016 4:55 PM
Deleted: strategize

Author 3/20/2016 4:55 PM
Deleted: analyses[1-3].

Author 3/20/2016 4:55 PM
Deleted: gene

Author 3/20/2016 4:55 PM
Deleted: ,

Author 3/20/2016 4:55 PM
Deleted: 3´-

Author 3/20/2016 4:55 PM
Deleted: ,

Author 3/20/2016 4:55 PM
Deleted: hours

Consortium papers. See support letters from the PIs of several current production groups and the DCC of ENCODE: Drs. Mike Snyder, Brad Bernstein, Brent Graveley, Tom Gingeras, Bing Ren, and Mike Cherry.

INNOVATION

This project requires a different type of innovation from traditional R01s. The RFA requires that we follow the directions set by the AWG, although we can contribute to setting the directions. Thus, we are fundamentally limited in being innovative in taking direction. Nonetheless, we are highly innovative in putting together a multifaceted team to meet the enormous challenge of facilitating and performing the integrative analysis activities in ENCODE, a large and complex consortium. At present, our team consists of eight highly talented investigators with expertise covering broad biological areas: transcriptional regulation (Weng, Kellis, Gerstein, Guigo, and Liu), epigenetics (Kellis, Weng, Gerstein, Kundaje, and Liu), evolution (Kellis), genomics and proteomics (Noble, Gerstein, Weng, and Guigo), regulatory RNA (Kellis and Weng), and biophysics (Gerstein and Weng). The team also has a variety of expertise in computational biology, e.g., machine learning (Noble, Kellis, and Kundaje), biostatistics (Liu and Irizarry), networks (Gerstein and Weng), and gene annotation (Guigo and Gerstein).

We are confident that the assembled team can produce innovative science in a consortium framework. This is manifested by the many Consortium publications that the team members have led and participated in. In particular, the team members have been highly innovative in building and applying state-of-the-art methods, with examples described in **Aims 1-4**, and a long history of producing computational tools that are widely used by the broader community. Importantly, most team members have a rich experience in various consortia: ENCODE (Weng, Gerstein, Kellis, Kundaje, Guigo, Noble, Irizarry, and Liu), modENCODE (Gerstein, Kundaje, Kellis, and Liu), the Epigenome Roadmap (Kundaje and Kellis), the 4D Nucleome Consortium (Noble), the 1000 Genomes Project (Gerstein), the PsychENCODE Project (Gerstein and Weng), the Extracellular RNA Communication Consortium (Gerstein), DOE KBase (Gerstein), and the 29 Mammals Project and the 12 Flies Project (Kellis). In summary, we have assembled a team with the right expertise, that has worked in large collaborative consortia and has delivered in that environment. The details of the team members are provided as follows:

Zhiping Weng (U. Mass. Medical School) Professor Zhiping Weng has an engineering background, and has worked for the last decade on biological problems ranging from genomic to protein-protein interaction analysis. She has participated in the ENCODE project since its inception in 2003, leading a technology development project in the pilot phase of ENCODE (2003-2007), a pilot project during the scale-up phase of ENCODE (2007-2011), and the DAC for ENCODE3 (2011-). She has participated in the integrative analysis in ENCODE since 2003, co-chaired the transcription regulation analysis group in the pilot phase of ENCODE, and was a member of the DAC in ENCODE2. She currently heads the ENCODE DAC and co-chairs the AWG. She has worked on transcription factor binding site detection, integrating histone modification signals to predict gene expression, and integrating chromatin features to predict enhancers and their target genes.

Manolis Kellis (MIT and the Broad Institute) Manolis Kellis is a professor of Computer Science and an Institute Member of the Broad Institute and has a background in computational genomics with 10+ years experience in comparative genomics, epigenomics, regulatory genomics, and disease genetics. He has led or co-led a large number of large-scale genomic studies, including the comparative analysis of 29 mammals, the integrative analysis of ENCODE chromatin datasets, the fly modENCODE integrative analysis, the comparative analysis of 12 *Drosophila* species, the comparative analysis of eight Candida genomes, and the first comparative studies of four yeast species.

Mark Gerstein (Yale U.) Professor Gerstein has been an integral part of the ENCODE Project since its inception, and within the project he has assumed a number of leadership roles. For instance, in ENCODE2 he

Author 3/20/2016 4:55 PM
Deleted: Page Break

Zhiping Weng 3/20/2016 9:26 AM
Comment [1]: ENCODE2 and ENCODE3 throughout.

co-directed the Networks/Elements Group, which resulted in his co-leading one of the ENCODE high-profile companions{Gerstein:2012fq}, and in ENCODE3, he co-lead the ENCODE & Cancer subgroup. He was co-chair of the AWG in modENCODE and in this capacity led the worm integrative paper{Gerstein:2010bu} and the comparative transcriptome paper{Gerstein:2015jr}. He also led and participated in a number of sub-analyses resulting in >15 companion papers in various phases of ENCODE, particularly those focusing on pseudogenes, variant interpretation, ncRNAs, TF binding sites and networks{Gerstein:2007bz}{Trinklein:2007io}{Rozowsky:2007ky}{Zhang:2007jm}{Zheng:2007gd}{Washietl:2007hk}{Euskirchen:2007ho}{Lu:2011jp}{Cheng:2011ir} {Negre:2011gg} {Gerstein:2012ck}{Yip:2012cd}{Cheng:2012cg}{Pei:2012ji}{Yan:2014jh}{Boyle:2015bq}.

He is also a member of a number of other genomics consortia, including the 1000 Genomes, PsychENCODE, exRNA, CMG, PCAWG and DOE Kbase and in these groups he has worked on connecting the ENCODE annotations with other genomic resources (eg {Khurana:2013em}, {Fu:2014jc}).

Roderic Guigo (Center for Genomic Regulation; Universitat Pompeu Fabra) Professor Guigo has been active in the field of computational genomics for more than 25 years. He has developed widely used tools for gene finding and annotation, such as GENEID and SGP, as well as benchmark metrics and datasets to evaluate the accuracy of gene prediction methods. He has participated in many genome projects, often in leadership positions. He has been involved in the ENCODE Project since the pilot phase, in which he lead the GENCODE efforts to delineate the gene and transcript reference annotation of the human genome. The bulk of his current research focuses on the mechanisms underlying RNA production (transcription) and post-processing (splicing). He also participates in other large scale functional genomics projects such as GTEx and BluePrint.

Rafael Irizarry (Dana Farber Cancer Institute; Harvard U.) Professor Irizarry has over 15 years of experience developing methods for high-throughput genomics data. His dedication to producing tools that are useful to biologists and the wider research community is evidenced by the popularity of the methods he has developed, such as RMA, fRMA, GCRMA, CRLMM, CHARM, CQN and bumphunter. He has made these tools freely available through open source code. His expertise in dealing with bias, systematic errors, batch effects, and other unwanted variation in biological data will be extremely valuable for improving the quality of ENCODE data via statistical methods.

Anshul Kundaje (Stanford U.) Prof. Kundaje's lab develops computational approaches to decipher the context-specific regulatory architecture of the human genome and the genetic and molecular basis of complex traits and disease by integrating diverse types of large-scale functional genomic and genetic data. Dr. Kundaje was trained as a computer scientist with expertise in data mining and machine learning coupled with extensive experience with large-scale integrative analysis of functional genomic data. He was the lead computational analyst of ENCODE2 and the Roadmap Epigenomics Project. He is currently an active member of the ENCODE3 DAC and co-leads the Regulation Subgroup. Dr. Kundaje has developed machine learning methods for deciphering comprehensive maps of cell-type-specific regulatory elements; deconvolving sequence, structural and functional heterogeneity of elements; learning predictive regulatory network models and interpreting the regulatory impact of natural and disease-associated non-coding genetic variation. He has developed state-of-the-art quality control measures and automated data processing pipelines for diverse functional genomic data types. Dr. Kundaje also has significant experience developing software infrastructure and web portals for rapid, cloud-based data mining and visualization of large-scale regulatory genomics data.

X. Shirley Liu (Dana-Farber Cancer Institute; Harvard T.H.Chan School of Public Health) Professor Liu is a computational biologist with expertise in algorithm development and integrative modeling of gene regulation. She has been a member of the mod/ENCODE consortia and the director of the Center for Functional Cancer Epigenetics at the Dana-Farber Cancer Institute. Her group developed a number of widely used algorithms for

Author 3/20/2016 4:55 PM
Deleted: companions[8],

Author 3/20/2016 4:55 PM
Deleted: paper[9]

Author 3/20/2016 4:55 PM
Deleted: [[PMID: 25164755]].

Author 3/20/2016 4:55 PM
Deleted: networks[10-19]. [[Add these PMIDs 22972285 22950945 22955978 22951037 25249401 25164757]]

Author 3/20/2016 4:55 PM
Formatted: Space After: 0 pt, No widow/orphan control, Tabs: 0.56", Left

Author 3/20/2016 4:55 PM
Formatted: Not Highlight

Author 3/20/2016 4:55 PM
Deleted: PMIDs 24092746 & 25273974).

Author 3/20/2016 4:55 PM
Deleted: and University

Author 3/20/2016 4:55 PM
Deleted: biologist

Author 3/20/2016 4:55 PM
Deleted:

Author 3/20/2016 4:55 PM
Deleted: on

transcription factor (TF) motif finding, ChIP-chip/seq, MNase/DNase-seq, and recently on CRISPR screen data analysis. Through integrating genome-wide transcription factor binding, chromatin dynamics, gene expression profiles, genetic and chemical screens, they model the specificity and function of transcription factors, chromatin regulators and lncRNAs in tumor development, progression, drug response and resistance. She has been at the forefront of identifying and correcting biases in next-generation sequencing and epigenomics data, a key component of big data integration.

William Noble (U. Washington) Professor Noble has a background in computer science and has applied a series of machine learning and statistical techniques to biological problems. He is one of the creators of the widely used MEME Suite of motif-based analysis tools, and he has led the use of support vector machines, wavelet analyses and dynamic Bayesian networks in genomic and proteomic analysis. Working in the Department of Genome Sciences at the University of Washington, he interacts with many high-throughput experimental groups and continues to have strong links to the machine learning community. He has been involved in the ENCODE Consortium since its inception in 2003.

A combined multi-site DAC with multiple investigators in geographical separation offers challenges for team cohesion, but we are prepared to meet these challenges and discuss them in the section titled "Risk Assessment and Leadership" in the Multi-PI Leadership Plan. In fact, the geographic distribution of the members of the DAC is ideally suited for a consortium that is itself expected to be geographically distributed, and has been geographically distributed in the previous three phases of the ENCODE Project. Importantly, we have a strong track record of working together over the past thirteen years of analysis during the ENCODE Project, modENCODE Project, and other large-scale consortia. Overall, the difficulties of coordination are offset by the benefits of the many different environments and complementary expertise provided by the participating groups.

APPROACH

Aim 1. Analyzing and integrating data and metadata from a broad range of experimental and computational functional genomics projects.

1.1 Exploratory analyses of diverse types of ENCODE data

1.1.1 Dataset completion by imputation for histone marks, DNA accessibility, RNA-seq, and DNA methylation.

The ENCODE project has balanced the priorities of employing many types of experimental assays and performing these assays in as many different biological contexts as possible. The solution has been to conduct a small set of assays (e.g., mapping histone marks by ChIP-seq and chromatin accessibility by DNase-seq) in a large variety of cell types, and a much larger set of assays (e.g., mapping transcription factors or TFs by ChIP-seq and RNA-binding proteins or RBPs by eCLIP-seq) in a few cell types. This leaves many types of experiments that have not been conducted in all biological contexts, reducing the ability to carry out integrative analyses that require complete data matrices, and reducing the utility of ENCODE datasets for cell types and/or assays that were not profiled experimentally. To remedy this situation, we will develop and apply methods for imputation of missing datasets by exploiting the correlational structure of functional genomic data across cell types and across assays (**Fig. 1.1.1a**). In our preliminary studies, we showed that a regression tree approach, ChromImpute{Ernst 2015ep}, recovers missing datasets for diverse cell types and assays (**Fig. 1.1.1b**), giving us confidence that we can complete missing datasets systematically.

Author 3/20/2016 4:55 PM
Deleted: A key component of big data integration is data normalization and bias correction, she

Author 3/20/2016 4:55 PM
Deleted: .

Author 3/20/2016 4:55 PM
Deleted: might potentially prevent the DAC from working as a cohesive

Author 3/20/2016 4:55 PM
Deleted: . We

Author 3/20/2016 4:55 PM
Deleted: going

Author 3/20/2016 4:55 PM
Deleted: and mitigate the risk

Author 3/20/2016 4:55 PM
Deleted: Section

Zhiping Weng 3/20/2016 9:54 AM
Comment [2]: need some verbiage to tie together the four sub-sections.

Author 3/20/2016 4:55 PM
Deleted: context

Author 3/20/2016 4:55 PM
Deleted: correlation

Author 3/20/2016 4:55 PM
Deleted: , 2015 #PMID:25690853

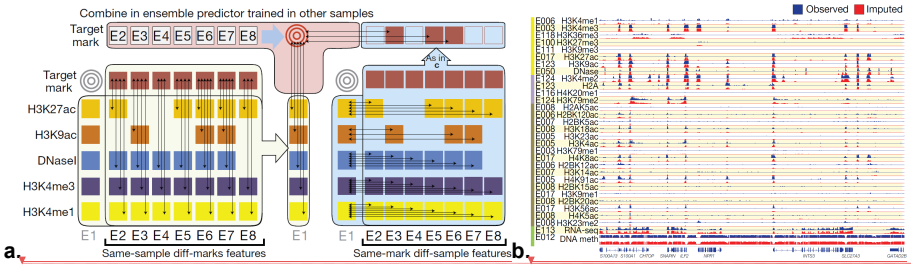


Figure 1.1.1: Imputation of missing datasets. **a. Ensemble strategy for signal track imputation** using features that exploit correlations between marks in the same sample (left) and correlations between samples for a given mark (right). We assume that no information is available for the target mark in the target sample (gray targets). Thus, we learn relationships between marks (left side) in other samples (column of E1 sample is not used) and learn relationships between samples (right side) using other marks from which we then compute same-mark features. The ensemble predictor that combines features across marks and across samples is learned only in other samples (top), and the marks in the target sample are used only during the actual application of the trained ensemble predictors to compute the imputed signals. **b. Imputed data are a close match to observed datasets.** Visualization of a randomly selected 200-kb region illustrates high-resolution concordance between observed (blue) and imputed (red) signal tracks. Imputed tracks are generated at 1-bp resolution for DNA methylation and 25-bp resolution for all other marks and trained without using the observed track.

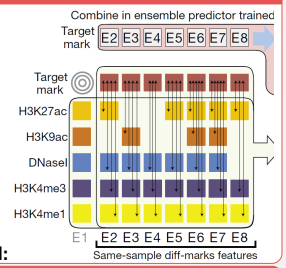
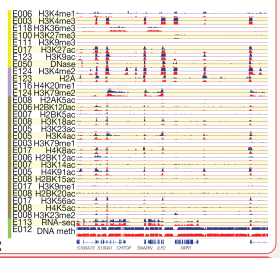
1.1.2. Prediction of TF binding sites in a new cell type

One area where predictive models may help infer information that we do not determine directly is the prediction of TF binding sites in cell types for which there are no ChIP-seq data for the TF but there are data on chromatin accessibility (DNase-seq) or histone modifications. Several DAC members have strong expertise in this area. We developed a method to predict TF binding sites by integrating histone modification data and TF motif information and applied it to modENCODE data on worm{Gerstein:2010bu} and also in yeast{Cheng:2011jz}. This model uses chromatin features (histone modifications, DNase hypersensitivity etc.) to infer the local accessibility of DNA regions, and then searches inferred accessible regions for TF binding motifs to predict binding sites. Similarly, we developed models to predict TF binding sites using DNAse-seq data. This model, LR-DNase, uses logistic regression to combine several features captured by DNase-seq data (e.g., DNase reads, footprints in DNase read coverage, and strand bias in DNase cleavage) with the DNA sequence and TF ChIP-seq data from a separate cell type, to predict TF binding sites in a cell type of interest. LR-DNase outperforms state-of-the-art methods including CENTIPEDE, PIQ, and BinDNase. Application of these models will enable inference of TF binding profiles in the hundreds of cell types where DNAse-seq and histone modification ChIP-seq are being conducted.

We discovered that chromatin dynamics coupled with TF motif discovery could be used to infer functional TF binding and gene expression changes by taking into account the characteristics of the TFs{He:2010cf}. For example, estrogen receptor binding sites already show DNase hypersensitivity prior to estrogen activation, while androgen receptor binding sites only show DNase hypersensitivity upon activation. We will systematically identify distinct chromatin dynamics relative to differing TF binding, differentiate signatures between transcriptional activators versus repressors, and model signatures associated with slow versus fast transcriptional responses upon perturbation.

1.1.3 Application of generalized machine learning methods to ENCODE data

The genome-wide experimental assays produced by ENCODE provide opportunities for data-driven research to complement traditional hypothesis-driven investigations, but also present unique challenges. We

Author 3/20/2016 4:55 PM
Deleted:

Author 3/20/2016 4:55 PM
Deleted:

Author 3/20/2016 4:55 PM
Deleted: don't
Author 3/20/2016 4:55 PM
Deleted: is
Author 3/20/2016 4:55 PM
Deleted: worm[9]
Author 3/20/2016 4:55 PM
Deleted: yeast[62].
Author 3/20/2016 4:55 PM
Deleted: Their
Author 3/20/2016 4:55 PM
Deleted: CENTEPEDE
Author 3/20/2016 4:55 PM
Deleted: TFs[66].
Author 3/20/2016 4:55 PM
Deleted: hypersensitivity
Author 3/20/2016 4:55 PM
Deleted: different

will implement a range of unsupervised machine learning methods to search for novel connections between elements of the ENCODE data, focusing on methods that scale to large data sizes and that are designed for the diverse statistical properties of data generated by different assays.

The premier method for integrating heterogeneous data is to use probability models, which enable inclusion of prior knowledge either through prior distributions or through the inherent structure of the chosen model. Applications of these models in ENCODE include hidden Markov models and their generalized form, dynamic Bayesian networks, for the semi-automated genome annotation (SAGA) methods ChromHMM and Segway. In addition to extending these SAGA methods (**Section 3.2.3**), we will, as appropriate, explore the use of unsupervised techniques such as Bayesian model-based clustering {Fraley:2002bg}, Gaussian process models (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.86.3414) {Rasmussen:2006vv}, probabilistic topic models {Blei:2012dk} and other learning procedures for graphical models (M. I. Jordan, *Learning in Graphical Models* 1996) {Jordan:1998us}.

Neural networks provide an increasingly popular alternative framework for analyzing massive datasets. Such networks use a series of non-linear transformations to project data into higher- or lower-dimensional spaces, allowing for learning in arbitrary classification, regression, or structured output problems. So-called "deep" neural networks, which contain many layers and typically require massive training datasets, have recently revolutionized a wide variety of fields, ranging from natural language processing and image analysis {LeCun:2015dt} to out-playing the world champion of the game "Go" {Silver:2016hl}. Deep neural networks have also been applied recently to genomics, including predicting the effects of non-coding variants {Heim:1989tf}, identification of replication domains {Liu:2015iq}, population genetic inference {Sheehan:2015us}, and predicting DNase hypersensitivity {Kelley:2015va}. Neural networks handle heterogeneity in a data-driven way by leveraging very large datasets to learn appropriate data projections that weight different data sources in a non-linear fashion. Unsupervised training of deep neural networks can be accomplished via the use of "auto-encoder" architectures, in which the same data is provided as both input and output to the model. Intermediate layers of the network can often be used for data visualization or compression, or to provide a unified input format to downstream analyses. Another class of unsupervised neural network, the self-organizing map, was used successfully in ENCODE2 to explore the latent state space of functional elements in the genome {Mortazavi:2013ks}, and we envision revisiting these maps using a deep architecture. Overall, for any unsupervised analysis of ENCODE data, deep neural networks will be an important component of the proposed DAC's analytical toolbox.

Kernel methods represent a third class of methods for learning from heterogeneous data. These techniques use a class of mathematical functions, known as kernel functions, to map data from its native format (the "input space") to a (typically) higher-dimensional "feature space." Kernel methods have been used very successfully to solve a huge variety of problems in computational biology {Scholkopf:2004ul}. Furthermore, kernels provide a modular framework for representing heterogeneous data, because separate kernel functions can be defined for each data type {Lanckriet:2004ha}. Importantly, many "classical" analysis methods can be adapted to operate using kernel representations, including supervised methods like Fisher's linear discriminant {Scholkopft:1999tz} as well as unsupervised methods like principal components analysis (PCA) {Scholkopf:2005jm} and kernel canonical correlation analysis (CCA) (http://www.ics.uci.edu/~welling/classnotes/papers_class/kCCA.pdf, {Ay:2014kh}).

We will also apply dimensionality reduction and manifold learning methods to glean structure from large, heterogeneous datasets. Dimensionality reduction methods are widely used for visualization, and include aforementioned techniques such as PCA and deep auto-encoders, as well as a variety of methods that project data non-linearly to a low-dimensional manifold. Recently, new dimensionality reduction methods such as t-SNE {VanderMaaten:2008tm} have proven useful for visualizing, for example, single-cell RNA-seq data

Author 3/20/2016 4:55 PM
Deleted: a …robability model. Thes ... [1]

Author 3/20/2016 4:55 PM
Formatted: Highlight

Author 3/20/2016 4:55 PM
Deleted:).

Author 3/20/2016 4:55 PM
Deleted: (PMID:26017442)…LeCu ... [2]

Author 3/20/2016 4:55 PM
Deleted: "…." Kernel methods have ... [3]

Author 3/20/2016 4:55 PM
Deleted: (Laurens van der Maaten, Geoffrey Hinton; JMLR; 9(Nov):2579--2605, 2008)

{Amir:2013dc}. DDRTree is a similar technique that imposes a tree-like prior on the learned projection, enabling visualization of cancer datasets {Mao:2015wp} or developmental trajectories from single-cell RNA-seq data (http://rpackages.ianhowson.com/cran/DDRTree/).

Finally, we will, as appropriate, apply standard clustering methods to help find structure in ENCODE data. Such methods include hierarchical clustering, k-means and k-medoids clustering, as well as a variety of biclustering methods on data defined in two dimensions (e.g., across cell types and assay types).

1.1.4 Inter-relating chromatin & splicing

We will also study the role of chromatin structure and modifications in the regulation of RNA processing, particularly splicing. Using integrative analysis of RNA-seq, ChIP-seq of histone modifications and of RNA polymerase, and ChIA-PET data produced during ENCODE2, we found significant positive association between H3K4me3, H3K9ac, and H3K27ac, on the one hand—histone marks characteristic of active promoters—, and exon inclusion, on the other hand, in a small but well-defined class of exons, representing approximately 4% of all regulated exons. These exons are systematically maintained at comparatively low levels of inclusion across cell types, but their inclusion is significantly enhanced in particular cell types when in physical proximity to active promoters (Figure 2, {Curado:2015ep}) .

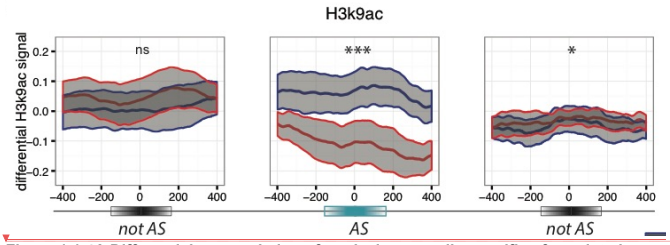


Figure 1.1.4A Differential accumulation of marks is generally specific of regulated exons. Differential ChIP-seq signals (average and standard error of the mean) for H3K9ac are represented for "more included" exons (blue) and "less included" exons (red) in pairwise comparisons between cell lines. The differential signal is computed on a 800 bp window around the regulated exon (AS) and the flanking not regulated (notAS) upstream (left) and downstream (right) exons. Significance levels are indicated by * (0.05>p>0.01), ** (0.01>p>0.001), *** (0.001>p) and ns (p>0.05). From {Curado:2015ep}

In order to elucidate how chromatin state effects and modulates RNA processing, we have built a series of random forest models to predict exon inclusion based on chromatin marks, chromatin state, and RNA binding protein binding assays (eCLIP) in the K562 cell line. The highest predictive power was achieved when using RBP binding data and histone modifications as predictors. We also found signatures from RNA Pol-II ChIP that affect choices between *cis*- and *trans*-splicing and post-transcriptional exon duplication and rearrangement. Higher RNA Pol-II density after cassette exons, likely indicating accumulation of transcripts, correlates with increased chances of local *trans*-splicing over *cis*-splicing (**Figure 1.1.4B**). Meanwhile, the density of Pol-II is also significantly lower downstream of *trans*-splicing junctions that produce circular RNA products, compared to those that lead to exon shuffling (**Figure 1.1.4B**).

Author 3/20/2016 4:55 PM
Deleted: (PMID: 23685480).

Author 3/20/2016 4:55 PM
Deleted: (Qi Mao, Li Wang, Steve Goodison, and Yijun Sun. Dimensionality Reduction via Graph Structure Learning. The 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'15), Hilton, Sydney, Australia, August 19, 2015)

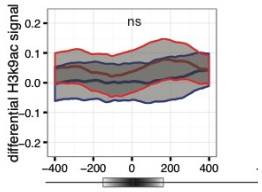
Zhiping Weng 3/20/2016 10:02 AM
Comment [3]: Cut this entire section? It does not have any planned activities for ENCODE4

Author 3/20/2016 4:55 PM
Deleted: Chromatin & Splicing

Author 3/20/2016 4:55 PM
Deleted: ,

Author 3/20/2016 4:55 PM
Deleted: [

Author 3/20/2016 4:55 PM
Deleted: J., 2015 PMID: 26498677])

Author 3/20/2016 4:55 PM


Deleted:

Author 3/20/2016 4:55 PM
Deleted: [

Author 3/20/2016 4:55 PM
Deleted: J., 2015, PMID: 26498677]

Author 3/20/2016 4:55 PM
Deleted:

Author 3/20/2016 4:55 PM
Deleted: to

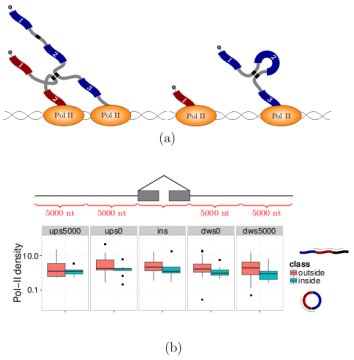
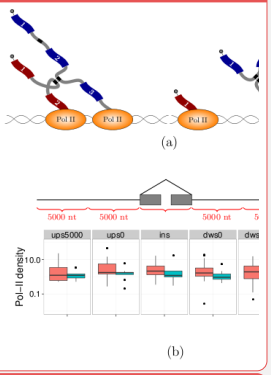


Figure 1.1.4.B Two regimes of RNA Pol-II lead to different splicing outcomes. Accumulation of Pol-II leads to collision and promotes trans-splicing. **C.** The density of RNA Pol-II (y-axis) is indeed higher downstream (x-axis) of cassette exons that have evidence of trans-splicing (green), compared to cassette exons that are spliced in cis- with the formation of circular products (red).

Author 3/20/2016 4:55 PM



Deleted:

Author 3/20/2016 4:55 PM

Deleted: pol…ol-II leads to collisior ... [4]

1.2 Analysis of biological network structure & dynamics.

We have developed a number of approaches for constructing and studying biological networks that can be applied to analyze ENCODE4 datasets. We integrated multiple genomic datasets to construct gene regulatory networks consisting of various regulatory factors including transcription factors and micro-RNAs and their target genes {Gerstein:2012fq}{Boyle:2015bq}{Cheng:2011dx}. For constructed gene regulatory networks, we developed methods to construct and analyze human and model organism gene regulatory networks {Yan:2010fu}{Gerstein:2010bu}{Cheng:2011dx}{Negre:2011gg}{Gerstein:2012fq} using ENCODE and modENCODE datasets. We also analyzed hierarchical structures of gene regulatory networks and found that hierarchy rather than centrality ("hubiness") better reflects the importance of regulators {Gerstein:2012fq}{Yu:2006jg}{Bhardwaj:2010fr}{Bhardwaj:2010jj}{Bhardwaj:2010em}. We also developed a novel and general purpose algorithm to determine and measure the hierarchical structure of any type of gene regulatory network {Cheng:2015dl}. In addition, we integrated regulatory networks with gene expression to uncover different types of functional modules {Luscombe:2004ei}{Cheng:2009ks}{Yu:2003jd}{Qian:2003wh}. We also introduced several software tools for network analysis including Topnet {Yu:2004cv}, tYNA {Yip:2006kv}, and PubNet{Douglas:2005ky}.

Author 3/20/2016 4:55 PM

Deleted: \cite{22955619, 25164757, 22125477}….Gerstein:2012fq}{Boyl ... [5]

In addition to the global features of regulatory networks such as hierarchy, we also analyzed their local topologies such as network motifs. For example, we analyzed feed-forward loops (FFLs), a common regulatory network motif in human, worm and fly gene regulatory networks {Gerstein:2012fq}{Gerstein:2010bu}{Boyle:2015bq}. The motif analyses will enable characterization of key regulatory mechanisms in each species and the comparison between species will enable us to observe how these mechanisms evolve.

Author 3/20/2016 4:55 PM

Deleted: the…regulatory networks ... [6]

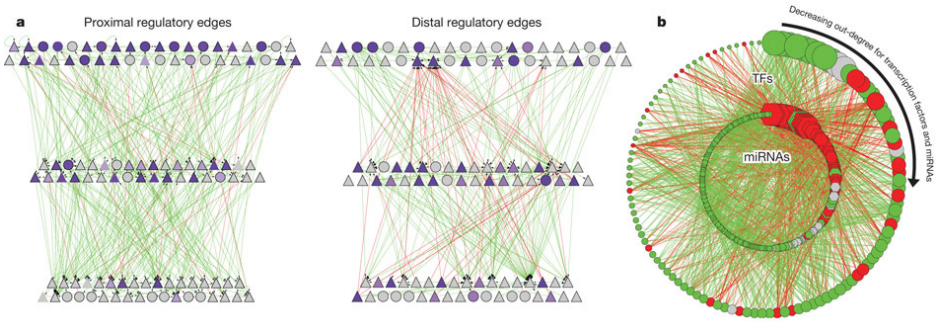
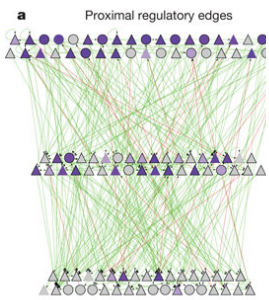


Figure Hierarchy of the human regulatory network derived from ENCODE data
a, Close-up representation of the transcription factor hierarchy. Nodes depict transcription factors. TFSSs are triangles, and non-TFSSs are circles. Left: proximal-edge hierarchy with downward pointing edges coloured in green and upward pointing edges coloured in red. The nodes are shaded according to their out-degree in the full network (as described in Table 1). Right: factors placed in the same proximal hierarchy but now with edges corresponding to distal regulation coloured green and red, and nodes re-coloured according to out-degree in the distal network. The distal edges do not follow the proximal-edge hierarchy. **b**, Close-up view of transcription-factor–miRNA regulation. The outer circle contains 119 transcription factors, whereas the inner circle contains miRNAs. Red edges correspond to miRNAs regulating transcription factors; green edges show transcription factors regulating miRNAs. Transcription factors and miRNAs each are arranged by their out-degree, beginning at the top (12:00) and decreasing in a clockwise order. Node sizes are proportional to out-degree. For transcription factors, the out-degree is as described in Table 1; for miRNAs, it is according to the out-degree in this network. Red nodes are enriched for miRNA–transcription factor edges and green nodes are enriched for transcription factor–miRNA edges. Grey nodes have a balanced number of edges (within ±1).

In this aim, we will study the structure and dynamics of inferred regulatory networks and relate these to other cellular networks. Our core networks will primarily consist of mixed TF/miRNA directed regulatory networks discovered in Aim 3.5, but will also be populated by undirected networks, constructed using information regarding protein interaction and biological-relatedness, such as those utilized or discovered in Aim 3.2. Figure 9 illustrates the TF-miRNA regulatory network generated with previous ENCODE data {Gerstein:2012fq}. Extending our previous work, we will use graph algorithms to discover clusters of highly connected genes within these networks, network motif algorithms to discover recurrent patterns of connectivity, and specifically search for recurrent regulatory feedback and feed-forward subgraphs.

Furthermore, we can try to identify gene regulatory mechanisms such as cooperative logic between multiple regulatory factors in the regulatory networks. We have recently developed a new computational method, Loregic to characterize the gene regulatory logic in complex systems {Wang:2015hn}. We already used this method to identify regulatory cooperative logic among TFs binding to promoters and TFs binding to distal regions like enhancers and miRNAs in leukemia by integrating ENCODE and TCGA data.

Author 3/20/2016 4:55 PM


Deleted:

Zhiping Weng 3/20/2016 11:51 PM
Comment [4]: What is this?

Zhiping Weng 3/20/2016 10:07 AM
Comment [5]: What is this?

Anurag Sethi 3/20/2016 11:51 PM
Comment [6]: Sequence-specific TFs

Author 3/20/2016 4:55 PM
Deleted: ones

Author 3/20/2016 4:55 PM
Deleted: the

Author 3/20/2016 4:55 PM
Deleted: factor

Author 3/20/2016 4:55 PM
Deleted: order

Author 3/20/2016 4:55 PM
Deleted: director

Zhiping Weng 3/20/2016 10:09 AM
Comment [7]: inchoate text.

Author 3/20/2016 4:55 PM
Deleted: \cite{22955619

Author 3/20/2016 4:55 PM
Deleted: the

Author 3/20/2016 4:55 PM
Deleted: logics

Author 3/20/2016 4:55 PM
Deleted: logics

Author 3/20/2016 4:55 PM
Deleted: \cite{25884877

Author 3/20/2016 4:55 PM
Deleted: logics

Author 3/20/2016 4:55 PM
Deleted: ,

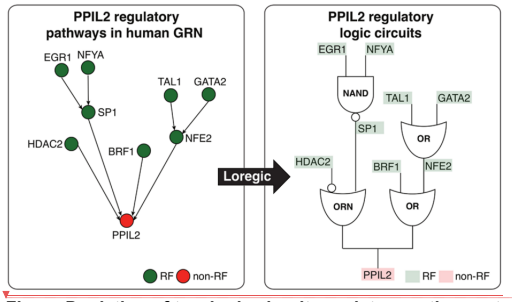
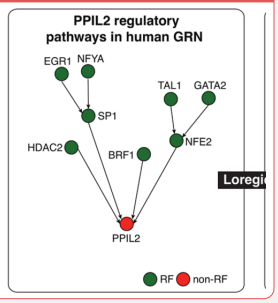


Figure Depiction of two logic circuit regulatory pathways targeting PPIL2.

Author 3/20/2016 4:55 PM



Deleted:

1.3 Integration & Harmonization of ENCODE data with datasets from other consortia

The overall goal of the DAC is to make the ENCODE annotation and analyses a broad and useful resource for researchers focusing on the human genome. To ensure this, the ENCODE annotation needs to compatible with other large genomics datasets. The integration of the ENCODE annotation with other datasets typically involves one or more of three different processes: 1) normalization and calibration of the ENCODE datasets relative to a somewhat differently normalized corpus; 2) relation of the more developed annotations from other consortia to the ENCODE corpus; 3) inference of evolutionary pressure within different ENCODE elements via genomic variation (germline and/or somatic) inferred from other large genomics datasets. In the sections below we go through a number of the large genomics datasets currently available GTEx, Roadmap, Psych ENCODE, 4D Nucleome, TCGA etc.) and talk about ways in which we plan to integrate the ENCODE data with these other large datasets. Below we list the DAC investigators who are also members of these other consortia and will leverage these relationships to help connect the datasets and annotations in a positive way.

Author 3/20/2016 4:55 PM
Deleted: forms

Author 3/20/2016 4:55 PM
Deleted: can be related

Author 3/20/2016 4:55 PM
Deleted: can be used to infer evolutionary pressure within different ENCODE elements.

Author 3/20/2016 4:55 PM
Deleted: ,

Author 3/20/2016 4:55 PM
Deleted: .

Author 3/20/2016 4:55 PM
Deleted: that

Author 3/20/2016 4:55 PM
Deleted: of

Author 3/20/2016 4:55 PM
Deleted: to

The Roadmap Epigenomics Project and International Human Epigenome Consortium. The DAC is working with the DCC to import consolidated epigenomes and unconsolidated experiments from the Roadmap Epigenomics Project and the International Human Epigenomics Consortium into the ENCODE portal. The DAC compiled ~7,000 consolidated and ~15,000 unconsolidated data files, including alignments, signal, and peaks, into 127 epigenome objects in ENCODE.

The Genotype-Tissue Expression Project (GTEx). GTEx is a large-scale data resource that studies human gene expression and regulation and their relationship to genetic variation. Manolis Kellis and Roderic Guigo are also part of GTEx and will help integrate the GTEx RNA-seq datasets into ENCODE pipelines for joint integrative analyses by the ENCODE AWG. In addition, the expression quantitative trait loci (eQTLs) identified by the GTEx project will be integrated with various ENCODE annotations.

PsychENCODE and BrainSpan. The PsychENCODE and BrainSpan projects are developing resources for genomic, regulatory, epigenomic, and proteomic landscapes in healthy and diseased brains, developing and adult brains, and in neural cell cultures. Mark Gerstein and Zhiping Weng are members of the PsychENCODE consortium and will ensure consistency of data processing and pipelines between the PsychENCODE and ENCODE consortia. In addition, the eQTLs identified in the PsychENCODE project will be integrated into the ENCODE annotations.

Author 3/20/2016 4:55 PM
Deleted: genomics

Author 3/20/2016 4:55 PM
Deleted: as well as

The Extracellular RNA (exRNA) Communication Consortium (ERCC). Mark Gerstein is in charge of the DAC for the exRNA consortium to develop the standardized RNA-seq pipelines for the analysis of exRNAs. We have developed a custom pipeline (exceRpt) for the analysis of small exRNA-seq data for the ERCC that is similar to the ENCODE small RNA-seq pipeline.

The 1000 Genomes Project. The Gerstein Lab aims to integrate different non-coding annotations with the germline variants identified by the 1000 Genomes Consortium. In this way, the non-coding annotation of each variant will be available to all users of the 1000 Genomes data and should serve as a valuable resource for the genetics community to identify causal variants in GWAS.

The Cancer Genome Consortium (TCGA) and International Cancer Genome Consortium (ICGC).
The Gerstein and Liu labs will integrate different non-coding annotations with the somatic variants identified by TCGA and ICGC. We will also integrate the RNA-seq and miRNA gene expression in the TCGA and ICGC datasets with ENCODE data by running the ENCODE pipelines on these datasets.

The 4D Nucleome Nuclear Organization and Function Interdisciplinary Consortium

Bill Noble is part of the 4D Nucleome consortium and is co-chairing its steering committee. He is thus in an excellent position to ensure coordination between the ENCODE and 4D Nucleome consortia. He is currently developing methods to integrate information from Hi-C and imaging experiments from the the 4D Nucleome consortium with functional genomics datasets during genome segmentation. These methods will be applied to the ENCODE Hi-C data.

1.4 Detailed Integration of Variants: Personal genomes and their use to develop allelic annotations

1.4.1 Personal genome construction

The alignment of assay reads is one of the main steps in processing functional genomic datasets. Conventionally, reads are aligned to the human reference genome. However, a systematic reference bias is introduced when reads are mapped to this haploid human reference sequence since reads that harbor an alternate allele are less likely to be aligned. In addition, reads can be improperly mapped to the reference genome in regions (or samples) with more genetic variation, especially when indels and larger structural variants are involved. This reduced mappability impairs estimation of read abundance and therefore compromises any downstream analyses. We propose to use a diploid personal genome as a more accurate representation of an individual's genome.

For personal genome construction, we have developed a computational tool, *vcf2diploid* {Rozowsky:2011gx}. The tool integrates an individual's genomic variation data (SNVs, indels, and SVs) into the reference genome. Phase information of heterozygous variants is also incorporated, producing maternal and paternal haplotypes. Chain files generated by the program can be used to account for coordinate offsets between the individual's parental haplotypes and the original reference genomic sequence. The versatility to convert between reference and personal genome coordinates allows mapping of genomic annotated regions (e. g. gene or peak coordinates for RNA-seq and ChIP-seq, respectively) between the genomes using available tools, such as the UCSC LiftOver tool {Rhead:2010if}.

Author 3/20/2016 4:55 PM
Deleted: of

Author 3/20/2016 4:55 PM
Deleted: the

Author 3/20/2016 4:55 PM
Deleted: variant

Author 3/20/2016 4:55 PM
Deleted: and these

Author 3/20/2016 4:55 PM
Deleted: of

Author 3/20/2016 4:55 PM
Deleted: \cite

Author 3/20/2016 4:55 PM
Deleted: , 2011 21811232

Author 3/20/2016 4:55 PM
Deleted: \cite

Author 3/20/2016 4:55 PM
Deleted: , 2010 19906737}

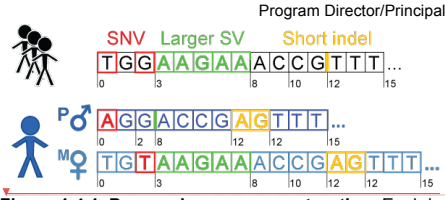


Figure 1.4.1. Personal genome construction. Each haplotype in the diploid personal genome is derived by incorporating phased or unphased variants (SNVs, indels and SVs) into the human reference genome. The coordinates can be mapped back to the human reference coordinates to facilitate comparisons with other reference-based resources, such as gene annotations from GENCODE.

We have previously constructed the personal diploid genome, splice-junction libraries and personalized gene annotations for NA12878 (also known as GM12878). We have made this assembly available as a resource at alleleseq.gersteinlab.org and have been updating it as new versions of the human reference genome, genomic annotations, and NA12878 genetic variation data are released. Furthermore, the availability of a computational tool enables the construction of personal genomes in a high-throughput fashion, as demonstrated in a recent publication\cite{Chen et al. Nat. Commun., in press} where we built 382 personal genomes using the variant call sets from the 1000 Genomes Project.

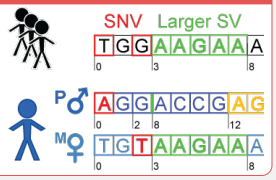
Since it has been demonstrated that using a diploid genome with an individual's variants improves both mappability of the reads{Rozowsky:2011gx} and results of the downstream analyses (particularly for SVs) {Sudmant:2015kz}, we propose to incorporate our personal genome construction tool, *vcf2diploid*, into ENCODE pipelines.

1.4.2 Developing allelic annotations

We also have extensive experience using personal genomes in analyses involving RNA-seq and ChIP-seq datasets. In particular, allele-specific (AS) analyses are very sensitive to mapping biases. Therefore in our pipeline *AlleleSeq*{Rozowsky:2011gx} the functional assay reads are aligned to a diploid personal genome, which alleviates the reference bias. We have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project {Djebali:2012hc}{Gerstein:2012fq}{Khurana:2013em}{Sudmant:2015kz}.

More recently\cite{Chen et al. Nat. Commun., in press}, we annotated variants associated with allele-specific expression (ASE) and binding (ASB) in a large pool of individuals from the 1000 Genomes Project. For this analysis, we integrated matching functional datasets (955 RNA-seq and 165 ChIP-seq in total), which include ChIP-seq datasets from 14 lymphoblastoid cell lines in ENCODE{ENCODEProjectConsortium:2012gc}. We developed a standardized framework and statistical approaches to aggregate and uniformly process the datasets. Using the beta-binomial test, we first estimated overdispersion in the allelic ratios distribution for each dataset (filtering ones with a high overdispersion parameter) and then identified statistically significant AS events. We have also incorporated strategies to alleviate ambiguous mapping bias – bias occurring when reads originating from one of the two alleles map to multiple locations. Overall, we detected more than 6K and 63K SNVs associated with ASB and ASE, respectively. These results were made available as an online resource, AlleleDB (alleledb.gersteinlab.org) and serve as an allele-specific annotation for the 1000-Genome variant catalogue.

1.5 Comparative analyses between human & mouse

Author 3/20/2016 4:55 PM


Deleted:

Author 3/20/2016 4:55 PM
Formatted: Highlight

Author 3/20/2016 4:55 PM
Deleted: \cite

Author 3/20/2016 4:55 PM
Deleted: , 2011 21811232

Author 3/20/2016 4:55 PM
Deleted: \cite

Author 3/20/2016 4:55 PM
Deleted: , 2015 26432246

Author 3/20/2016 4:55 PM
Deleted: in the use of

Author 3/20/2016 4:55 PM
Deleted: \cite

Author 3/20/2016 4:55 PM
Deleted: , 2011 21811232

Author 3/20/2016 4:55 PM
Deleted: \cite

Author 3/20/2016 4:55 PM
Deleted: , 2012 22955620;

Author 3/20/2016 4:55 PM
Deleted: , 2012 22955619;

Author 3/20/2016 4:55 PM
Deleted: , 2013 24092746;

Author 3/20/2016 4:55 PM
Deleted: , 2015 26432246

Author 3/20/2016 4:55 PM
Formatted: Highlight

Author 3/20/2016 4:55 PM
Deleted: \cite{ENCODE Consortium, 2012 22955616}.

Author 3/20/2016 4:55 PM
Deleted: them

Mapping orthologous genes: We will work closely with the AWG, the DCC and the phylogenomics community to create a resource of human and mouse orthologs in both genic and intergenic regulatory regions. Such a resource will be of great use to help transfer the biological knowledge gained from mouse models to human and to improve our understanding of human biology and health. We will start with the ortholog resources from established methods, such as Inparanoid, OrthoMCL and TreeFam, evaluate the different methodologies, and then propose a gold standard set of functional orthologs through integrating a variety of information from sequences, functional genomics and phylogenomics. We have developed a highly accurate tree-based phylogenomic pipeline for identifying orthologs and paralogs that overcomes the limitations of existing methods. In particular, we overcome common inaccuracies in the topology of individual gene trees, TreeFix{Wu:2013iv}, which, given a sequence alignment, maximum likelihood gene tree, and known species tree topology, outputs a highly accurate and error-corrected version of the gene tree. Using gene trees corrected by TreeFix, we minimized the effect of erroneous gene trees and greatly improved the accuracy of inferring orthologs and paralogs. Finally, we have developed an approach for integrating multiple ortholog sets using evidence from other species in order to benefit from the potentially complementary information available in discordant inferences.

Mapping orthologous non-coding regions: Given the strong focus of ENCODE on non-coding functional DNA elements, we will also develop and apply methods for mapping orthologous non-coding regions between human and mouse. Specifically, we will use sequence homology in the context of a syntenic map to recognize orthologous regions, and we will confirm their orthology based on functional genomics information. We have mapped promoter and enhancer regions between human and mouse based on orthologous annotations of chromatin states in the two species. We used peaks in H3K4me3 (promoter), H3K27ac (enhancer) and H3K27me3 (polycomb repressed) regions to define putative regulatory elements in mouse, and mapped each of them to the human genome using the UCSC multiple alignment chain files{Hinrichs:2006eq}. We calculated the human and mouse pairwise alignment for each multiple alignment, and selected the highest-scoring pairwise alignment as the base for constructing an orthologous region in human, by extension on either side using lower-scoring multiple alignments. We found that the chromatin state of orthologous regulatory regions was highly concordant (**Fig. 1.5**), giving us confidence in our mapping strategy, and indicating that the function of orthologous non-coding regions can be conserved across such evolutionary distances.

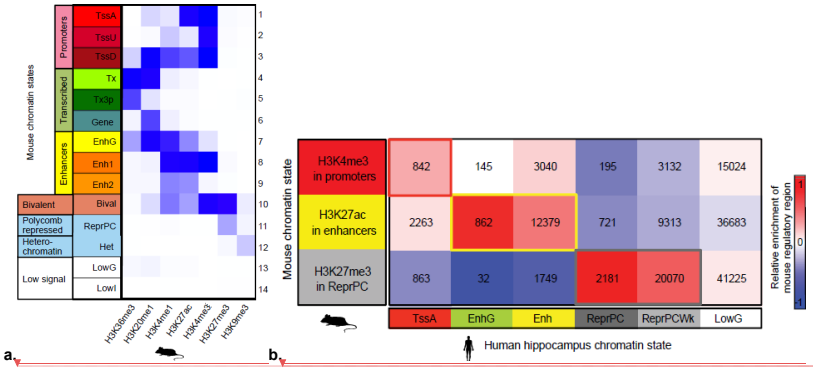


Figure 1.5. Chromatin state conservation in orthologous regulatory regions between human and mouse. **a.** Combinatorial patterns of the seven histone modifications profiled were used to define promoter, gene body, enhancer, bivalent, repressed Polycomb, heterochromatin, and low signal chromatin states. Darker blue indicates a higher enrichment of the measured histone mark (x axis) to be found in a particular state (y axis). **b.** Promoter, enhancer and repressed chromatin states in mouse hippocampus (rows), as profiled in this study, align to matching chromatin states in human (columns), as

Author 3/20/2016 4:55 PM
Deleted: ,

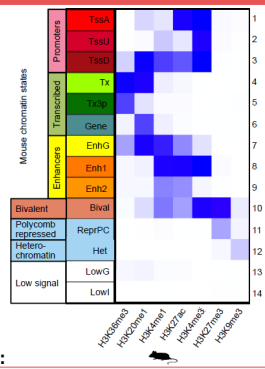
Author 3/20/2016 4:55 PM
Deleted: the

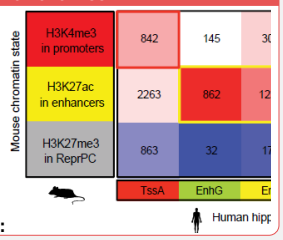
Author 3/20/2016 4:55 PM
Deleted: , 2013 #PMID:22949484

Author 3/20/2016 4:55 PM
Deleted: ,

Author 3/20/2016 4:55 PM
Deleted: a

Author 3/20/2016 4:55 PM
Deleted: , 2006, #PMID:16381938}.

Author 3/20/2016 4:55 PM

Deleted:

Author 3/20/2016 4:55 PM

Deleted:

profiled by the Roadmap Epigenomics Consortium. Shading indicates enrichment relative to human chromatin state abundance (columns). The number of regions overlapping is shown in each cell of the heatmap.

Aim 2. Serving as an informatics resource by supporting the activities of the ENCODE Analysis Working Group (AWG)

2.1 Assist the AWG in defining and prioritizing integrative analysis

With the large amount of data generated by the ENCODE production groups, it is particularly important to prioritize and accurately define the types of computational analyses performed by the Consortium. Therefore, we will work very closely with the AWG to coordinate activity of the DAC. The AWG will formulate the most important and informative biological questions and the DAC will assist with the prioritization by providing feedback on resource and time requirements. Following directions set by the AWG and coordinating with the DCC, data producers, and other Consortium members involved in computational analysis, the DAC will develop infrastructure for processing, analysis, and integration of ENCODE data. The leadership will assign new AWG requests to one or more components of the DAC. Depending on the type of analysis and data, the DAC investigation groups will assess existing or develop new tools and build computational frameworks for processing and analysis of the data. The DAC will work closely with the DCC to implement robust pipelines for automated processing and analysis of subsequent datasets generated by the production groups.

To facilitate this close interaction with the AWG (as well as other members of the Consortium), we will organize semimonthly conference calls for the entire AWG, small group conference calls, as well as one-to-one Skype sessions. During the biweekly AWG calls, the DAC will report progress on active tasks and receive new tasks from the AWG. For the latter, the DAC will provide assessments of resource requirements, suggest specific groups to work on the task, and estimate timelines, thus assisting the AWG in prioritization of current and future analysis directions. The DAC will be entirely open to communication and collaborations with other Consortium members and allow their participation in any analysis.

2.2 Define and further develop uniform processing pipelines for major ENCODE data types

Thousands of genomic experiments have been performed by the ENCODE members. Gaining novel biological insights using integrative analysis of these datasets relies on the fact that they are processed using accurate and consistent pipelines. The DAC has worked with various working groups during previous rounds of ENCODE to develop standardized reproducible pipelines for RNA-seq, TF and histone mark ChIP-seq, and whole genome bisulfite sequencing (WGBS) experiments. The DAC has also developed guidelines and practises for assessing the accuracy and replicability of each experiment. The current guidelines also assess antibody validation, experimental replication, sequencing depth, data and metadata reporting, and data quality. We have also worked with other international consortia such as IHEC and ICGC to ensure that the datasets produced from different consortia are processed in a similar fashion and can be compared with one another. The pipelines are available to the larger scientific community on DNAnexus, with all pipeline source code available on GitHub.

During ENCODE4, the DAC will continue to evaluate and further develop existing pipelines and quality metrics, as new computational algorithms and experimental techniques continue to emerge. We will also work with various working groups to develop new uniform processing pipelines for additional data types, e.g., fRIP-seq, eCLIP-seq, Hi-C, and ATAC-seq. In addition, the DAC is fully prepared to work with new data types and develop pipelines for them. Below we give an overview of the current state of ENCODE3 pipelines and the work that is planned for ENCODE4.

Mark Gerstein 3/19/2016 10:00 AM
Comment [8]: +pigwdf@gmail.com "Aim 3" (Or is it a 2)

Zhiping Weng 3/19/2016 10:00 AM
Comment [9]: This is Aim 2.

Author 3/20/2016 4:55 PM
Deleted: on

Author 3/20/2016 4:55 PM
Deleted: because

2.2.1 An RNA-seq pipeline

During ENCODE3, the DAC has been very active in the development of the RNA-seq pipeline, working closely with the DCC and data producers, in the context of the RNA working group. The DAC organized a benchmarking experiment to evaluate the RNA-seq pipeline, and performed extensive testing to guarantee reproducibility of the pipeline results independent of the computational platform used. The DAC has designed GRAPE, a robust, efficient and scalable open source software system for the storage, organization, access, and analysis of RNASeq data {Knowles:2013df} (https://github.com/guigolab/grape-nf). GRAPE has been implemented using the Nextflow (http://www.nextflow.io) and the Docker (https://www.docker.com/) technologies. Nextflow is a domain specific language for computational pipelines based on the Dataflow concurrency model which allows implicit task parallelization. Moreover, it supports the Docker container technology for improved reproducibility and software deployment, which allows effortless distribution and installation of the pipelines in different environments. The ENCODE3 pipeline, based on STAR {Dobin:2015by} for read mapping and RSEM {Li:2011cf} for transcript quantification has been implemented within GRAPE, and can be seamlessly deployed at any site supporting Docker.

The DAC also attempted to harmonize RNA-seq processing pipelines across different projects. The Gerstein lab is part of the transcriptome working group in the ICGC/PCAWG project and is working closely with them on unifying the processing of RNA-seq datasets. The distribution of the reproducibility metric scores across ICGC samples can also be used, as in the case of GTEx, as a guide to set the appropriate reproducibility thresholds for the ENCODE samples. Similarly the Guigo lab has been working with the Blueprint consortium (the European component of IHEC) to compare RNA-seq pipelines, and as a result, Blueprint has adopted the ENCODE RNA-seq pipeline.

The current RNA-seq pipeline focuses on gene (and transcript) quantification. However, other biologically relevant measures can be extracted from RNA-seq data, such as the use of alternative Transcription Start and Termination Sites (TSS and TTS), and most notably, the usage of alternative exons. The DAC has been involved in the implementation of the Integrated Pipeline for Splicing Analysis (IPSA) pipeline for detection and quantification of splice junctions (https://github.com/pervouchine/ipsa). IPSA is designed to provide a uniform and standardized processing protocol for different types of RNA-seq data, including stranded and unstranded data, data with or without biological replicates, etc. Its mission is to count split-mapped alignments corresponding to annotated and novel introns in a position-specific manner. Read counts are aggregated over offsets for each intron using Shannon entropy to control for the support level by distinct staggered positions and to exclude artefactual large counts. Introns receive a number of descriptors reflecting the annotation status, splice site nucleotides, and read counts in support of splicing as well as local intron retention at each splice site. These descriptors are used to compute the PSI (percent-spliced-in) metric for exons and introns and coSI (completeness of splicing index). The results of IPSA were compared to exon inclusion rates computed by the MISO software on the set of ~15K human cassette exons and the results were consistent (Pearson correlation r=0.91). An earlier branch of the pipeline was applied to the comparative analysis of conservation of splicing events between human and mouse using ENCODE2 human and Mouse ENCODE data (https://github.com/pervouchine/hm-splice-pipe).

2.2.2 Pipeline for analyzing ChIP-seq data of transcription factors

During ENCODE3, the DAC has worked closely with the ENCODE Binding Working Group to define and prototype a state-of-the-art computational analysis pipeline for transcription factor (TF) ChIP-seq data. Changes in approaches for generating and analyzing TF ChIP-seq data required us to revise the well-

Author 3/20/2016 4:55 PM
Deleted: independently from

Author 3/20/2016 4:55 PM
Deleted: [

Author 3/20/2016 4:55 PM
Deleted: D.G., 2013 PMID: 23329413]

Author 3/20/2016 4:55 PM
Deleted: to

Author 3/20/2016 4:55 PM
Deleted: [

Author 3/20/2016 4:55 PM
Deleted: A., 2015 PMID: 26334920]

Author 3/20/2016 4:55 PM
Deleted: [

Author 3/20/2016 4:55 PM
Deleted: B., 2011 PubMed PMID: 21816040]

Author 3/20/2016 4:55 PM
Deleted: metrics

Author 3/20/2016 4:55 PM
Deleted: focus

Author 3/20/2016 4:55 PM
Deleted: that

Author 3/20/2016 4:55 PM
Deleted: ISPA

Author 3/20/2016 4:55 PM
Formatted: Highlight

Author 3/20/2016 4:55 PM
Deleted: in ENCODE

Author 3/20/2016 4:55 PM
Deleted: pipelines

Author 3/20/2016 4:55 PM
Deleted: protocols

Author 3/20/2016 4:55 PM
Deleted: methods for

established and robust ENCODE2 pipeline. Particularly, the pipeline needed updating for the analysis of paired-end (PE) data that was not generated in ENCODE2. The updated pipeline also includes some of the latest computational peak calling algorithms that substantially improve the resolution and accuracy of TF binding site identification. In addition, the DAC (Kundaje), in collaboration with the Bickel lab, developed a new implementation of the Irreproducible Discovery Rate (IDR) framework {Li:2011kd} for evaluating and thresholding datasets based on reproducibility (https://github.com/nboley/idr). The implementation is a significant refactoring of the older version and is over 100x faster. An average run (comparison of 300K peaks across 2 replicates) takes less than 2 minutes in the new implementation. This collaboration is just one example of the many productive interactions between the DAC and computational groups, and will be continued in DAC4 (see the support letter from Drs. Bickel and Brown, and the support letter from Dr. Keles, the leader of another computational PI).

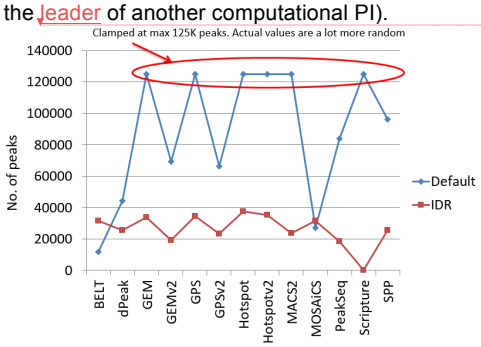
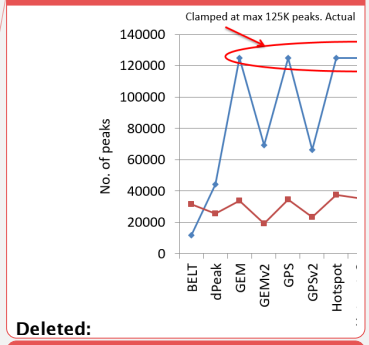


Figure 2.2.2a. The Irreproducible Discovery Rate framework stabilizes peak calls across most peak callers. The results of peak calling with and without IDR for several popular peak callers are shown for a Pol2 ChIP-seq dataset in H1 embryonic stem cells. The default thresholds (shown in blue) for a majority of peak callers are highly unstable and indicate vast differences between peak callers. After IDR (shown in red), the differences in the result values are significantly reduced.

The DAC led a coordinated effort across multiple labs (as part of the ENCODE Binding Working Group), to systematically compare and evaluate 10 peak calling algorithms (**Figure 2.2.2.a)**. A collection of benchmarking datasets were selected, spanning a diversity of TFs with different binding characteristics and of different data quality. The peak callers were evaluated based on 3 criteria: (1) reproducibility of peak calls across replicate datasets; (2) accuracy and resolution of peak calls, assessed by comparing predicted point binding sites to known motif locations; (3) the ability to deconvolve closely spaced and overlapping binding events. The analysis was conducted by Kundaje and two Computational groups (Keles and Gifford). After extensive analyzes and discussions, the Binding Working Group converged on three robust peak callers (SPP, PeakSeq and GEM) that scored well in the evaluations and are complementary to each other. SPP (Park lab) and PeakSeq (Gerstein lab - in the DAC) are both backward compatible with ENCODE2 and have been heavily tested. SPP takes into account read distributions accounting for peak shape. GEM (Gifford lab) deconvolves closely spaced binding events and integrates peak calling with motif discovery. PeakSeq uses a relatively different peak calling algorithm that adjusts well to highly punctate binding events and broader regions of enrichment.

Author 3/20/2016 4:55 PM
Deleted: to be updated to ... [7]

Author 3/20/2016 4:55 PM


Deleted:

Author 3/20/2016 4:55 PM
Deleted: (IDR)…framework stabi ... [8]

Author 3/20/2016 4:55 PM
Deleted: and…as part of the ENC ... [9]

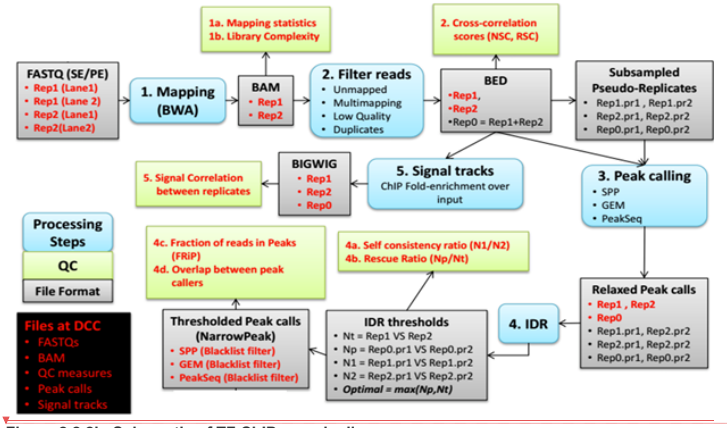


Figure 2.2.2b. Schematic of TF ChIP-seq pipeline.

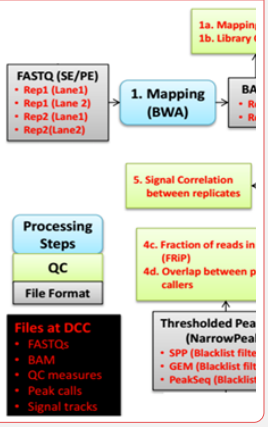
The DAC has developed a free, open-source implementation of the pipeline specification https://github.com/kundajelab/TF_chipseq_pipeline which is a mirror of the DCC's open-source official pipeline implemented on the cloud-based DNAnexus platform. A publication on the protocol specification of the ENCODE3 TF ChIP-seq pipeline is currently in revision at *Nature Protocols*. The key steps of the pipeline are summarized in **Figure 2.2.2.b.**

2.2.3 A pipeline for analyzing ChIP-seq data of histone modifications

Histone modification ChIP-seq data is substantially different to TF ChIP-seq data requiring its own unique analysis pipeline. The ENCODE2 Chromatin ChIP-seq pipeline was relatively underdeveloped due to the lack of effective peak callers and statistical methods for evaluating the highly variable sizes and patterns of thus called peaks. Also, the IDR method (while highly effective on TF ChIP-seq data) cannot be transfer directly to the analysis of histone marks due to a broad range of peak sizes and the instability of peak callers to differences in sequencing depth between replicates. During ENCODE3, the DAC had to develop a new method for evaluating the reproducibility of histone mark ChIP-seq data.

The DAC led a collaborative effort including members of the Binding Working Group, to evaluate six histone ChIP-seq peak callers and to define a uniform histone ChIP-seq analysis pipeline. High quality, deeply sequenced benchmark datasets for six key histone modifications with narrow (H3K27ac, H3K9ac, H3K4me3, H3K4me1) and broad (H3K36me3, H3K27me3, and H3K9me3) regions of enrichment in the human lymphoblastoid cell line GM12878 were obtained from the Snyder lab. These were distributed to the participating groups, and peak calls were generated and analyzed. The peak callers were evaluated based on (1) reproducibility of identifications, (2) stability of the peak calls to sequencing depth fluctuations (splitting and merging of peaks), and (3) accuracy and resolution assessed by comparison to gold standards (e.g., H3K4me3 peaks compared with promoters, H3K36me3 peaks compared with transcribed gene bodies identified using RNA-seq data). After careful evaluations, MACS2 and MoSAICS-HMM were found to be the top performing peak callers.

The ENCODE3 histone ChIP-seq pipeline is similar to the ENCODE3 TF ChIP-seq pipeline outlined in the

Author 3/20/2016 4:55 PM

Deleted:

Author 3/20/2016 4:55 PM
Deleted: DCC

Author 3/20/2016 4:55 PM
Deleted: by the Data Coordination Center

Author 3/20/2016 4:55 PM
Deleted: has

Author 3/20/2016 4:55 PM
Deleted: properties compared

Author 3/20/2016 4:55 PM
Deleted: a

Author 3/20/2016 4:55 PM
Deleted: that is

Author 3/20/2016 4:55 PM
Deleted: at adaptively thresholding

Author 3/20/2016 4:55 PM
Deleted: does not

Author 3/20/2016 4:55 PM
Deleted: differences

Author 3/20/2016 4:55 PM
Deleted: evaluated

Author 3/20/2016 4:55 PM
Deleted: peak caller

Author 3/20/2016 4:55 PM
Deleted:)

Author 3/20/2016 4:55 PM
Deleted: comparing with

previous section with several key differences. MACS2 and MoSAICS-HMM are the peak callers for calling narrow peaks and broad domains respectively. Instead of using IDR, we have designed a simple heuristic procedure to identify reproducible peaks and domains. For each sample, peaks/domains are called using reads pooled across replicates including those from individual replicates and pseudo-replicates generated using all pooled reads. The only peaks retained are those from the pooled data that significantly overlap (>50% length) peaks in both biological replicates or in both pseudo-replicates.

2.2.4 A motif pipeline for validating TF ChIP-seq data

Per request of the AWG, the DAC developed a motif-based antibody validation pipeline for ChIP-seq data of sequence-specific TFs. The pipeline evaluates the enrichment of known motifs in a given ChIP-seq dataset and provides a probabilistic assessment on how well the enrichment matches the labeled TF. The pipeline comprises the following steps: (1) The motif models (position weight matrix or PWM) are collected from public databases: TRANSFAC, JASPAR, PBM, and HT-SELEX (**Fig. 2.2.4a**). (2) The motif instances are defined by scanning the reference genome using the PWMs, excluding repeats, transposons, CDS and 3´-UTR. (3) The average of three motif enrichment scores are computed based on the motif instances in ChIP-seq peaks: a global enrichment score (obtained by comparing the motif PWM with its shuffled version), positional bias score (obtained by comparing peak center with the flanking regions) (**Fig. 2.2.4b**), and peak rank bias score (obtained by comparing the high peak score regions to the low peak score regions) (**Fig. 2.2.4c**). (4) The known motifs are grouped by their PWM similarities and each motif group is ranked by the highest average enrichment score within the group. We found that known motifs ranked highest in 80% of TF ChIP-seq datasets, while ~8% datasets failed to show enrichment for their corresponding known motifs, suggesting antibody problems, or indirect binding of these TFs. Because different TFs (e.g., TFs belonging to the same protein family) may share the same motif, or one TF may use multiple motifs, we developed a Bayesian framework for antibody validation that takes into account this ambiguity. Our approach sets specific cutoffs in accord to the characteristics of each TF and its corresponding motifs. This framework also points out antibodies with TFs that cannot be validated by our motif enrichment method due to their diverse binding preferences.

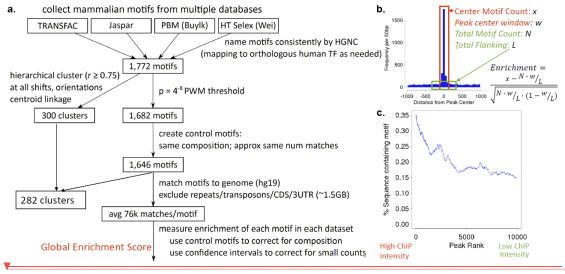


Figure 2.2.4. Criteria for assessing the validating the antibody of TF binding experiments based on regulatory motif enrichments. (a) Construction of the ENCODE motif catalog {Kheradpour,2014in} by processing and clustering multiple databases and experimental techniques. (**b**) Positional enrichment score based on regulatory motif instances relative to cluster centers. **(c)** Rank-based enrichment test based on motif instances relative to TF binding intensity distribution using a Mann Whitney rank sum test.

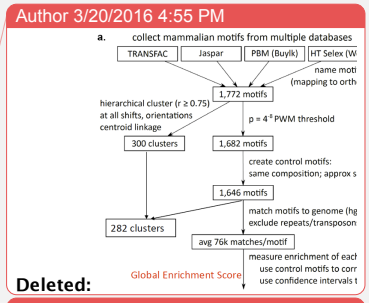
2.2.5 A pipeline for analyzing whole genome bisulfite sequencing (WGBS) data

The DAC developed a uniform analysis pipeline for WGBS data. We tested the combinations of three types of pre-processing methods, five widely used WGBS mapping tools (Bismark with Bowtie or Bowtie2, BSMAP, and

Author 3/20/2016 4:55 PM
Deleted: simpler…imple heuristic ... [10]

Author 3/20/2016 4:55 PM
Deleted: compare…btained by co ... [11]

Author 3/20/2016 4:55 PM
Deleted:



Author 3/20/2016 4:55 PM
Deleted: , NAR 24335146…2014i ... [12]

GSNAP, LAST and BRAT) with four different parameters of seed length for Bowtie and Bowtie2, and five types of post-processing criteria based on cytosine coverage and quality scores of mapping reads. Our results showed that Bismark+Bowtie and Bismark+Bowtie2 were the most accurate mapping algorithms throughout single-end and paired-end simulated datasets respectively. LAST showed comparable performance to Bismark. Due to the larger memory requirement for LAST, we selected Bismark+Bowtie and Bismark+Bowtie2 pipelines for the ENCODE single-end and paired-end WGBS uniform pipelines. To assist the DCC in implementing the WGBS pipelines on DNAnexus, we tested memory usage, disk usage, and run time for both human and mouse datasets. We also advised the DCC on the file format required for DNAnexus and the use of the lambda genome for quality control. The DCC finalized the uniform WGBS analysis pipeline on DNAnexus and deployed the source code on github (https://github.com/ENCODE-DCC/Bismark-ENCODE-WGBS).

Author 3/20/2016 4:55 PM
Deleted: Because of…ue to the la… [13]

2.2.6 Future developments of uniform processing pipelines

During the next ENCODE phase, the DAC will continue to assess the performance of existing uniform processing pipelines on the new experimental datasets. In collaboration with members of various ENCODE working groups, we will update the existing pipelines and quality metrics. We plan substantial extensions and improvements on the RNA-seq, histone mark ChIP-seq, DNase-seq, and WGBS pipelines as described below. In addition, we will develop new pipelines for new data types depending on the needs of ENCODE4. We provide two example pipelines for Hi-C and ATAC-seq, as we expect these data types to be prominent in ENCODE4. We point out that the DAC has a broad coverage of expertise and should be able to develop analysis pipelines for a wide variety of new data types. We will also draw upon talents in the Consortium, by collaborating with computational groups. In addition, we have set aside budget for recruiting one more lab into the DAC to meet unforeseen needs. See a letter from Prof. Uwe Ohler, who has developed several algorithms for analyzing CLIP-seq and related data types for RNA binding proteins.

Author 3/20/2016 4:55 PM
Deleted: in ENCODE… we will u… [14]

2.2.6.1 The RNA-seq pipeline and pipelines for other RNA types

The DAC will continue to benchmark the ENCODE3 RNA-seq pipeline in response to new developments of computational methods and sequencing technologies. Moreover, we will also continue to work with other consortia (e.g., ICGC, Blueprint, GTEx), by reaching out to the leaderships of these projects in an effort to harmonize RNA-seq pipelines and develop methods for normalizing the pipeline outputs across projects. An important challenge that we foresee is updating to new releases of the reference genome. Mapping of raw reads to the new genome assemblies for thousands of datasets in a large consortium like ENCODE demands considerable computing resources. The DAC will develop methods to lift-over transcript quantification across human and mouse genome assemblies that will not require remapping of the reads. The underlying idea is that only reads originating from altered regions of the genome will be remapped, while the quantification of genes and transcripts located in other regions can be updated without read remapping. The DAC will apply these "intelligent" lift-over methods to a variety of data types, as well as to personalized genome analyzes, e.g., running the RNA-seq analysis pipeline using reads mapped to the diploid personal genome corresponding to an individual from whom the samples were obtained (such as the ENTEx samples).

Author 3/20/2016 4:55 PM
Deleted: a…he reference genome… [15]

The DAC will also develop uniform processing pipelines for other RNA types (including small-RNA-seq, RAMPAGE, eCLIP-seq, fRIP-seq, etc.) that are currently in use in ENCODE3, as well as other technologies that could be used in ENCODE4, such as nascent RNA-seq, ribosome profiling, etc. Two technological developments appear to be particularly relevant in the field of RNA sequencing: single cell RNA-seq and sequencing of longer reads (i.e. those produced by technologies such as PacBio, Oxford Nanopore, 10x Genomics). We will pay particular attention to these areas, and plan our effort in accordance with the needs of

Author 3/20/2016 4:55 PM
Deleted: ,…(including small-RNA… [16]

ENCODE4. As an example, we outline a pipeline for long-read RNA-seq using the PacBio platform (**Figure 2.2.6.1**).

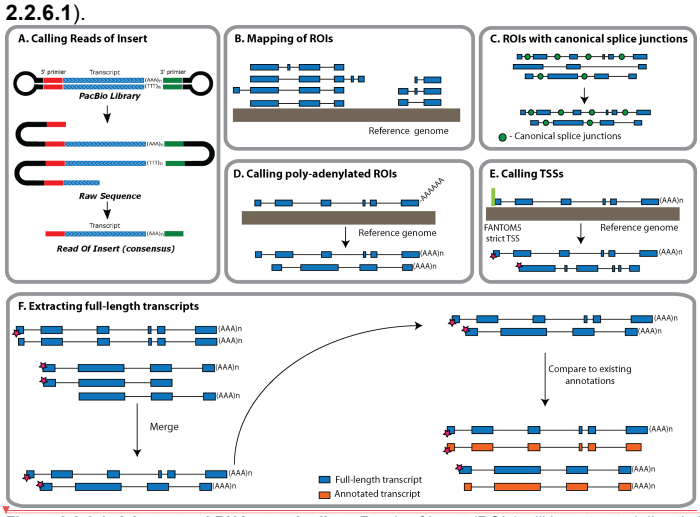


Figure 2.2.6.1. A Long-read RNA-seq pipeline. Reads of Insert (ROIs) will be extracted directly from the raw sequences using PacificBio's SMRT software and put into a FASTQ file. Only those reads fulfilling the "full-length" Iso-seq criterion (i.e. reads bounded by a library adapter at each end) will be kept for further analysis. Mapping of ROIs to the reference genome will be carried out using STAR or a comparable mapper. Given the relatively high error rate of the SMRT technology, we will analyze further only those spliced ROIs that are composed entirely of canonical introns, and take advantage of these splicing motifs to 'strand' the transcript. Mapped ROIs with polyA tails will be further analyzed, and used together with the splice junction information to infer the genomic strand of the corresponding transcript. The 5' ends of ROIs will be collapsed into TSS clusters and compared to the FANTOM5 strict TSS set. Redundant ROIs will be merged into full-length transcripts using cuffmerge or a similar piece of software. Next, the intron structures, TSSs and TTSs of full-length Iso-seq transcripts will be compared to existing annotations.

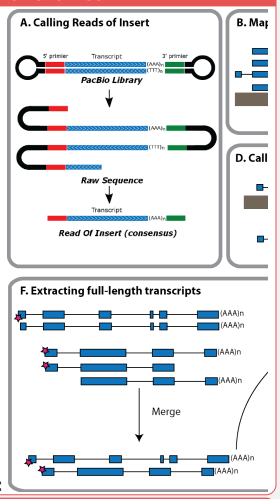
2.2.6.2 Further development of the pipelines for ChIP-seq of histone marks and DNase-seq

Despite substantial effort of the DAC, the ENCODE3 histone marks pipeline in underdeveloped. The accuracy of calling weak peaks is low. Peak boundaries are unstable and highly affected by low sequencing depth. Furthermore, merging across replicates leads to peak splitting. Despite the DAC's initial effort on developing a DNase-seq pipeline, it was decided to simply adopt John Stamatoyannopoulos' lab pipeline, since they are the only ENCODE3 lab generating DNase-seq data. The DAC is performing a systematic evaluation of ten histone ChIP-seq and five DNase-seq peak callers, aiming to improve the pipeline for punctate histone marks that are enriched in enhancers (H3K27ac, H3K9ac, H3K4me2 and H3K4me1) and the DNase-seq pipeline. During our effort in identifying enhancers, we observed that strong peaks of these histone marks and strong DNase peaks are highly concordant in a cell-type-specific manner, with DNase signal peaking in the troughs of the histone mark signals. Thus we can use this property to evaluate a pipeline's performance. Furthermore, we can use the results of functional assays, such as mouse transgenic assays and massively parallel reporter assays, to guide our effort. See support letters from two pioneering scientists of these functional assays, Len Pennacchio at LBNL and Nadav Ahituv at UCSF, for their willingness to share their data and collaborate with the DAC.

The ENCODE3 pipeline on broad histone marks (H3K36me3, H3K9me3, and H3K27me3) will be improved using a different approach. The DAC developed a peak caller called MUSIC {Harmanci:2014gx} after the ENCODE3 histone mark pipeline evaluation. MUSIC utilizes multi-scale decomposition of the ChIP-seq signal

Author 3/20/2016 4:55 PM
Deleted: reads

Author 3/20/2016 4:55 PM
Deleted:



Author 3/20/2016 4:55 PM
Deleted: output…ut into a FASTQ ... [17]

Author 3/20/2016 4:55 PM
Deleted: pipeline for …istone mar ... [18]

Author 3/20/2016 4:55 PM
Formatted: Font color: Custom Color(RGB(34,34,34)), Highlight

Author 3/20/2016 4:55 PM
Deleted: of John Stamatoyannopoulos, because… since they are the only ... [19]

Zhiping Weng 3/20/2016 3:21 AM
Comment [10]: super enhancer detection?

Author 3/20/2016 4:55 PM
Deleted: needs to

profile in conjunction with a novel method for correcting the effects of the multiple mapping reads. MUSIC outperforms other methods in identifying enriched regions within broad histone marks such as H3K36me3 and in correlating these enriched regions with expression profiles measured using RNA-seq. For ENCODE4, we plan to develop a principled measure for evaluating reproducibility of broad histone ChIP-seq data using a multi-scale binning approach (similar to MUSIC) coupled with the IDR framework. This approach will be less sensitive to the problems that plague most histone ChIP-seq peak callers—instability of peak boundaries and the effects of peak splitting and merging across replicates. The method will compute signal enrichment scores in replicate datasets for non-overlapping bins at different scales. The IDR method will then be used to evaluate the number of reproducible and rank consistent bins across the true replicates. Contiguous reproducible bins will be merged, and reproducible bins across multiple scales will be reconciled using a hierarchical merging strategy to obtain narrow and broad regions of variable size. These regions will be refined using signal processing methods for boundary detection. An analogous analysis will be performed on pairs of pseudo-replicates obtained by randomly sub-sampling reads from each replicate or by sub-sampling reads pooled across all true replicates. The reproducible peaks identified across true replicates will then be compared to those obtained from the pseudo-replicates to obtain a robust measure of reproducibility.

2.2.6.3 Detection of Differentially Methylated Regions

The DAC aims to develop in ENCODE4, a method for identifying statistically significant differentially methylated regions (DMRs) between two conditions (or two cell types) that is currently lacking in the ENCODE3 pipeline for whole-genome bisulfite sequencing (WGBS) data as described above (Section 2.2.5). The method will assess the statistical significance of DMRs even when only two biological replicates are available, which is the case for all ENCODE3 data.

Several existing methods are able to identify individual differentially methylated CpGs, and group these, in some ad-hoc way, into regions {Hansen:2012gr}{Wu:2015hm}{Sun:2014fk}{Lee:2015jf}{Dolzhenko:2014bo}. Although most of these methods attach adjusted p-values to the single CpGs, they do not offer a rigorous assessment of significance for the regions themselves. Methods that do assess statistical significance on regions require large sample sizes due to parametric assumptions {Park:2014ho}{Hebestreit:2013gb}. Other methods ignore biological variability all together {Saito:2014gd}{Akalin:2012cm}, or ignore correlation of methylation states of nearby loci {Park:2014ho} which might lead to misleading conclusions or result in loss of power.

We propose a new two-stage approach that first groups loci into candidate regions and then explicitly evaluates statistical significance at the region level while accounting for variability among biological replicates. In the first stage, we will quantify differential methylation levels between two conditions at each CpG locus. We will pool the reads from all technical replicates in each condition to achieve high coverage. Because the methylation levels of neighboring CpGs are highly correlated {Hansen:2012gr}, we will smooth the signal to further combat low coverage. Next we will define candidate DMRs by segmenting the genomes into groups of CpGs that show evidence of differential methylation in the same direction (e.g., higher in condition 1 than in condition 2). In the second stage, we will compute a t-statistic for each candidate DMR, which takes into account variability between biological replicates and adjusts for overall trends within the region. Finally we will compute an empirical p-value for each candidate DMR by permutation, and we will control the false discovery rate using the Benjamini-Hochberg procedure {Benjamini:1995tx}.

To demonstrate the feasibility of this approach we show encouraging preliminary results when applied to the mouse WGBS data during development, produced by Joe Ecker's lab in ENCODE. We identified 313,927

Author 3/20/2016 4:55 PM
Deleted: then

Author 3/20/2016 4:55 PM
Deleted: The ENCODE3 pipeline on whole-genome bisulfite sequencing (WGBS) data described above (Section 2.2.5) does not have the capability of

Author 3/20/2016 4:55 PM
Deleted:), and the DAC plans to develop such a method in ENCODE4.

Author 3/20/2016 4:55 PM
Deleted: then

Author 3/20/2016 4:55 PM
Deleted: PMID 23034175, PMID 26184873, PMID 24565500, PMID 26559505, PMID 24962134

Author 3/20/2016 4:55 PM
Deleted: PMID 24836530, PMID 23658421}.

Author 3/20/2016 4:55 PM
Deleted: PMID 24423865, PMID 23034086}

Author 3/20/2016 4:55 PM
Deleted: PMID 24836530

Author 3/20/2016 4:55 PM
Deleted: first

Author 3/20/2016 4:55 PM
Deleted: PMID 23034175

Author 3/20/2016 4:55 PM
Deleted: We then

Author 3/20/2016 4:55 PM
Deleted: and Hochberg, 1995

DMRs between forebrain and liver at E11.5, ten times more than between forebrain and hindbrain at the same developmental time point. We also detected many more DMRs during early (E11.5 vs. E16.5) than during late hindbrain development (E16.5 vs P0). One example is shown in **Figure 2.2.6.3**. We observed a great gain in power from statistically borrowing strength across adjacent CpGs, including those with low coverage—DMRs cover 59-70% of CpGs found by a single-CpG approach; in contrast, only 12-23% of the CpGs covered by the DMRs are significant using the single-CpG approach. Another advantage of our proposed approach is that the statistical inference is based on a well-defined parametrized function. Therefore we can extend this approach to compare multiple conditions. We can also extend the model to include terms that explain batch effect or systematic errors.

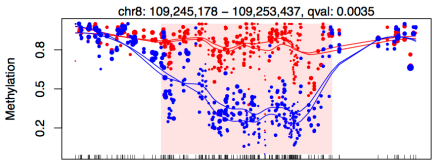


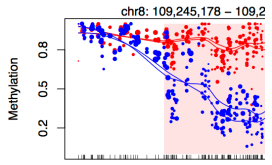
Figure 2.2.6.3. An example DMR. The points represent single CpG methylation level estimates which are plotted against genomic location. The size of the points is proportional to their coverage. The two colors denote the two conditions being compared. The shaded red area denotes the DMR reported to be statistically significant.

2.2.6.4 Development of a uniform processing pipeline for Hi-C

A standard Hi-C pipeline was not part of ENCODE3 and we plan to develop it in ENCODE4. The DAC (Noble) has been involved in the development of methods for high-throughput characterization of chromatin 3D architecture since the 2010 publication {Duan:2010jf}. Subsequently, the Noble lab has collaborated on the development of several novel chromatin conformation assays {Ma:2014dg}{Ay:2015ha}, and we have developed analytical methods for identifying colocalized sets of genomic loci {Witten:2012gs}, calling statistically significant contacts {Ay:2014bf}, inferring centromere positions {Varoquaux:2015kw}, and inferring 3D structures from Hi-C data {Varoquaux:2014ga}. We will bring this expertise to bear in the development and deployment of an analysis pipeline for Hi-C data generated by the ENCODE Consortium. The details of this pipeline will necessarily be worked out in collaboration with other members of ENCODE4, and based upon the pipelines being developed by the nascent 4D Nucleome Consortium. Noble is a co-chair of the Nuclear Organization and Function Interdisciplinary Consortium Steering Committee of 4D Nucleome, and he will work towards mantaining close ties between 4D Nucleome and ENCODE to ensure pipelines compatibility, interoperability of data, and data standards. See support letters from Bing Ren and Job Dekker, PIs of two production centers of the 4D Nucleome Consortium, expressing enthusiasm about collaborating with the DAC.

We envision a pipeline with a number of core components (see our recent review {Ay:2015gv}): a mapping module that incorporates sophisticated prior knowledge about sources of noise in Hi-C assays {Ay:2015gv}, an efficient normalization procedure {Yaffe:2011er}{Imakaev:2012dd}{Hu:2012es}{Cournac:2012iw}{Li:2015km}{Yang:2014br}{Shavit:2014kb}, a QC module that calculates various data quality measures, a method such as Fit-Hi-C {Ay:2014bf} for identifying significant contacts at a specified false discovery rate threshold, a complementary method such as HICCUPS for identifying statistically significant chromatin loops {Rao:2014eo}, and a method for calling domains at various scales, from fine-scale chromatin loops, through topologically associating domains (TADs) up to chromatin compartments {Dixon:2012fb}{Filippova:2014kc}{LevyLeduc:2014ka}. Importantly, the pipeline will be connected to a powerful visualization engine. Our recent evaluation suggests that the Wash U

Author 3/20/2016 4:55 PM
Deleted: hindbrain development. ... [20]

Author 3/20/2016 4:55 PM


Deleted:

Author 3/20/2016 4:55 PM
Deleted: Develop

Author 3/20/2016 4:55 PM
Deleted: in 2010 (…Duan 2010 ... [21]

Author 3/20/2016 4:55 PM
Deleted: (Ay 2015 ... [22]

William Stafford N…, 3/17/2016 11:00 PM
Comment [11]: [15] E. Yaffe and A. Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to charac-
terize global chromosomal architecture. Nat Genet, 43:1059–1065, 2011.
[16] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods, 9:999–1003, 2012.
[17] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu. HiCNorm: removing biases in Hi-C data via
Poisson regression. Bioinformatics, 28(23):3131–3133, 2012.
[18] A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, and J. Mozziconacci. Normalization of a chromosomal contact map. BMC Genomics, 13:436, 2012.
[19] W. Li, K. Gong, Q. Li, F. Alber, and X.J. Zhou. Hi-Corrector: a fast, scalable and memory-efficient package
for normalizing large-scale Hi-C data. Bioinformatics, 31(6):960–962, 2015.
[20] Ei-Wen Yang and Tao Jiang. GDNorm: An improved poisson regression model for reducing biases in Hi-C data. In Proceedings of the 14th International Workshop of Algorithms in Bioinformatics, volume 8701 of ... [23]

Author 3/20/2016 4:55 PM
Deleted: about the data… a meth ... [24]

Epigenome Browser {Zhou:2011js} is the best current visualization tool, though we will consider other alternatives as they become available. Current methods for inferring three-dimensional structures from Hi-C data do not scale to the full human genome and do not handle diploidy in a principled fashion. We are actively working on developing such methods, and if they become available, then they may be added to the pipeline.

2.2.6.5 Development of a uniform processing pipeline for ATAC-seq

The assay for transposase-accessible chromatin using sequencing (ATAC-seq) {Buenrostro:2013bc} was recently developed to profile chromatin accessibility in low input samples. We anticipate that ATAC-seq experiments will be performed in ENCODE4 and the DAC proposes to develop a uniform processing pipeline. The DAC (Kundaje) in collaboration with the Greenleaf and Chang (the inventors of ATAC-seq) have developed a prototype. The pipeline will generate appropriately mapped and filtered alignment files, peak calls, genome-wide signal coverage tracks as well as a detailed report containing mapping and several novel quality control measures.

The planned ATAC-seq pipeline will require as input raw FASTQ files and tentatively will contain the following steps: (1) preprocessing of the FASTQ files to trim adapters and align the reads/read-pairs to a reference genome using the Bowtie2 aligner {Langmead:2012jh} allowing a maximum insert size of 2Kb. Next we will filter the alignments to remove reads that are unmapped, unpaired, multi-mapping (MAPQ < 30) or are not primary alignments. We will use PICARD {cite} to mark and remove duplicates. We will also remove all mitochondrial reads. The read coordinates will then be adjusted by 4 bp in a strand-specific manner to obtain Tn5 insertion positions. (2) We will use the MACS2 {Zhang:2008gm} peak caller to obtain narrow peaks and sub-peaks of enriched signal. Specifically, we will use parameters that do not shift read positions (unlike ChIP-seq) and smooth the aggregate read counts by 150 bp (the approximate size of a nucleosome). We will also evaluate other peak callers for DNase-seq and punctate histone mark ChIP-seq data, as described in Section 2.2.6.2. The peaks will be then filtered against the ENCODE blacklist regions {Kundaje:GfK5zJ23}. (3) In order to take advantage of the two replicates per sample, we will use the IDR {Li:2011kd} framework to estimate high confidence, reproducible, rank-consistent peaks which is identical to the procedure outlined in the TF ChIP-seq pipeline (see Section 2.2.2). (4) We will also use MACS2 to generate normalized genome-wide signal coverage tracks representing the fold-enrichment and -log10(*p*-value) of ATAC-seq signal at each base pair relative to a Poisson background distribution adjusted for local biases.

2.3 DAC's Contribution to ENCODE working groups

The DAC carries out focused analysis in the context of various ENCODE working groups. As described in Section 2.2, the DAC participates actively in the Binding Working Group, and it built the two ChIP-seq pipelines (one for TF and the other one for histone marks) and the motif pipeline for validating antibodies used in TF ChIP-seq experiments by working closely with the group members. Below we briefly summarize DAC's contribution to other ENCODE working groups.

2.3.1 The functional characterization working group:

During ENCODE3, H3K27ac peaks in ChIP-seq assays for two mouse tissues (heart and forebrain) at the E11.5 developmental stage were chosen and tested for regulatory activity by the Pennacchio group using transgenic assays. Approximately 40 regions were tested in each tissue and 50% of them showed regulatory activity on average. Working in the context of the functional characterization working group, the DAC organized the ENCODE Enhancer Challenge and solicited predictions from the Consortium prior to the release of experimental results. Twelve computational groups made predictions in response to the challenge.

Author 3/20/2016 4:55 PM
Deleted: (

Author 3/20/2016 4:55 PM
Deleted: 2011 22127213)

Author 3/20/2016 4:55 PM
Deleted: develop

Author 3/20/2016 4:55 PM
Deleted: Develop

Author 3/20/2016 4:55 PM
Deleted: , 2013

Author 3/20/2016 4:55 PM
Deleted: we plan

Author 3/20/2016 4:55 PM
Deleted: requires

Author 3/20/2016 4:55 PM
Deleted: We first preprocess

Author 3/20/2016 4:55 PM
Deleted: then

Author 3/20/2016 4:55 PM
Deleted: cite

Author 3/20/2016 4:55 PM
Deleted: We

Author 3/20/2016 4:55 PM
Deleted: then

Author 3/20/2016 4:55 PM
Deleted: are

Author 3/20/2016 4:55 PM
Deleted: cite

Author 3/20/2016 4:55 PM
Deleted: are

Author 3/20/2016 4:55 PM
Deleted: cite

Author 3/20/2016 4:55 PM
Deleted: Irreproducible Discovery Rate (

Author 3/20/2016 4:55 PM
Deleted:) {cite

Author 3/20/2016 4:55 PM
Deleted: of the Binding Working Group.

The DAC assessed the accuracy of different methods as well as whether the predictions from different groups could be combined to make more accurate predictions. In particular, we developed an unsupervised ensemble approach to combine predictions from different algorithms for ENCODE. We compared the accuracy of different unsupervised ensemble approaches from different machine learning algorithms. Our preliminary results indicated that a number of computational methods predict enhancers more accurately than individual epigenomic datasets, and an unsupervised ensemble method consistently outperformed the most accurate individual machine learning models (**Figure 2.3.1**).

The DAC will continue to organize such challenges. The blind-test nature of these challenges greatly stimulates the development of new computational methods, and the critical assessment of existing methods. In the remaining time of DAC3, the DAC is planning to organize two more challenges, one for predicting the target genes of 15 enhancers and 15 super enhancers based on CRISPR knock-out results from Bing Ren's lab, and one for predicting enhancers based on the STARR-seq results from Kevin White's lab. With the newly established Functional Characterization Centers in ENCODE4, we anticipate a rapid accumulation of functional results for future challenges. The DAC also coordinates participation of prediction challenges outside ENCODE, e.g., the DREAM challenge described below. See support letter from Steven Brenner, the organizer of the Critical Assessment of Genome Interpretation (CAGI), expressing great interest in working with the DAC on future CAGI challenges.

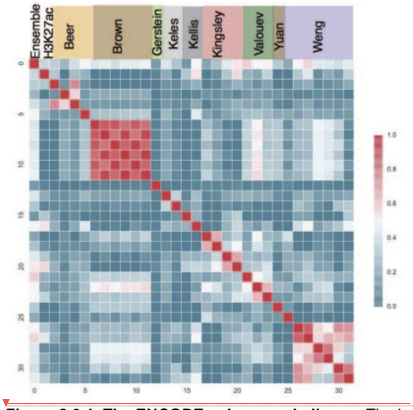


Figure 2.3.1. The ENCODE enhancer challenge. The heatmap shows comparison of the predictions made by different groups for H3K27ac peaks and the ensemble method. Different predictions from the same group were highly correlated, while predictions from different groups were very diverse.

2.3.2 The RNA working group

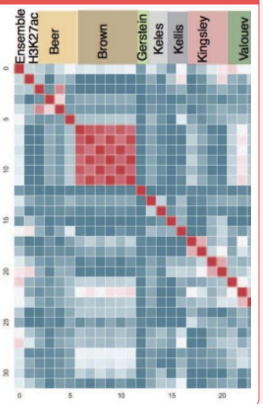
Working with the RNA working group, the DAC has been heavily involved in the development, implementation, and benchmarking of the ENCODE3 RNA-seq pipelines. We have computed gene expression matrices for human and mouse long RNA-seq experiments to store information about gene expression level across different ENCODE samples. The gene expression values are collected from the individual files processed with ENCODE3 long RNA-seq pipeline. Current version of this matrix includes TPM and FPKM values for annotated GENCODE genes per each long RNA-seq experiment. We have been exploring tools for visualizing the RNA-seq expression values across multiple samples. We proposed prototype plots and developed specifications for the generation of graphical representations of gene expression data. We explored the impact of moving to the

Author 3/20/2016 4:55 PM
Deleted: the

Author 3/20/2016 4:55 PM
Deleted: encyclopedia

Author 3/20/2016 4:55 PM
Deleted: the other

Author 3/20/2016 4:55 PM
Deleted:



Author 3/20/2016 4:55 PM
Deleted: The different

Author 3/20/2016 4:55 PM
Deleted: but

Author 3/20/2016 4:55 PM
Deleted: visualization of

GRCh38 reference genome in the estimates of gene expression. The DAC processed the ENCODE3 benchmark dataset using different genome assemblies and GENCODE annotation versions. While variations were observed for some gene families, overall gene expression values remain stable across human genome assemblies. In addition, we will make the personal genomes for each of the ENCODE cell lines and assess impacts of using personal genomes for RNA-seq analyses.

2.3.3 The analysis working group (AWG) and four AWG sub-groups [[DC2?: Throughout, both "sub-groups" and "subgroups" are used -- best to keep consistent]]

As described above, most of the activities within the DAC are carried out in the context of the AWG. The entire AWG holds semimonthly conference calls. Many of the activities of the AWG are related to ENCODE, described in Aim 3. AWG also has four sub-groups (cancer, GWAS, 3D Nucleome, and regulation). One DAC investigator co-chairs each of these four sub-groups. Below, we summarize the DAC activities within the four subgroups.

The cancer AWG subgroup: Recent progress made by The ENCODE Consortium has provided detailed annotation of non-coding regions within the human genome, and whole-genome sequencing of disease genomes has identified large volumes of variants therein, thus making this an opportune time to interpret the function of non-coding variants. We have summarized the current understanding of non-coding variants in cancer {Khurana:2016da}, and we also organized the ENCODE data as resource for cancer research. (1) In order to maximize the utility of ENCODE data for investigating cancer biology, we extracted the cancerous tissue/cell lines from ENCODE and matched them to specific tumor types. We found that 1,423 cell lines with ENCODE data (out of the 2,055 human cell lines) are actually cancer cell lines (69.2%). Each of the cancerous cell lines has been carefully matched to the specific cancer type, and the possible TCGA abbreviation was provided to enable further studies within the cancer community. (2) The current genome-wide functional characterization data from ENCODE provides the most comprehensive annotations of the human genome. Hence, we tried to collect a full category of non-coding annotations from ENCODE, and used them to annotate the newly-discovered mutational hot spots, thereby helping to provide insights into the biological mechanisms of potential cancer driver events. (3) Many genomic features from ENCODE may help to provide explanations of more than 90% of the variations associated with background mutational processes {Lawrence:2013fs}. Such data may thus directly aid in discriminating true driver events from those that result from mutational heterogeneity caused by external process {Lochovsky:2015bv}. However, the cancer community still faces the challenge of extracting such correlated features from the raw data produced by thousands of ENCODE experiments. Therefore, we carefully explored the entire set of ENCODE data and specifically selected all potentially related features. Results from uniformly processed pipelines were extracted and reformatted for direct community use. (4) We have recently developed a new computational method (Loregic) to characterize the gene regulatory mechanisms in complex systems {Wang:2015hn}. We used this method to identify genome-wide regulatory cooperativity in leukemia by integrating ENCODE and TCGA data.

The GWAS AWG subgroup: The focus of this subgroup is to maximize the impact of the ENCODE resource on studies related to human diseases, specifically in terms of interpreting results of genome-wide association studies (GWAS). This has been one of the most active AWG subgroups, with 40-50 participants on the semimonthly conference calls. These calls include presentations from individual labs seeking to coordinate activities, share methodological developments, publicize resources, establish guidelines, and receive feedback. Topics include guidelines on variant selection, expanding upon the use of linkage disequilibrium (LD) in statistical enrichment tests, as well as the use of LD for selecting regulatory annotations for enrichment analysis of genetic variants associated with complex traits and human disease. By making presentations and giving tutorials on the computational tools developed by the DAC at the annual CHARGE meeting, the DAC

Author 3/20/2016 4:55 PM
Deleted: DAC's…he activities with … [25]

Author 3/20/2016 4:55 PM
Deleted: of…ithin the human genc … [26]

Author 3/20/2016 4:55 PM
Formatted: Not Highlight

Author 3/20/2016 4:55 PM
Deleted: , …(Loregic) to character … [27]

Author 3/20/2016 4:55 PM
Deleted: This…he focus of this su … [28]

has also helped coordinate efforts with the ENCODE-CHARGE collaboration. One DAC lab also collaborates directly with the adiposity group within CHARGE. We are currently working with CHARGE to generate a phenotypic matrix for multi-phenotype analysis. The GWAS subgroup has also established a very active collaboration with the eMERGE consortium. This activity provides a unique opportunity to directly collaborate with researchers carrying out GWAS studies, and the ability to share intermediate results across different ENCODE groups. This activity has been the product of bi-weekly calls which are well attended and highly productive, with active discussions on the specifics of GWAS analyses, variant selection, LD, and interpreting the resulting associations.

The 3D nucleome AWG subgroup: This subgroup meets twice a month, and calls typically involve 25–35 participants. A primary function of the subgroup is to familiarize members of ENCODE with new datasets and novel methodologies relevant to the analysis of 3D chromatin conformation. The 3D Nucleome subgroup is currently working on two manuscripts. One effort, led by David Gilbert and Job Dekker, focuses on experimentally characterizing changes in topologically associated domain architectures in rearranged genomes. The manuscript exploits a panel of Hi-C datasets from various cancer cell lines generated by the Dekker lab, as well as complementary assays measuring rearrangements (from BioNano Genomics) and measuring replication timing. The second manuscript focuses on developing and comparing various measures of quality and reproducibility for Hi-C and ChIA-PET data. Work on the second manuscript is being led by the DAC (Noble) and a postdoc in the Noble lab, Gurkan Yardımcı. Dr. Yardımcı is carrying out a blind evaluation of a variety of metrics by providing members of the 3D Nucleome subgroup with collections of real and simulated Hi-C contact matrices.

The regulation AWG subgroup: This subgroup is developing integrative computational methods to predict high-resolution context-specific regulatory elements by integrating diverse types of ENCODE data and to characterize high-resolution combinatorial transcription factor co-binding patterns and regulatory sequence grammars underlying cell-type specific regulatory elements. A diversity of methods for predicting *in-vivo* transcription factor binding events from DNA sequence and chromatin accessibility data is now available. However, there is a need for systematic evaluation of these methods using matched datasets and performance measures. This subgroup is conducting an internal evaluation of methods developed by consortium members. We are also teaming up with the organizers of the DREAM Challenge to establish an open *in-vivo* transcription factor binding challenge using ENCODE data. This subgroup will train models on the Tier1 and Tier 2 reference cell-lines that have DNase-seq data and a large number of TF binding profiles based on ChIP-seq. Cross-cell type prediction performance will be evaluated. Well-calibrated models will be used to infer high-confidence binding maps of TFs in ~300 cell types and tissues that have DNase-seq data but lack comprehensive ChIP-seq experiments. These predicted maps will complement the primary functional element lists (derived primarily for specific experimental assays) in the Encyclopedia and will also provide high-resolution, base-pair level cell-type specific annotations of combinatorial binding events in regulatory elements.

2.4 Facilitating the production of Consortium papers and reports

Key ongoing roles of the DAC are to facilitate the analysis and writing associated with integrative Consortium papers (including standards and reports), integrate the data to be produced in ENCODE4, and to leverage these data to help interpret GWAS variants. The DAC will facilitate each of these objectives by performing integrative analyses and providing technological infrastructure for the analyses specific to each paper.

Author 3/20/2016 4:55 PM
Deleted: in…ithin CHARGE. We a… [29]

Author 3/20/2016 4:55 PM
Deleted: group…ubgroup meets t… [30]

Author 3/20/2016 4:55 PM
Deleted: The…his subgroup is … [31]

Author 3/20/2016 4:55 PM
Formatted: Font:Italic

Author 3/20/2016 4:55 PM
Deleted: ….is now available. How… [32]

Author 3/20/2016 4:55 PM
Formatted: Font:Italic

Author 3/20/2016 4:55 PM
Deleted: The Regulation…his sub… [33]

Author 3/20/2016 4:55 PM
Deleted: writing

Author 3/20/2016 4:55 PM
Deleted: A key continuing role…e… [34]

With respect to helping to develop the technological infrastructure needed for the collaborative writing of consortia papers, it is important to note that many genomics papers are complex (often involving tens to hundreds of authors) and they often contain many main figures, tables and supplementary exhibits. Tracking and sharing these materials is a non-trivial task. To address this, in addition to coordinating the production of papers with mailing lists and conference calls, we will make use of various online document management and collaboration systems for editing and collaboration (by using Wikis and Google Docs), content management (by using Drupal and Joomla), references management (by using Endnote, Bookends and Papers), references and figures tracking (by using Mendeley and BibTex), and for distribution of files and figures (by using Dropbox). We will provide expertise in developing this infrastructure to the future consortium publications.

The DAC also helps to integrate the underlying data, genomic assays, analysis results and documents. We consider the papers of a given roll out (i.e., those based on a single data freeze) and their associated data to constitute and informational hierarchy. At the top of the hierarchy is the main paper, which summarizes everything for the encyclopedia. The main paper refers to companion papers devoted to specific studies. Each individual paper discusses a large dataset and connects it hierarchically to the tables and figures in the main paper. The datasets that are most commonly referenced, as well as the secondary analyses results (e.g., peak calls and segments), which point to the actual underlying raw data (usually sequencing data), are stored in central repositories (such as short-read archive). We will hierarchically link the tables and figures in each paper to a chain of specific analysis results, programming scripts, corresponding versions of subsidiary results, and to the raw data. These will be stored in paper supplements and project web sites. Hierarchical information structures will enable us to organize data intuitively. In addition, the clear documentation will be essential for reproducibility.

The consortium publications on which the DAC has worked have been highly cited. The DAC has analyzed the patterns of dissemination of ENCODE publications to study how the outside scientific community benefits from ENCODE data. In particular, using publication data related to the ENCODE consortium {ENCODEProjectConsortium:2012gc}, we constructed co-authorship networks using temporal data from 2004 to 2014 (**Figure 2.4**). The networks show how the information from the consortium has diffused through authorship relationships. We found that the Consortium works as a community, whereas non-members collaborate on a smaller scale, and a few brokers initiate the connections between the consortium and non-members. Thus, large scientific consortia should set up formal outreach groups to communicate with outside researchers {Wang:2016gr}.

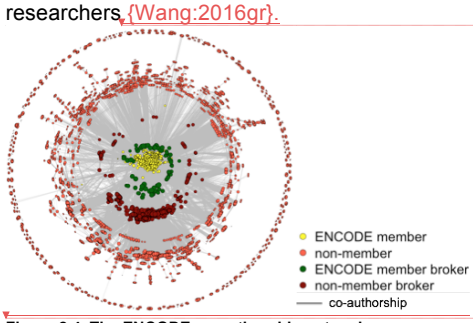
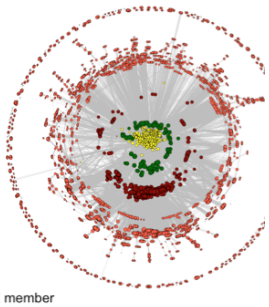


Figure 2.4. The ENCODE co-authorship network.

Author 3/20/2016 4:55 PM
Deleted: The DAC will set up a …[35]

Author 3/20/2016 4:55 PM
Deleted: Another important responsibility of the DAC in writing consortium papers is connecting…he DAC also helps to …[36]

Author 3/20/2016 4:55 PM
Deleted: that…n which the DAC h…[37]

Author 3/20/2016 4:55 PM


Author 3/20/2016 4:55 PM
Deleted:

Aim 3. Creating high-quality Encyclopedias of DNA elements in the human and mouse genomes

The ENCODE Consortium is generating data that is rich in terms of both quantity and variety. Many integrative analyses are being performed on these data, especially across multiple data types. It is a major challenge to present the analysis results in a comprehensive and yet concise way while simultaneously facilitating the further research and clinical activities within the broader scientific community. An important function of the DAC is to produce standardized output from the various pipelines and to extract the basic elements into the ENCODE Encyclopedia. Working under the direction of the AWG, the DAC seeks to achieve maximal utility, easy interpretability, and high accuracy when we choose a result to be included in the Encyclopedia. For instance, functional elements can be used for variant interpretation, or for analyzing more elaborate structures such as chromatin loops. We provide the Encyclopedia as downloadable files and supply tools for their visualization. Working as an integral part of the AWG, the DAC generates files associated with the Encyclopedia and submits these files to the DCC, along with metadata describing the analysis methods, including versions of programs and pipelines. The DAC also develops prototypes of visualization tools and works with the DCC to implement these tools at the ENCODE portal (https://www.encodeproject.org/data/annotations/).

In order to most accurately reflect primary experimental data, we propose to build the Encyclopedia using a bottom-up approach which consists of three annotation levels (**Figure 3**). The ground level of annotations stay close to raw experimental data, and are typically derived from individual types of experimental data. The middle level of annotations require the integration of multiple types of experimental data, and the top level requires even greater integration. Below, we describe the current and planned components of Encyclopedia from the ground up.

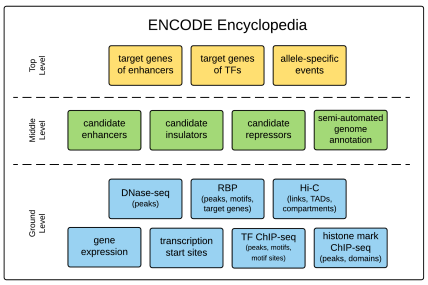


Figure 3. The structure and composition of the ENCODE Encyclopedia.

3.1 The ground level of annotations in the Encyclopedia

The main data types of the ENCODE Consortium currently include RNA-seq, RAMPAGE, ChIP-seq of TFs and histone marks, DNase-seq, WGBS, Hi-C, and eCLIP-seq, fRIP-seq, and Bind-n-seq of RBPs. We propose to extract the most direct annotations from these data to include in the Encyclopedia. Here, we describe our proposed Encyclopedia components based on existing ENCODE data types. We anticipate new data types in ENCODE4, and will build the Encyclopedia components for them using similar approaches to what is described in this section.

Mark Gerstein 3/19/2016 9:58 AM
Comment [12]: +michael.rutenbergschoenberg@yale.edu Aim2 starts here

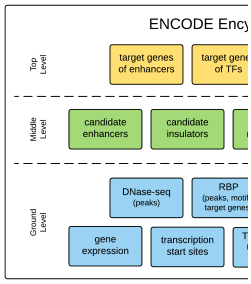
Zhiping Weng 3/19/2016 9:58 AM
Comment [13]: I swapped Aims 2 and 3. Now the Encyclopedia Aim is Aim 3.

Author 3/20/2016 4:55 PM
Deleted: .

Author 3/20/2016 4:55 PM
Formatted: Space After: 5 pt

Author 3/20/2016 4:55 PM
Deleted: data,…n terms of both in… [38]

Author 3/20/2016 4:55 PM
Deleted: stay as truthfully as possible to the…ost accurately reflect primary … [39]

Author 3/20/2016 4:55 PM


Deleted:

Author 3/20/2016 4:55 PM
Deleted: , ChIP-seq of…and histo… [40]

3.1.1 Transcription

The most direct result of analyzing a long-RNA-seq dataset (sequencing of RNA fragments longer than 200 nt) is the expression levels of all annotated genes. Thus, for a collection of RNA-seq data in multiple cell types, we can build a gene expression matrix, with genes in one dimension, cell types in the other dimension, and each entry of the matrix being the expression level of a gene in a cell type. Working closely with the RNA group (**Section 2.3.2**), the DAC has released two versions of gene expression matrices at the ENCODE Portal. The first is for all ENCODE2 RNA-seq datasets, and the second additionally includes ENCODE3 RNA-seq datasets up to Oct 15, 2016.

RAMPAGE {Batut:2013kc} is an assay for identifying transcription start sites (TSSs) at base-pair resolution and quantifying their associated expression levels. ENCODE has released RAMPAGE data for 71 human cell types. Genomic regions that are significantly enriched for RAMPAGE signal indicate TSSs, and the DAC has developed a computational pipeline to confidently identify such regions. The pipeline outputs a list regions with associated expression and confidence scores. The confidence scores are calculated by comparing the number of reads in a TSS to the background transcription rate estimated from a matching RNAseq experiment. Regions that are not consistently expressed between replicates are filtered using the IDR procedure (described in the TF ChIP-seq pipeline in **Section 2.2.2**), and the remaining TSSs are merged across biological samples to produce a unified TSS list. Finally, a TSS expression matrix is produced by quantifying the RAMPAGE signal in each of these regions for each sample. We propose to curate lists of cell type specific TSSs and a TSS expression matrix (akin to the gene expression matrix) to be included in the Encyclopedia.

The DAC has also established a prototype for visualizing the gene and TSS expression matrices, and we are currently working with the DCC to implement this at the ENCODE portal. In the interim, the DAC has worked with Prof. Feng Yue to display the gene expression matrix (http://promoter.bx.psu.edu/ENCODE/). Prof. Yue's website can display the expression levels of a user-specified gene across all cell types with ENCODE RNA-seq data. In addition to displaying the expression profile of one gene as a barplot or boxplot, the visualizer being built by the DAC and DCC will also display the expression profile of a set of user-specified genes as a heatmap. The user will be able to toggle between different scales of expression levels (linear or log), cellular compartments (cytoplasm, nuclear, or nucleolar) and RNA-seq protocols (total, polyA+, or polyA–). To make the visualizer more accessible and user-friendly to those who are already familiar with the GTEx portal, it will be similar to the GTEx portal in terms of appearance and functionality. As will become apparent soon, the visualizer for expression matrices will be adapted to visualize other Encyclopedia components expressed in the form of a matrix.

Going forward, we will update the gene and TSS expression matrices and the visualizer for each data freeze. There are two main areas of development that we envision. First, the DAC is currently evaluating algorithms that quantify the expression levels of transcripts, with the goal of identifying a method for building a transcript expression matrix. The major challenge is that the RNA-seq analysis algorithms need to discriminate the reads that come from different transcripts of the same gene. Developments in sequencing technology (such as longer read lengths) will greatly alleviate this challenge, and in **Section 2.2.6.1** we describe our plan for building a long-read RNA-seq analysis pipeline. Secondly, it remains a challenge to normalize expression levels across multiple cell types (especially with respect to dealing with batch effects). In ENCODE3, most human (mouse) RNA-seq data are produced by the Gingeras (Wold) lab. We need to minimize the batch effects so that we can perform the comparison between human and mouse. Similar considerations apply to RNA-seq data generated using different protocols, e.g. rRNA-depleted total RNA vs. polyA+ mRNA. We will perform normalization and calibration to minimize batch effects (see **Section 3.1.6**).

Author 3/20/2016 4:55 PM
Deleted: matrix…atrices at the EN… [41]

Author 3/20/2016 4:55 PM
Deleted: PMID: 22936248…atut:2… [42]

Author 3/20/2016 4:55 PM
Deleted: it…his at the ENCODE p… [43]

Zhiping Weng 3/20/2016 1:32 AM
Comment [14]: grouping of cell types by ontology

Zhiping Weng 3/20/2016 1:32 AM
Comment [15]: grouping of cell types by ontology

Author 3/20/2016 4:55 PM
Deleted: form

Author 3/20/2016 4:55 PM
Deleted: lies in…s that the RNA-s… [44]

Author 3/20/2016 4:55 PM
Formatted: Not Highlight

Author 3/20/2016 4:55 PM
Deleted: lab effect…atch effects s… [45]

In addition to the expression matrix for annotated genes and transcripts from long-RNA-seq datasets, we plan to generate similar expression matrices for genes and transcripts using short-RNA-seq datasets (RNA fragments shorter than 200 nt), for miRNAs using small-RNA-seq datasets (RNAs shorter than 35 nt). The processing pipelines for these other types of RNAs are not yet established, and the DAC will work closely with the labs that produce these data to develop these pipelines, typically in the context of a working group (**Section 2.3** describes how the DAC conducts its work in various working groups). In addition, ENCODE may perform new types of experimental assays for RNA-level quantification, such as single-cell RNA-seq, GRO-seq, etc. As described in **Section 2.2.6.1**, the DAC will develop uniform processing pipelines for these new data types. We also plan to generate expression matrices for these data to include in the Encyclopedia. The visualizer that we will develop for RNA-seq expression matrices can be applied to these expression matrices directly or with small adjustments (e.g., in the case of miRNA), and we will design a consistent and interactive platform for visualizing and comparing all types of expression matrices.

To summarize, the Encyclopedia components for transcription-related data types (e.g., RNA-seq, RAMPAGE, small-RNA-seq, GRO-seq etc.) will be a series of 2D matrices, each with one dimension representing genes or transcripts and the other representing cell types. These matrices will be carefully normalized to avoid batch effects, and they may readily be downloaded. They can also be visualized in an interactive manner for individual genes or transcripts as bar plots, or for a set of genes or transcripts as heatmaps.

3.1.2 Transcription factor binding regions (ChIP-seq peaks), motifs, and motif sites

ENCODE has released over 1000 ChIP-seq datasets for transcription factors in human and mouse. The genomic regions that have significantly more reads in the ChIP sample (relative to the input sample) signify TF binding. These regions are commonly called ChIP-seq peaks (or peaks in short), and uniform processing pipelines established by the DAC can reliably identify these peaks. As described in **Section 2.2.2**, the pipeline outputs a list of peaks, each is associated with a score that is computed based on the enrichment of reads in the ChIP sample over the input sample. The pipeline also takes advantage of two biological replicates per experiment using the irreproducibility discovery rate (IDR) framework, which requires that the ranks of a peak in the two replicates be "reproducible" with respect to each other. For ChIP-seq of TF data, the Encyclopedia will contain lists of ChIP-seq peaks, enriched motifs, and motif sites within the peaks. Users can query the ENCODE portal for a set of TF ChIP-seq data and download their corresponding peak lists, or visualize these peak lists using the UCSC genome browser or the WashU epigenome browser. The DAC will work closely with the DCC to implement the functionalities of selecting specific datasets for intuitive viewing using these browsers. Specifically we will build a system to suggest datasets in biologically related cell types or TFs with overlapping binding regions in the genome.

The DAC has developed a TF centric, web-based, wiki-style system for visualizing the entire set of ChIP-seq peaks for each TF at http://factorbook.org. Factorbook currently contains all of the released ENCODE ChIP-seq datasets for TFs, arranged in two TF-by-cell-type matrices (one for human and one for mouse; **Figure 3.1.2a**). Users can select a TF, or a ChIP-seq dataset of a TF in a particular cell type in the matrices, to get to a web page for that specific TF. The main idea behind Factorbook is to gather all ENCODE ChIP-seq data (on both TF and histone marks) related to a query TF and to display them in an organized fashion. Each TF page contains several panels: 1) functional annotations of the TF from major public sources; 2) line plots showing average histone mark signal profiles in a ±2 kb window centered on the TF peak summits computed from ChIP-seq data in the same cell type (**Figure 3.1.2b**); 3) sequence motifs enriched in top-ranked peaks (**Figure 3.1.2c**); 4) heatmaps showing histone mark levels at individual ±2 kb windows around TF peaks, ranked in the

Author 3/20/2016 4:55 PM
Deleted: …level quantification, su… [46]

Author 3/20/2016 4:55 PM
Deleted: …related data types, … [47]

Author 3/20/2016 4:55 PM
Deleted: compared with…relative … [48]

Author 3/20/2016 4:55 PM
Formatted … [49]

Author 3/20/2016 4:55 PM
Deleted:

Author 3/20/2016 4:55 PM
Deleted: , …(one for human and o… [50]

same order as the TF peaks (**Figure 3.1.2d**); 5) heatmaps showing ChIP-seq levels of other TFs at individual ±2 kb windows around the query TF peaks, ranked in the same order as the query TF peaks; and 6) average nucleosome occupancy signals around TF peaks.

The motif section contains up to five of the most highly enriched motifs in the top 500 ChIP-seq peaks, as computed using the MEME-ChIP method. Usually the cognate motif of the TF (i.e., the motif dictated by the biophysical binding preferences of the TF) is the most enriched motif, and additional enriched motifs may correspond to co-factors. In some cases however, the most enriched motif corresponds to a co-factor, and MEME-ChIP also tends to predict low-complexity false positives. In order to filter out spurious motifs, we require a motif to be enriched in the entire set of ChIP-seq peaks, compared to flanking regions and GC-matched non-peak regions in the genome. In the motif section of Factorbook, the motifs that are filtered out are "gray out". In addition, we implemented the crowd-sourcing capability for users to vote on a motif. Commends and references that support the vote are provided (**Figure 3.1.2c**).

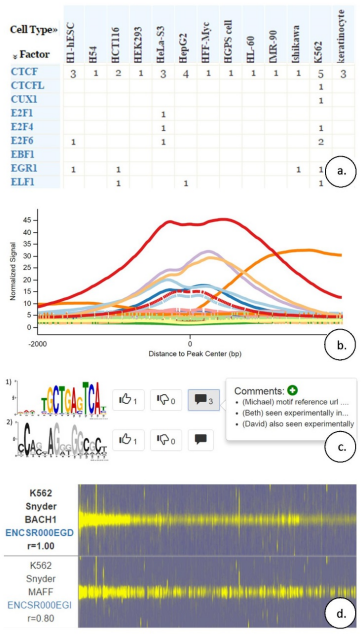
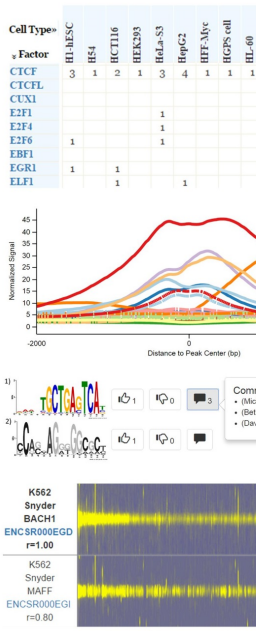


Figure 3.1.2. Factorbook: a TF-centric information resource (http://factorbook.org). a. A TF-by-cell-type matrix. Users can select a TF (an entry in the factor column) or an entry of the matrix, indicating the number of ChIP-seq datasets that ENCODE has released for a TF in a cell type (when the number is greater than 1, multiple ENCODE labs have released datasets for the same TF in the same cell type). **b.** Average histone mark signals centered on the summits of TF peaks. Different histone marks are shown in different colors, and the data are cell type specific. **c.** Sequence motifs enriched in ChIP-seq peaks. The lower motif is grayed out because it did not pass our filtering criteria. Users can vote on each motif and provide comments. **d.** Heatmaps show ±2 kb windows centered on the peaks of the BACH1 TF, with the strongest peak at left. TF ChIP-seq datasets for other TFs in the same cell type (K562 in this example) are shown as subsequent heatmaps, sorted in descending order of the correlation coefficient (r) of signals across the peaks of the query TF (BACH1 in this example).

Author 3/20/2016 4:55 PM
Deleted:)….; 5…. heatmaps show ... [51]

Author 3/20/2016 4:55 PM
Deleted: significant…nriched motif ... [52]

Author 3/20/2016 4:55 PM
Deleted:



Author 3/20/2016 4:55 PM
Deleted: The…actorbook: a TF-C ... [53]

Author 3/20/2016 4:55 PM
Formatted: Not Highlight

Author 3/20/2016 4:55 PM
Deleted: the …escending order by ... [54]

We envision several directions to improve Factorbook. (1) We will implement the capability to allow users to visualize the underlying ChIP-seq signal of the TF of interest along with other ENCODE data at specific loci by connecting to the UCSC and the WashU genome browsers directly from Factorbook. We have built a visualizer for candidate enhancers (**Section 3.2.1**), which can be incorporated into Factorbook. (2) The DAC is currently evaluating several existing methods for predicting motifs and motif sites, and it is developing deep learning methods that can learn the motif and their sites in the entire set of ChIP-seq peaks (not just the top-ranked peaks) using ChIP-seq signals at nucleotide resolution. Another direction is to explore more complex motif models (such as k-mers {ref}) or to incorporate DNA shape {ref}. (3) We will expand the crowd-sourcing capability in the motif section of Factorbook to other sections. (4) We will link the orthologous TFs in human and mouse, and we will develop tools to compare their binding characteristics. (5) We will implement a system that provides suggested datasets to the user (e.g., TFs with similar binding characteristics in a different cell type, but does not have ChIP-seq data in the cell type being queried). Such a system will become increasingly desirable as more datasets become available.

Author 3/20/2016 4:55 PM
Deleted: in…t specific loci by con… [55]

3.1.3 histone modification enriched regions and DNase hypersensitive sites

The DAC has developed a uniform processing pipeline for identifying genomic regions enriched in histone marks (**Section 2.2.3**). These regions are commonly called peaks, analogously with TF peaks. This pipeline is particularly effective for histone marks that have punctate signals, and especially for ones are enriched in promoters or enhancers (H3K4me3 and H3K27ac), but is less effective for those that have broad signals (e.g., H3K4me1) or that occupy large domains (H3K27me3 and H3K9me3). As described in **Section 2.2.6.2**, the DAC will continue to develop and evaluate algorithms for identifying histone mark peaks or domains, and improve the uniform processing pipeline. In that section we also described plans to further improve the DNase-seq pipeline.

Author 3/20/2016 4:55 PM
Deleted: in an analogous way to … [56]

For histone mark ChIP-seq and DNase-seq datasets, the Encyclopedia contains peak lists for a large collection of cell types. We provide these as downloadable files which can be opened with genome browsers, again in analogy with peak lists of TF ChIP-seq data. We propose extending the Factorbook framework by adding histone mark ChIP-seq and DNase-seq datasets to show aggregate analysis results. The capacity planned in the previous section will also allow visualization of individual loci.

Author 3/20/2016 4:55 PM
Deleted: Thus for…or histone ma… [57]

Zhiping Weng 3/20/2016 3:22 AM
Comment [16]: gene-centric views, like expression matrices

3.1.4 RNA binding regions and the target genes of RNA binding proteins (RBPs)

The Gravely team in ENCODE3 has released eCLIP-seq data for a large panel of RBPs in two cell lines (K562 and HepG2), as well as RNA-seq data for the corresponding cell types, with and without each RBP being knocked down by shRNAs. The binding motifs of these RBPs are currently being investigated through in vitro binding assays (Bind-and-seq). Working with the Gravely group (see support letter), the DAC will incorporate the following four components into the Encyclopedia.

Author 3/20/2016 4:55 PM
Deleted: in…or the corresponding… [58]

The first component will be two matrices expressing which genes or transcripts the RBPs bind to. Some RBPs prefer to bind mature transcripts with exons spliced out, while others bind to primary RNAs (which include introns). We will construct an RBP binding matrix for each of the two cell types from the eCLIP-seq experiments. The two dimensions index annotated genes/transcripts and RBPs, and the elements of the matrix denote the binding strength of an RBP to a gene. These two matrices can be visualized using a tool we are currently building for this purpose (**Section 3.1.1**) with small modifications (the cell type dimension of the gene expression matrix is changed to the RBP dimension). The second component will be a set of peak lists for RBPs - a catalogue of the peaks in the genes that RBPs bind to. These peaks are typically ~50-nt wide and can be identified by comparing eCLIP-seq datasets with their matching inputs, akin to peak finding in TF ChIP-

Author 3/20/2016 4:55 PM
Deleted: One…he first componen… [59]

seq datasets. We currently do not have a uniform processing pipeline for eCLIP-seq data, so the DAC will work with the Yeo lab (the producer of eCLIP-seq data) to establish one. The third component will be the sequence motifs of RBPs. These can be estimated from the Bind-and-seq experiments and validated by comparing the enrichments of reads in the eCLIP-seq peaks with the input. The fourth component will be the target genes or transcripts of each RBP, which are the genes that show differential expression or differential splicing upon shRNA knockdown of the RBP. These results can be expressed in two matrices (differential expression matrices), similar to the two RBP binding matrices and with the same dimensions. For the RBP knockdown matrices, the entries represent the differential expression of each annotated gene or transcript (expressed as log-fold-change) as measured by the RNA-seq experiments. Similar matrices can be created to quantify differential splicing, using quantities such as percent-spliced-in (PSI).

In summary, for RBPs, the Encyclopedia will contain binding matrices, differential expression matrices, differential splicing matrices, motifs, and peaks. These files all can be downloaded. As we mentioned, these matrices can be visualized using a tool we are currently developing, and the peaks can be visualized using the same methods as those for ChIP-seq peaks. One unique aspect of the RBP peaks is that they are restricted to lie within single genes, which allows gene-based visualization across multiple RBPs. We also propose to build a resource like Factorbook to present the results for RPBs.

3.1.5 Enhancer-target gene links, topologically associated domains (TADs) and compartments.

ENCODE3 (Dekker lab) has performed Hi-C experiments on 12 human cell types and expect to produce data for 8 more. The resolution of this data is only 40 kb, which is not high enough to detect links between promoters and distal regulatory elements, but it is sufficient to detect TADs and compartments. TADs are approximately 100-500 kb in size, and regulatory elements are more likely to regulate genes in the same TAD than genes in a different one. Working with the Dekker lab (see support letter from Prof. Job Dekker), the DAC has calculated insulation scores at varying resolutions for each Hi-C experiment to detect TADs throughout the genome {Lajoie:2015ea}. Genomic compartments represent a large-scale pattern of chromatin organization, typically 1-10 mb in size. Based on these interaction patterns, the genome can be divided into two alternating compartments, referred to as A and B {LiebermanAiden:2009jz}. Compartment A is typically associated with active transcription, while compartment B is often associated with repressed transcription. By transforming Hi-C contact matrices to correlation matrices, these compartments may be identified via principal component analysis, and we have done so for the cell types in each Hi-C experiment. As described in **Section 2.2.6.4**, the DAC plans to establish a uniform processing pipeline as part of the 3D Nucleome Working Group in ENCODE, in collaboration with Hi-C labs in 4D Nucleome consortia. The identified TADs and compartments are provided as BED files for download, and the DAC will provide for visualization via the Juicebox and WashU genome browser. We anticipate that high-resolution Hi-C datasets will be produced in ENCODE4, allowing detection of links between distal regulatory element and their target genes.

3.1.6 Normalizing across experiments and protecting against batch effects

As described in **Section 3.1**, assembling the foundational components of the Encyclopedia requires quantitative comparisons between expression and binding signals across different cell types, between human and mouse data. The DAC will implement robust methods to normalize measurements across experiments and between species. Even after proper normalization, many technical and organizational reasons can still lead to systematic biases. Batch effects are an important type of systemic bias; they are due to, for instance, differences in protocol at different processing centers. The DAC will implement robust methods to minimize batch effects. In this section we showcase results of our methods when applied to ENCODE3 data.

Author 3/20/2016 4:55 PM
Deleted: (except that TF peaks are bigger, several hundred bps).… W… [60]

Author 3/20/2016 4:55 PM
Deleted: Thus or RBP…n summa… [61]

Author 3/20/2016 4:55 PM
Deleted: 3.1.5 Topologically

Author 3/20/2016 4:55 PM
Deleted: cell types. These datasets are at 40 kb… The resolution,…of this … [62]

Zhiping Weng 3/20/2016 1:07 AM
Comment [17]: +rairizarry@gmail.com +kdkorthauer@gmail.com

Hi Keegan, Rafa. Can you proofread this section and fill in the XXX?

Author 3/20/2016 4:55 PM
Deleted: above …n **Section 3.1**, … [63]

Figure 3.1.6a shows the quartile signal levels of H4K3me3 (a histone mark enriched in promoters) at the promoters of all protein-coding genes across 89 cell types. Before normalization (top panel in the figure) we detect a 10-fold difference in the median values across cell types and high variability in the signal distributions. After performing a Principal Component Analysis (PCA) we found that the first principal component is highly correlated with the number of promoters in each sample with signal value 0 (correlation coefficient r=XXX), indicating that the variance is due to technical rather than biological effects. We will implement a normalization approach that will treat the large variability in the number of reported 0s as a missing data problem. The bottom panel of the figure shows the results of applying a preliminary version of our approach.

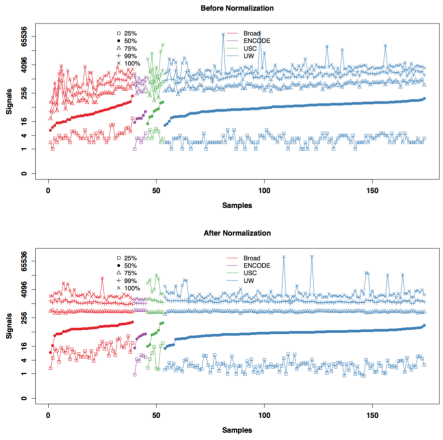
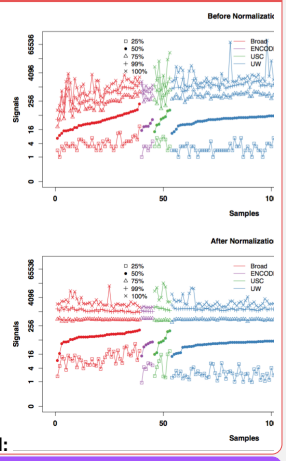


Figure 3.1.6a: Normalization across cell types. For each sample we compute and plot the quartiles, the 99th percentile and the highest value. Colors indicate production labs.

We use the ChIP-seq data of the insulator binding protein CTCF to illustrate our approach. We analyzed CTCF data from 9 human cell types, where the data were produced by several different ENCODE production centers. We first normalized the ChIP signals using the TMM algorithm {Robinson:2010dd} and log-transformed them. Some CTCF peaks displayed different signals in different cell types, representing true biological variability, whereas other peaks show consistently high signals for specific production centers regardless of cell types, suggesting strong batch effects (left panel in **Figure 3.1.6b**). Preliminary analysis shows that GC-bias may account for the batch effect—the signals in some datasets correlate positively with the QC content of the genomic region (correlation coefficient r=XXX) while in other datasets the correlation is equally strong but negative. After we applied some of our previously developed methods {Leek:2010jq} to correct for this batch effect, the samples clustered by cell type instead of processing location (right panel of the figure). In ENCODE4 we will develop methods that minimize GC-bias and other confounding effects. In conjunction with the production centers we will also develop experimental design strategies that will permit us to estimate and correct for batch effects going forward.

Author 3/20/2016 4:55 PM
Deleted:),… we detect a 10-fold ... [64]

Author 3/20/2016 4:55 PM

Deleted:

Zhiping Weng 3/20/2016 1:08 AM
Comment [18]: I will email Keegan + Rafa a photo of my suggested changes on this figure.

Author 3/20/2016 4:55 PM
Deleted: to minimizing batch effects that may be caused by differences among production centers.… We analyzed ... [65]

Zhiping Weng 3/20/2016 1:02 AM
Comment [19]: Did I get this right? If so, please swap the two panels in Figure 3.1.6b and label on the top "Before Batch Effect Correction" and "After Batch Effect Correction" respectively.

Author 3/20/2016 4:55 PM
Deleted: We…n conjunction with t ... [66]

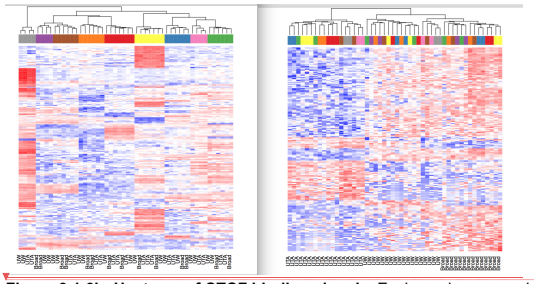
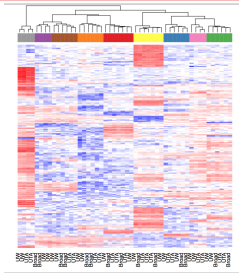


Figure 3.1.6b: Heatmap of CTCF binding signals. Each row is a genomic region (ChIP-seq peak) and each column is a sample. Color blocks on the top of the heatmap indicate cell types. Production centers are indicated at the bottom of the heatmap.

Author 3/20/2016 4:55 PM

Deleted:

3.2 The middle level of annotations in the Encyclopedia

In the middle level of the Encyclopedia we will include annotations that integrate multiple types of experimental data and require interpretation using biological knowledge. First we describe the process of identifying enhancers, then we describe the proposal for identifying repressors and insulators. Last, we describe the roadmap for semi-automated genome annotation using Segway and chromHMM.

Author 3/20/2016 4:55 PM
Deleted: require the integration ... [67]

3.2.1 Enhancers

One essential component of the Encyclopedia is the collection of enhancers, a major type of regulator element that enhances the transcription of a target gene, often at a distance and with cell type specificity {Maston:2006hb}. The number of active enhancers that have been experimentally tested (in transgenic assays in a single mammalian tissue) remains small, and our knowledge of how these enhancers become active in each cell type remains surprisingly incomplete {Visel:2007jw}{Shlyueva:2014ey}. Computational predictions of active enhancers have typically focused on identifying their associated properties (such as inter-species conservation) and identifying clusters of transcription factor binding sites {Hallikas:2006jr}. They have also attempted to discover their epigenetic signatures associated, such as nucleosome positioning, DNase-I hypersensitivity, histone marks associated with active regulatory regions of the genome, and/or eRNA transcription {ENCODEProjectConsortium:2012gc}{RoadmapEpigenomicsConsortium:2015gq}. It is challenging to integrate these disparate features to obtain a more accurate region predictions.

Author 3/20/2016 4:55 PM
Deleted: elements…lement that c ... [68]

Author 3/20/2016 4:55 PM
Formatted: Not Highlight

Author 3/20/2016 4:55 PM
Deleted: , 2006 16719718}. ... [69]

In an effort to stimulate this integration of genomic and epigenomic data, the DAC organized a Consortium-wide blind test of enhancer prediction algorithms (**Section 2.3.1**). Len Pennacchio's team performed mouse transgenic assays on roughly 80 putative enhancer regions identified by their H3K27ac signals in the forebrain and heart at the E11.5 stage of mouse embryos, but withheld the experimental results until the DAC had solicited computational predictions from the participating groups and submitted these predictions to the DCC. Only then did the DAC evaluate these predictions and release the evaluation and experimental results to the participating groups. The DAC found that for both the forebrain and the heart, sophisticated computational methods performed better than simply calling H3K27ac peaks, but the leading computational methods differed between the two cell types. Another important finding was that the H3K27ac peaks called using the uniform processing pipeline developed by the DAC significantly outperformed the peaks called by the generic algorithm that was used to select the 80 putative regions initially.

Author 3/20/2016 4:55 PM
Deleted: computational efforts in predicting enhancers by integrating ... [70]

In an attempt to develop an unsupervised method for detecting enhancers in a cell type-specific manner across as many ENCODE cell types as possible, in both human and mouse, the DAC undertook a systematic

Author 3/20/2016 4:55 PM
Deleted: defining…etecting enhar ... [71]

evaluation of DNase, histone mark (H3K27ac, H3K9ac, H3K4me1, H3K4me2, H3K4me3, H3K9me3, and H3K27me3), and DNA methylation data on four mouse cell types (midbrain, hindbrain, neural tube, and limb). These were the cell types with sufficient numbers of regions that have been tested using mouse transgenic assays. DNase and H3K27ac performed equally well on average if are used individually, and much better than several other histone marks (H3K4me1, H3K9ac, H3K4me3, H3K27me3, and H3K9me3) and DNA methylation. Yet DNase and H3K27ac performed best on different cell types. The DAC combined DNase and H3K27ac into one average ranking scheme which consistently outperformed DNase or H3K27ac alone across all four cell types. Another round of 150 regions have been newly identified based on the DNase-H3K27ac average rank and are currently being tested by the Pennachhio lab.

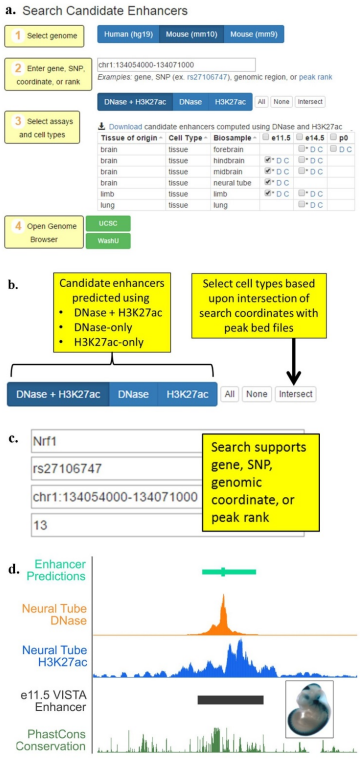


Figure 3.2.1A. An candidate enhancer visualizer. a. Four steps in using the visualizer. **b.** The user can choose the datasets depends on data availability. **c.** The user can choose a genomic region by specifying a gene, a SNP, the region's coordinates, or the rank of a predicted enhancer. **d.** A top ranked candidate enhancer coincides almost completely with a VISTA region previously tested positive in neural tube.

The DAC subsequently implemented the DNase-H3K27ac average rank across 52 human cell types and 20 mouse cell types for which both DNase and H3K27ac data are available. The DAC has also implemented a web server (http://zlab-annotations.umassmed.edu; **Figure 3.2.1A**) that allows users to query candidate

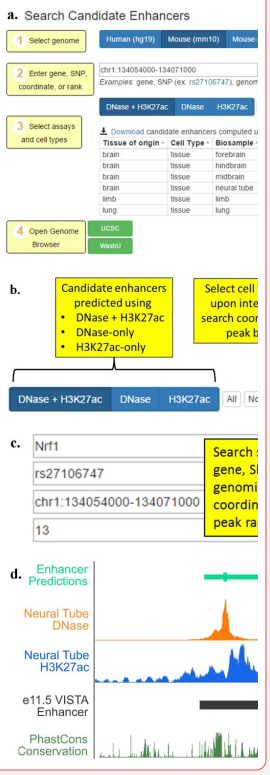
Author 3/20/2016 4:55 PM
Deleted:),

Author 3/20/2016 4:55 PM
Deleted: they

Author 3/20/2016 4:55 PM
Deleted: ,

Author 3/20/2016 4:55 PM
Deleted: claimed victories for

Author 3/20/2016 4:55 PM
Deleted:



Author 3/20/2016 4:55 PM
Deleted: developed

Author 3/20/2016 4:55 PM
Deleted: which

enhancers by coordinates, nearby genes, or SNPs and visualize them in the UCSC genome browser or the WashU epigenome browser, along with the supporting DNase and H3K27ac signals. The example figure shows a top enhancer prediction in the neural tube, which coincides almost perfectly with a VISTA enhancer (**Figure 3.2.1A.d)**.

Going forward, the DAC will continue to develop computational algorithms and visualization methods to identify candidate enhancers in all ENCODE cell types. We will continue to work with Dr. Pennacchio to iterate and validate our enhancer prediction algorithm (see support letter from Dr. Pennacchio). One natural extension is to develop a hidden Markov model which explicitly incorporates the interplay between DNase and H3K27ac signals - for instance, DNase peaks often occur in the troughs of H3K27ac signals. Meanwhile, a number of massively parallel reporter assays (MPRAs) such as STARR-seq and FIREWACh have been conducted recently {Arnold:2013di}{Vanhille:2015ho}{Murtha:2014bf}{Patwardhan:2012hy}, providing us with much larger datasets with which to study the precise combinations of genomic and epigenomic features that are most predictive of active regulatory regions. Prof. Nadav Ahituv is submitting a Characterization Center proposal to perform MPRAs on 100,000 putative regulatory elements inserted into the genome using a lentivirus-based approach. He has expressed a strong interest in collaborating with the DAC (see support letter from Prof. Ahituv).

One of the best-performing methods for predicting forebrain enhancers in the ENCODE challenge (described earlier in this section and in **Section 2.3.1**) was developed by a DAC lab (Gerstein). It use a pattern recognition algorithm called Matched Filter \cite{Book - Papoulis, S. U. Pillai, Probability, Random Variables, and Stochastic Processes 4th Edition} to identify a particular shape of H3K27ac ChIP-seq signal (two nucleosome-sized peaks flanking a nucleosome-sized valley) in the presence of noise. In ENCODE4, we plan to extend this method to incorporate additional histone marks and DNase-seq and transcriptions of enhancer RNA (eRNA). The algorithm will be trained using enhancers identified by MPRAs. In addition, we will also develop a framework to combine the results from Matched Filter on different epigenomic datasets to make more accurate predictions and understand the properties that most likely contribute to enhancer establishment, maintenance, and recognition. We will use feature selection methods within a supervised learning framework, such as random forests, support vector machines, Bayesian networks, and regression methods. While sequence information will not be used to make these predictions, we can test whether the predicted regions contain evolutionarily conserved clusters of TF motifs and identify TF families (with known motifs) that are associated with these regulatory regions (in a cell type-specific manner). We can then annotate both the regulatory regions active within each cell type, and the likely TF binding site within these regions. Our ultimate goal is not only the systematic annotation of each functional element but also obtaining new mechanistic insights into the interplay between different elements, and we will work closely with the AWG and Consortium members to design computational and experimental tests to evaluate the role of the most informative features.

Author 3/20/2016 4:55 PM
Deleted: such … top enhancer pr… [72]

Author 3/20/2016 4:55 PM
Deleted: further develop…terate a… [73]

Author 3/20/2016 4:55 PM
Deleted: During …ne of the best- … [74]

Author 3/20/2016 4:55 PM
Formatted: Highlight

Author 3/20/2016 4:55 PM
Deleted: ,…and DNase-seq and … [75]

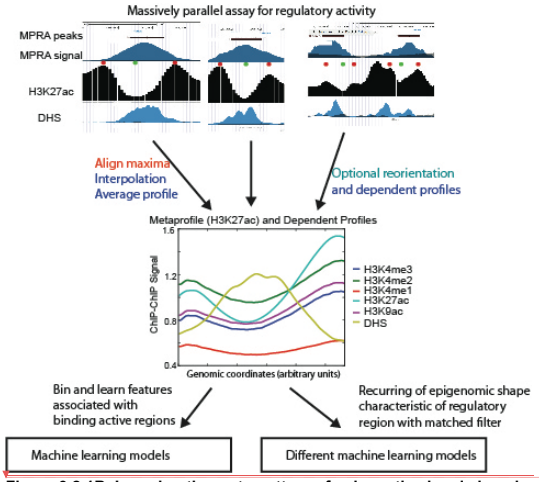
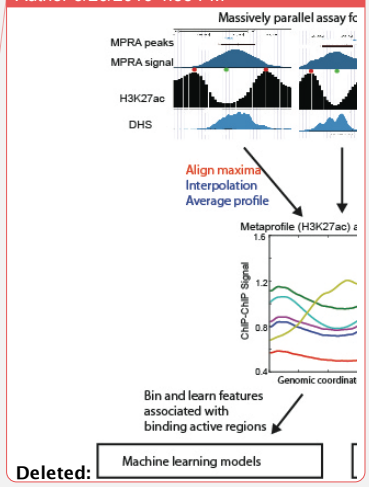


Figure 3.2.1B. Learning the meta-pattern of epigenetic signals in enhancers. The enhancers identified by massively parallel regulatory assays are aligned by the position of the maximal H3K27ac ChIP-seq signal and averaged to calculate the shape (meta-pattern) of enhancers. We can also calculate the meta-patterns on MNase-seq, DNase-seq, and other histone ChIP-seq datasets by simultaneously applying the same transformation to these datasets. We will combine the matched filter scores for each chromatin mark within a machine learning framework and this can improve the accuracy of the predictions.

Author 3/20/2016 4:55 PM



Deleted:

3.2.2 Insulators and repressors

While there are no high-throughput techniques specifically designed to identify insulators or repressors, there are low- and medium-throughput methods {Petrykowska:2008io}. We expect that many of the MPRAs can be modified to find these elements, e.g., by measuring a decrease in the reporter signal. If ENCODE4 has access to production or characterization centers that can produce data on insulators or repressors, then we can develop approaches for predicting insulators and repressors that are similar those for the prediction of enhancers.

Author 3/20/2016 4:55 PM

Deleted: technique…echniques ... [76]

3.2.3 Semi-automated genome annotation (SAGA)

In the previous two subsections we have described approaches for directly identifying regulatory elements such as enhancers, insulators and repressors by identifying genomic and epigenomic features that are predictive of each element and building computational models using these features. A complementary approach is to simultaneously partition and label the human genome based on a large number of genomic and epigenomic features, so that genomic regions with similar values for these features are assigned the same label. This approach, called semi-automated genome annotation (SAGA), uses unsupervised machine learning methodology without any prior knowledge of annotation information; it is only semi-automated because the labels produced by the model must be interpreted in a human post-processing step, and assigned to promoter, enhancer, insulator, repressed, or quiescent regions of the genome. These states show distinct functional enrichments, confirming their respective biological functions. They also provide a chromatin context for understanding binding of sequence-specific factors, motif enrichments, and other diverse functional elements, such as origins of replication or insulator regions.

Author 3/20/2016 4:55 PM

Deleted: silencers,…epressors by ... [77]

The DAC has developed a number of SAGA methods. During the pilot phase of ENCODE, we combined a hidden Markov model (HMM) with wavelet smoothing to produce annotate of the pilot regions as "active" or "repressed" {ENCODEProjectConsortium:2007fu}{Thurman:2007ft}, and we produced a software tool capable of automatically creating annotations using any number of distinct labels {Day:2007by}. A variety of SAGA methods have since been described, employing HMMs with a flat {Filion:2010dw}{Kharchenko:2010gm} or hierarchical structure {Jaschek:2009fr}{Larson:2013gw}{Biesinger:2013hs}, or generalizing the HMM to a hierarchical change-point model {Lian:2008kz}. During the second phase of ENCODE, two research groups within the consortium (Noble and Kellis) independently developed SAGA algorithms, ChromHMM {Ernst:2012ii} and Segway {Hoffman:2013im}, which employ closely related probabilistic models that offer multiple important advantages; they have efficient algorithms for carrying out inference, and a modeling paradigm in which the internal variables have well-defined semantics. The two approaches are complementary, in the sense that ChromHMM aims for a birds-eye view of the data, opting to collapse each 200 bp of data to a single Boolean value, whereas Segway provides a more detailed view, operating on the raw data at up to its native 1 bp resolution. These methods continue to be widely used, see, for example, the recent Roadmap Epigenomics Consortium papers (**Figure 3.2.3A)**. They are both currently being applied systematically to the combined compendium of ENCODE3 and Roadmap Epigenomics data in both human and mouse.

As part of the current annotation effort, our Segway annotation incorporates several recent improvements, including the ability to train on the entire human genome using a minibatch training mode, and the modeling of emission distributions within each state of the model using a mixture of Gaussians (rather than a single Gaussian). In addition, we have developed and validated a method for automatically assigning human-interpretable labels to each annotation. The method leverages the fact that a large number of manually interpreted SAGA annotations have been published. We trained a supervised classification algorithm to automatically assign one of ten labels ("Enhancer," "Promoter," etc.) to an integer label produced by Segway, using as features the pattern of histone modifications and gene enrichments associated with the labeled segments. In a cross-validated setting, this approach achieves an accuracy of 98.3%, and similar learned labels tend to cluster in a 2D projection.

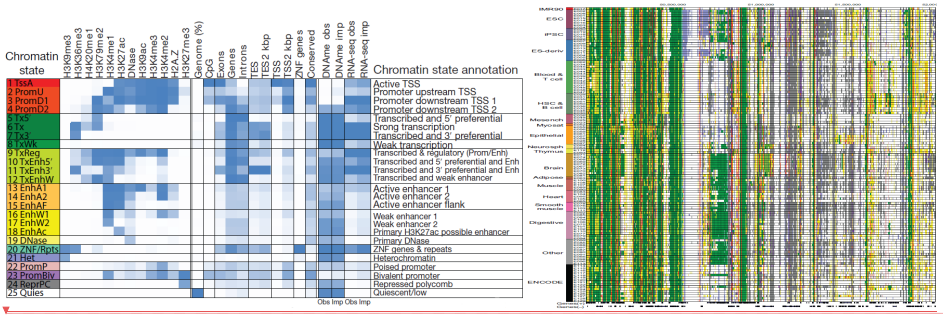


Figure 3.2.3A. Application of ChromHMM to analyzing Roadmap Epigenomic Consortium data. Left: Properties of 25 chromatin states. Right: the chromatin states in an example genomic region across 126 epigenomes (rows), showing regions with conserved states and cell type-specific states.

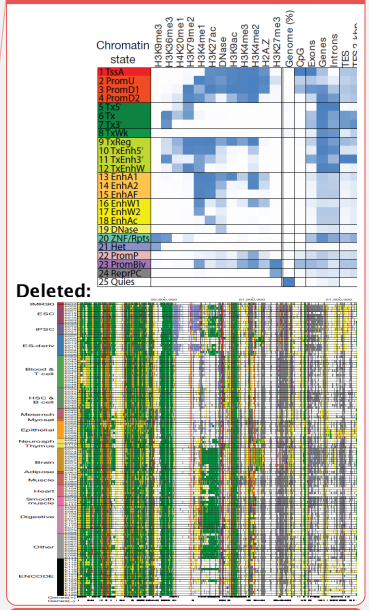
During the next phase of ENCODE, we intend to improve upon Segway in three significant ways. Many of these improvements will be adapted to ChromHMM with adjustment. Ultimately, these semi-automated learning

Zhiping Weng 3/20/2016 8:52 AM
Comment [20]: See this PDF file for the references in this section: https://www.dropbox.com/s/c0dnz91igq3cujc/encode-dac.pdf?dl=0

Author 3/20/2016 4:55 PM
Deleted: a two-label annotation… ... [78]

Author 3/20/2016 4:55 PM
Deleted: ….. In addition, we have ... [79]

Author 3/20/2016 4:55 PM
Deleted:



Author 3/20/2016 4:55 PM
Deleted: ,…we intend to improve ... [80]

methods will provide a scalable way to integrate all of the ENCODE data, along with human annotations, to provide a joint annotation that summarizes our knowledge of the functional landscape of the human genome in various cell and tissue types at multiple scales.

First, we will augment Segway with learned networks representing pairwise relationships between cell types and assay types. The idea is to down-weight correlated contributions from related assay types and, in the context of multi-cell type annotation, to encourage related cell types to have similar annotations. The feasibility of this type of network learning has been demonstrated in recent work on learning complex chromatin states {Zhou:2016bv}. A more scalable approach has also been outlined for large sets of histone modification assays {Lundberg:2015fr}. The Graphical Models Toolkit (GMTK) software, used in Segway's implementation, will allow us to easily incorporate priors derived from these learned networks into Segway. Thus, we can avoid the mean field approximation required by a previous method that infers a tree over cell types {Biesinger:2013hs}.

Second, we will extend Segway to implement a hierarchical state space. Our current Segway annotation effort is being performed at three different resolutions, 10 bp, 100 bp and 10 kbp. These complementary annotations capture phenomena at different scales, from fine-scale phenomena such as regulatory elements up to large-scale domains {Libbrecht:2015ck}. Several published reports describe hierarchical approaches that jointly capture phenomena at different scales {Knijnenburg:2014ke}, including methods that employ hierarchical probability models akin to the one used by Segway {Jaschek:2009fr}{Larson:2013gw}. We will implement our previously described conjugate priors approach 35 (), in which the parameters of a substate are generated by the emissions from its parent state. Each top-level label is associated with some number of sub-labels; a top-level label might be "enhancer," whereas a low-level label might be "H3K27ac region upstream of enhancer TF binding site." We will also experiment with applying hierarchical models in the context of multi-cell type annotation. In such a setting, the model might learn a latent vector of "potential activity types" at each position that specifies a distribution over possible labels. Doing so would discourage the algorithm from calling a position in one cell type as, say, "promoter" if no other cell types show evidence for having a promoter in that position.

Third, we will develop a variance stabilization method to cast the inputs to Segway into interpretable units. In general, the output of a genomics assay—the number of reads mapping to a given genomic position—has no natural units. In particular, a difference between 0 and 30 reads may have a very different statistical importance from a difference between 1,000 and 1,030 reads. This property is due to a relationship between the mean of the data-generating process (i.e. the genomics assay) and its variance. In practice, the different mean-variance relationships among the heterogeneous inputs to Segway make the statistical modeling problem much harder.

Currently, to attempt to normalize for the mean-variance relationship of the read counts, we preprocess the inputs using an inverse hyperbolic sine transform. However, there is no theoretical basis for using this particular transformation. Theoretically, given a known mean-variance relationship, a dataset can be put into interpretable units using the variance-stabilizing transformation, where $\sigma(x)$ is the standard deviation of a variable with a mean of x. The resulting signal is in units of standard deviation, so it has the useful property that all data points have a 95% confidence intervals of ±2 units. We will therefore investigate methods to learn the mean-variance relationship of functional genomics datasets using the variation among biological replicates. This problem involves optimizing an objective of the form, whereas most existing approaches for learning functions—such as using splines or Gaussian processes—are designed for a squared-error objective of the form. We will then use the learned mean-variance relationship in combination with a negative binomial distribution to transform the

Author 3/20/2016 4:55 PM
Deleted: , at multiple scales,

Author 3/20/2016 4:55 PM
Deleted: here…is to down-weight … [81]

Author 3/20/2016 4:55 PM
Deleted: ,… 10 bp, 100 bp and 10 … [82]

Author 3/20/2016 4:55 PM
Deleted: practise

Author 3/20/2016 4:55 PM
Deleted: sequencing …ead count … [83]

Zhiping Weng 3/20/2016 8:54 AM
Comment [21]: missing equation

Zhiping Weng 3/20/2016 8:54 AM
Comment [22]: missing equation

Zhiping Weng 3/20/2016 8:55 AM
Comment [23]: missing equation

Author 3/20/2016 4:55 PM
Deleted: place

Segway inputs into interpretable units. We expect this transformation to significantly improve the robustness and interpretability of the resulting models and annotations.

3.3 The top level of annotations in the Encyclopedia

We propose three components to be included in the top level of the Encyclopedia: the target genes of enhancers, the target genes of transcription factors, and allele specific binding or transcription events. This is not an exhaustive list and will be augmented upon formation of ENCODE4. For example, we can include the target genes of silencers if their experimental detection becomes reliable and there will be sufficient experimental data to train computational methods. The key feature of annotations in the top level of the Encyclopedia is that they require the most integration. Hence, these annotations are the furthest away from the underlying raw experimental data, yet the annotations are of broad interest which warrants their inclusion in the Encyclopedia. Furthermore, these annotations are key components for building regulatory networks.

3.3.1 Linking enhancers to their putative target genes

Many DAC labs have done significant work on linking enhancers to the promoters of their target genes. We correlated histone marks between enhancers and promoters across multiple cellular context to identify links, and search for additional features supporting these enhancer-promoter assignments {Ernst:2011kw}. We have also used ENCODE ChIP-seq data to identify enhancers and connect them to target transcripts {Yip:2012cd}. Other ENCODE labs have also developed correlation-based methods {Thurman:2012fe} {Sheffield:2013di}. Methods that are based on static correlations of histone marks or DNase signals and gene expression across multiple cell types are potentially limited in their abilities in detecting cell type specific and dynamic links, as well as non-linear behaviors. Recent work developed more sophisticated metrics than correlation {Marstrand:2014jg}{Corradin:2014eb}. Alternatively, supervised approaches (e.g., {He:2014gg}) train computational models using links detected in Hi-C and ChIA-PET experiments {ref}.

The lack of reliable experimental data on enhancer-gene links, especially cell type specific links which are the most interesting in terms of understanding the dynamics of transcriptional regulation. The chromatin conformation type of experiments are typically of low resolution. In **Section 3.1.5** we mentioned ENCODE Hi-C data from the Dekker lab on 12 cell types, but they are at 40 kb resolution, too low for reliably assigning enhancer-gene links. One study reported 1 kb resolution Hi-C maps on GM12878 and 5-kb resolution maps for eight other cell types {Rao:2014eo}. We anticipate that more high-resolution maps will become available from production centers of the 4D Nucleome consortium. In addition, modified Hi-C or ChIA-PET technologies that perform promoter capture can significantly improve the detection of enhancer-gene links. Another source of experimental data are expression quantitative trait loci (eQTLs). We expect consortia like GTEx and PsychENCODE to produce substantial eQTL data. The DAC will take advantage of these new data to train algorithms for detecting target genes.

The DAC recently evaluated correlation-based methods linking methods using promoter capture Hi-C data on GM12878 cells {Mifsud:2015en}. Requiring enhancers to be linked to only one gene, we created training, testing, and validation datasets from this Hi-C data resulting in over 13k positive links and 291k negative links in each dataset. However, the small percentage of positive links in each dataset (< 5%) will make classification difficult due to imbalances in the classes. Using our curated datasets, we found that correlation based methods using DNase or H3K27ac signal in GM12878 perform poorly (area under the ROC curve or AUROC was 0.60). We then developed a random forest method to incorporate additional features: distance from gene to enhancer, DNase signals in promoter and in enhancer, H3K27ac signals in promoter and in enhancer, gene expression in GM12878, and our previously calculated DNase and H3K27ac Pearson

Author 3/20/2016 4:55 PM
Deleted: , hence they

Author 3/20/2016 4:55 PM
Deleted: these

Zhiping Weng 3/20/2016 11:26 PM
Comment [24]: Add a couple of sentences of why is biologically important to link enhancers to target genes. Interpretation of GWAS variants.

Author 3/20/2016 4:55 PM
Formatted: Underline

Author 3/20/2016 4:55 PM
Deleted: PMID: 21441907

Author 3/20/2016 4:55 PM
Deleted: PMID: 22950945

Author 3/20/2016 4:55 PM
Deleted: 2012 PMID: 22955617

Author 3/20/2016 4:55 PM
Deleted: GR 2013, PMID: 23482648}.

Author 3/20/2016 4:55 PM
Deleted: mark

Author 3/20/2016 4:55 PM
Deleted: and Storey, 2014, PNAS, PMID: 24469817} {

Author 3/20/2016 4:55 PM
Deleted: GR 2014 PMID: 24196873}.

Author 3/20/2016 4:55 PM
Deleted: 2014 PNAS PMID: 24821768}

Jayanth Krishnan 3/21/2016 1:01 AM
Comment [25]: This sentence needs to be revised

Author 3/20/2016 4:55 PM
Deleted: 2014 Cell, PMID: 25497547}.

Zhiping Weng 3/20/2016 11:00 PM
Comment [26]: +jjem0808@gmail.com Jill Can you write one paragraph summarizing your recent results? Please include: 5% of the enhancer-promoter links are positives, hence the problem is difficult, the performance of correlation, and the improved performance of RF.

Author 3/20/2016 4:55 PM
Deleted: The DAC recently evaluated correlation-based methods using the capture Hi-C data on GM12878 cells {Mifsud ... [84]

correlation coefficients. With this model, we observed a dramatic increase in performance with an AUROC of 0.84. We also observed that the most important feature in the model was distance between the enhancer and gene's TSS. We are currently working on expanding this model by adding new features. For example, by incorporating the signal of additional histone marks across promoter and enhancers (H3K4me1/2/3, H2AFZ, H3K36me3, H4K20me1, H3K79me2, H3K9me3, and H3K27me3), the performance improves with an AUROC of 0.88. Moving forward, we hope to continue on improving our method as well as develop a method using a core set of features that can be applied across many ENCODE cell types.

The DAC will continue to develop innovative methods to evaluate the role of insulator regions, spatial proximity, chromatin marks, binding of complementary or synergistic TFs, motif content, and association with general and specific TFs in determining enhancer-promoter specificity. We will develop a probabilistic framework to integrate structural, static and dynamic genomic information. The sequence features are denoted by K-mer profile and co-occurrence matrix, while the dynamic features include DNase I, histone modification and TF binding information. We do the transformation for histone mark/DNase signals and extract informative features that can tell positive from negative datasets. Meanwhile, gene expression variance, is also integrated as an additional feature into a statistical model to predict enhancer gene linkage. For example, a TF can connect to their target genes if it bind to the enhancers and promoters of the target genes. To incorporate the eQTL effect, we will incorporate an eQTL score as a part of spatial structure feature to predict enhancer gene linkage. The LD region of trans-eQTL will be considered, and the mediation effect of a trans-eQTL acting as cis-eQTL on the proximate gene will be excluded if the proximate gene has high correlation with its distal target genes.

After careful evaluation of computational methods on a comprehensive set of experimental data, we will develop a resource of enhancer-gene links in specific cell types. These will be lists of links in the BED format, as the links derived from Hi-C experiments described in **Section 3.1.5**. We will develop the tools to visualize them using the WashU browser, along with the raw experimental data that support the predictions, in much the same way as our enhancer visualizer (http://zlab-annotations.umassmed.edu). In addition, the Hi-C or other experimental data in the same cell type, if available, will also be displayed.

3.3.2 Linking TFs to their putative target genes

One major goal of performing a TF ChIP-seq experiment is to identify the target genes of the TF of the cell type of interest. We will use the same logic as linking enhancers to target genes to link the TF peaks to target genes. This approach requires that there are histone mark ChIP-seq or DNase-seq data in the same cell type. To complement this approach, the DAC will also develop more generic approaches that take advantage of the rapid accumulation of ChIP-seq, DNase-seq and ATAC-seq data in human and mouse, both in ENCODE and in the public. As of Dec 31, 2015, there are ~23K sets of these data available to the public, including ~10k TF ChIP-seq datasets and excluding single-cell samples. The Liu lab of the DAC has been systematically collecting and curating these data in the Cistrome collection (http://cistrome.org/db/). In addition, we will integrate TF ChIP-seq data in the Cistrome with large-scale gene expression cohorts from the Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression Project (GTEx). Previously we developed a probabilistic model, referred to as target identification from profiles, that identifies a given TF's target genes based on ChIP-seq data {Cheng:2011bd}. For each TF, our model builds a characteristic, averaged profile of binding around the TSS and then uses this to weight the sites associated with a given gene, providing a continuous-valued 'regulatory' score relating each TF and potential target. Similarly, we calculate a regulatory potential (RP) which is the sum of significant binding (as measured by ChIP) weighted by an exponential decay as a function of their distance to the transcription start site (TSS) of a putative target gene {Wang:2013kn}.

In ENCODE4, we will build a probabilistic model to incorporate three factors in determining the strength of regulation between a TF and a putative target gene: the level of TF expression in the cell type of interest, the correlation of gene expression between the TF and the gene across multiple cell types, and the level of TF binding near the gene computed using the aforementioned regulatory potential. The decay rate in the regulatory potent, empirically estimated based on long-range chromatin interactions and the level of differential gene expression observed when transcription factors are perturbed, is set so that a peak 10kb from the TSS contributes half of that at the TSS. Better Hi-C data from the ENCODE and 4D Nucleome consortia will generate better resolution and condition-specific topologically associated domain (TAD) information, which we could incorporate into the RP so binding in one TAD domain would not influence the RP of a gene in the next TAD domain. We will extensively benchmark our approach using RNA-seq data on cell types for which a TF is knocked down. Note that some components of this model can be tested using the collection of ENCODE3 data on RNA binding proteins (**Section 3.1.4**), with their target genes determined by RBP knock down and then RNA-seq. The difference is that RBPs bind locally to genes and TFs often regulate from a distance. In addition, RBPs exert their impact by regulating expression and splicing. In ENCODE4, we will compile a list of target genes for each TF with ChIP-seq data. These genes will be included into Factorbook for visualization in a cell type specific manner, using tools described in **Section 3.1.2**.

[Goal: Manolis will turn this into at most 5.5 Word pages]

Aim 4. Assessing quality and utility of the ENCODE data and providing feedback to NHGRI and the Consortium. In Aim 4, we will develop and apply computational and statistical methods to assess the quality and utility of ENCODE datasets in a systematic and unbiased way, including: (1) data-type specific QC measures including for ChIP-seq, RNA-seq, ATAC-seq; (2) data-type agnostic measures of dataset quality; (3) measures of cumulative coverage and dataset completeness; (4) imputation-based and submodular-optimization-based methods for assay prioritization; (5) methods for cell type prioritization.

4.1. Data-type specific measures of dataset quality: The DAC will continue to work with the AWG and the PIs to standardize metrics of dataset quality, incorporating existing metrics (many of which were developed by the DAC in collaboration with ENCODE Working groups) and developing new metrics per Consortium needs.

QC for ChIP-seq: We use metrics based on antibody validation, replicate consistency, replicate rank consistency, strand cross-correlation analysis (**Fig. 4.1a-c**), and fraction of reads in peaks. We adopt a number of quality control (QC) measures to evaluate the quality of each ChIP-seq sample, mostly following ENCODE ChIP-seq QC standard with some additional criteria. They include sequence qualify (FASTQC), mapping quality (uniquely mapped read ratio), library complexity (PCR bottleneck coefficient), ChIP enrichment (number of ChIP-seq peaks passing false discovery rate and fold enrichment cutoffs), signal to noise (fraction of reads in peaks), regulatory region (overlap with union of all DNase-seq peaks), evolutionary conservation (PhastCons), and motif (enrichment in peak summit). Researchers could search ChIP-seq sample of interest, evaluate data quality on the site, browse the ChIP-seq signals on genome browsers, and download peaks and signals to their own computer for downstream analysis.

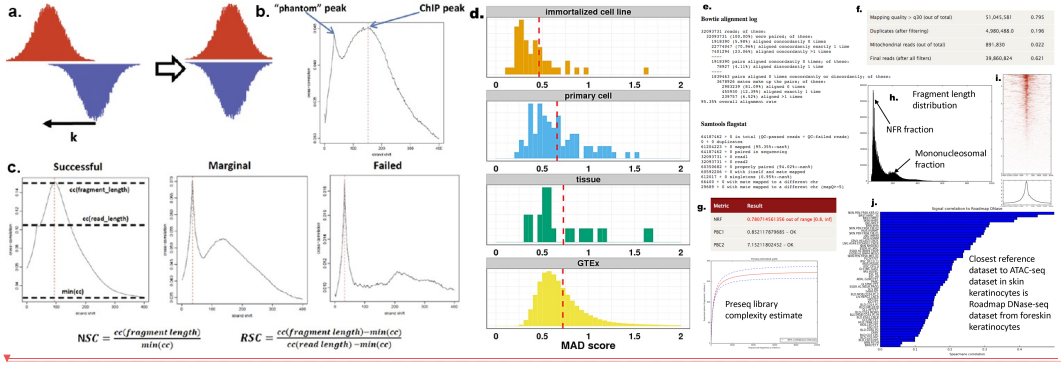
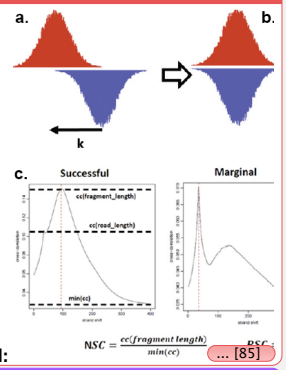


Figure 4.1. Data type-specific quality control critera. (a-c) Criteria for assessing the quality of a ChIP-seq experiment. **(a)** Cross-correlation analysis: reads are shifted in the direction of the strand they map to by an increasing number of base pairs and the Pearson correlation between the per-position read count vectors for each strand is calculated. Read coverage as wigglegram is represented. **(b)** Two cross-correlation peaks are usually observed in a ChIP experiment, one corresponding to the read length (phantom peak) and one to the average fragment length of the library. **(c)** Absolute and relative height of the two peaks help determine the success of a ChIP-seq experiment. A high-quality IP is characterized by a ChIP peak that is much higher than the phantom peak, while often very small or no such peak is seen in failed experiments. **(d)** RNA-seq quality control metric distribution for MAD scores (a measure of inter-replica reproducibility). The MAD score is lower for immortalized cell lines than primary cells than tissues for ENCODE data, where the distribution is comparable to that of GTEx. The dashed red line indicates the median. **(e-j) ATAC-seq QC metrics.** (**e**) Mapping statistics from Bowtie and Samtools (f) Read Filtering Statistics (**g**) Library complexity measures and Preseq library complexity estimate (**h**) Fragment length distribution including estimation of nucleosome free fraction (NFR) and mononucleosomal fraction (**i**) TSS enrichment plot (**j**) Estimation of nearest cell type based on similarity to reference DNase-seq datasets.

Mark Gerstein 3/19/2016 12:36 AM
Comment [27]: +baikang.pei@gmail.com +manoli@mit.edu +Zhiping.Weng@umassmed.edu "We Team" for aim 4

Author 3/20/2016 4:55 PM


Deleted: ... [85]

Zhiping Weng 3/19/2016 7:39 PM
Comment [28]: The figure panels needs to be relabeled to a-j

Zhiping Weng 3/21/2016 2:55 AM
Comment [29]: +purcaro@gmail.com Michael can you make one file out of the panels in this figure? Please make sure to preserve the quality. You need to p ... [86]

Michael Purcaro 3/21/2016 1:47 AM
Comment [30]: changed!

Zhiping Weng 3/21/2016 1:48 AM
Comment [31]: It is a big grainy. Can you do better?

Michael Purcaro 3/21/2016 2:22 AM
Comment [32]: How about this? I bumped resolution to 600dpi.

Michael Purcaro 3/21/2016 2:33 AM
Comment [33]: actually, the graininess was from copying and pasting the image; inserting the file improved sharpness

Zhiping Weng 3/21/2016 2:55 AM
Comment [34]: still looks kind of bad. Can you tell the difference?

QC for RNA-seq: For RNA-seq, we use metrics based on sequencing depth, alignment rate, duplicate read rate, compositional biases, ncRNA content, intronic vs. exonic coverage, positional bias, and coverage continuity. We have learned that different reproducibility metrics are needed when comparing replicates from isogenic samples, such as those originated from cell lines, vs. when comparing replicates of more different biological source, such as those from tissues obtained in different individuals. To develop the appropriate metrics threshold, the DAC is extensively analyzing data produced by the GTEx project, where RNA-seq is produced in tissue samples from multiple donors (**Fig. 4.1d**). The aim is to produce the appropriate background distribution for the metrics depending on the biological source.

QC for DNA accessibility. For DNase-seq and ATAC-seq, we use several QC measures based on enrichment, reproducibility and detection of potential sample swaps or mislabeling. (1) We use mapping logs with percent aligned reads, percent reads above MAPQ 30, percent duplicate reads, mitochondrial fraction and the final number of usable reads, and QC bias distribution (**Fig. 4.1e,f**). (2) We compute measures of library complexity, including PCR Bottleneck coefficient (PBC) and non-redundant fraction (NRF){Landt:2012cl}, and use PRESEQ{cite} and PICARD to estimate the true library complexity (**Fig. 4.1g**). (3) We use the fraction of reads in the nucleosome free fraction (NFR) relative to the mononucleosomal fraction (180-220 bp) using a mixture of gamma distributions (**Fig. 4.1h**). (4) We use the fraction of reads in all peaks, promoter peaks, non-promoter peaks, DNase-overlapping peaks from ENCODE and Roadmap Epigenomics, TSS bias, and fraction in ENCODE blacklist regions (**Fig. 4.1i**). (5) We use the nearest reference cell type/tissue to detect mislabeled samples: We compare a target ATAC-seq dataset to reference DNase-seq datasets from ~300 diverse cell types and tissues from ENCODE and the Roadmap Epigenomics project. Specifically, we obtain a universal set of DNase-seq peaks (a union) across all cell types. We then compute the correlation of the ATAC-seq signal to the signal from each reference DNase-seq dataset across the universal peak set. We rank the reference cell types in descending order to correlation. The top 10 most similar reference samples should be biologically similar to the target cell-type/tissue of the ATAC-seq sample. This analysis allows us to identify potential gross sample swaps or mislabeling of datasets (**Fig. 4.1j**).

QC for personal genomics. Specifically for personal genomes, we have developed RNA-seq and ChIP-seq metrics within the diploid framework{Chen et al. Nat. Commun., in press} that quantify the deviation of the read distribution of allelic ratios across all heterozygous sites from the binomial distribution. An overdispersion metric on the fraction of reads mapped to the reference allele can identify datasets with unusual allelic ratio distributions, that indicate lower QC issues such as uneven or sparse read coverage.

4.2. Data-type agnostic measures of dataset quality. In addition to these data type-specific metrics, we will develop and apply general methods for assessing experiment quality.

Imputation-based QC metrics. In particular, we have demonstrated the use of systematic imputation of histone marks, DNA methylation, RNA-seq, and DNase datasets based on their correlation with datasets of the same data type and of different datatypes within the same cell type and within other cell types. Briefly, we use the correlation between an experimentally observed data track and an imputed data track (**Fig. 4.2a**), evaluated either across the entire genome ("GWcorr"), or in the 1% of regions that have greatest signal strength ("Match1"). We found that the agreement of a dataset with the imputed signal based on these correlation metrics provides a powerful and unbiased metric for evaluating dataset quality, flagging low-quality datasets that are sometimes missed by other quality metrics (**Fig. 4.2b-d**), including sample or antibody swap problems. We will extend these methods to incorporate additional data types and additional features in the prediction.

Author 3/20/2016 4:55 PM
Deleted: useable

Author 3/20/2016 4:55 PM
Formatted: Highlight

Author 3/20/2016 4:55 PM
Formatted: Highlight

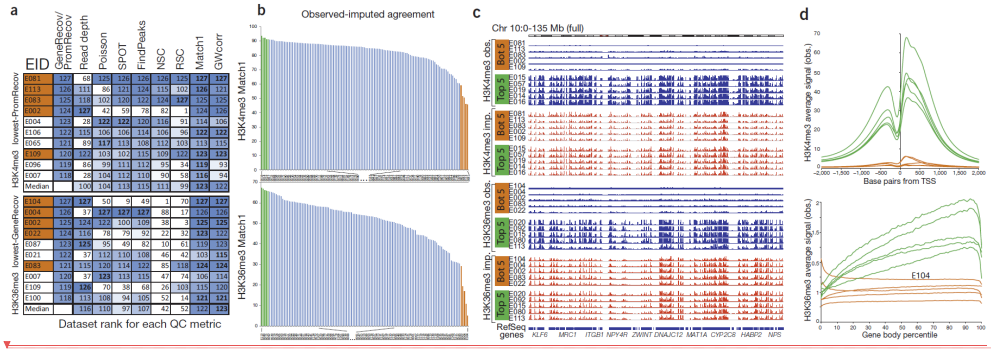
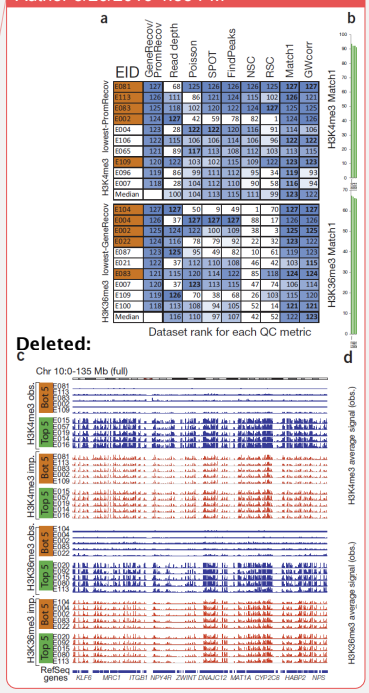


Figure 4.2. Imputation-based QC complements datatype-specific QC metrics. (**a**) Comparison of dataset-specific QC metrics (columns) for ten Roadmap Epigenomics datasets (rows) that show the lowest agreement with gene and promoter annotations, based on H3K4me3 (top) and H3K36me3 (bottom). Each entry shows rank (out of 127) for imputation-based QC (first column), read depth (second column) and each QC metric (Poisson statistic, Signal Proportion of Tags (SPOT), FindPeaks, Normalized and Relative Strand Correlation between forward and reverse strands (NSC and RSC)), and similarity between imputed and observed data, evalauted genome-wide (GWcorr) and in the highest-signal 1% regions (Match1). Orange-shaded EIDs denote the five worst-agreement datasets. (**b**) Distribution of match1 imputation-based QC metric highlights five worst-similarity (orange) and five highest-similarity (green) datasets. (**c**) Observed (blue) and imputed (red) signal tracks for worst-similarity (orange) and best-similarity (green) datasets for chromosome 10. Datasets with lowest imputation QC score show relatively flat signal, indicating successful flagging of low-quality datasets even when several datatype-specific QC metrics failed to detect them. (**d**) Aggregation of observed signal for H3K4me3 surrounding the TSS (top) and H3K36me3 in gene bodies (bottom) for the five best (green) and worst (orange) datasets according to the imputation-based Match1 QC metric, highlighting unusual profiles of flagged, indicating lower quality, even though they were not flagged by traditional datatype-specific QC metrics.

Developing new QC metrics by training on outlier datasets. Definitions of data quality depend on the specific applications for which the data is to be utilized. In the case of RNA-seq, the most common applications appear to be finding differentially expressed genes between two or more conditions, clustering transcripts or samples, and predicting sample types or outcomes. For ChIP-seq of histone marks and sequence-specific factors applications include the location of peaks, regions of differential binding, the inference of chromatin states, and of regulatory motifs. In our attempt to measure quality we will quantify the effect of removing bad quality data on the biological results. We will consider removing: one read from one sequencing run, all data points from mapped to a region across all samples, all data from an sequencing run, all data arising from a specific biological specimen, all data arising from a specific batch of samples, or even all data produced by a particular laboratory. We will identify cut-offs such that removing experiments that do not reach this cut-off improve downstream results. We will leverage the availability of dozens of ENCODE samples to derive QC metrics by jointly analysis the all samples, which can detect problems missing by single-sample quality metrics. For example, multi-dimensional scaling of all ENCODE gene expression data with Spearman correlation >0.9 between replicates shows outlier samples all processed by the same lab (**Fig. 4.3**), providing a benchmark target set for new measures of QC. Indeed, a new metric evaluating the “percent of genes quantified below 0.5 transcripts per million reads” (TPM) can flag these outliers.

Author 3/20/2016 4:55 PM



Deleted:

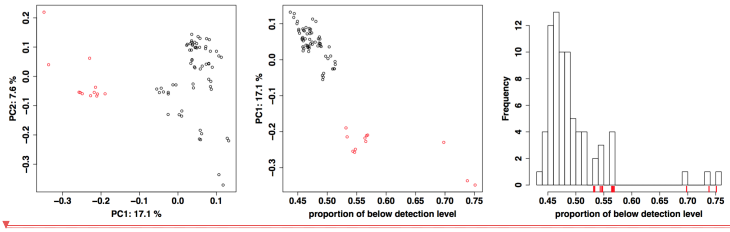


Figure 4.3. Exploiting outlier datasets for QC metric development: Left: Multidimensional scaling plot of ENCODE expression data on primary cells and cell lines with similar starting material for two different processing labs (color). Middle: First principal component (PC1) vs. proportion of genes with quantification below 0.5. Right: histogram of proportion of genes below detection level.

4.3. Measures of dataset completion. We will also develop metrics and methods for evaluating the progress and completeness of the entire ENCODE project, along multiple axes, including: (i) genomic coverage, towards identifying biochemically active nucleotides across all datasets and cell types; (ii) cell type coverage, towards identifying all distinct cellular states; (iii) data type diversity, for each cell type; (iv) activity pattern capture, for each functional element. We have previously applied these measures in ENCODE and modENCODE, to track the contribution of increasing numbers of experiments to the genomic coverage of human, Drosophila, and worm, for both conserved and non-conserved regions.

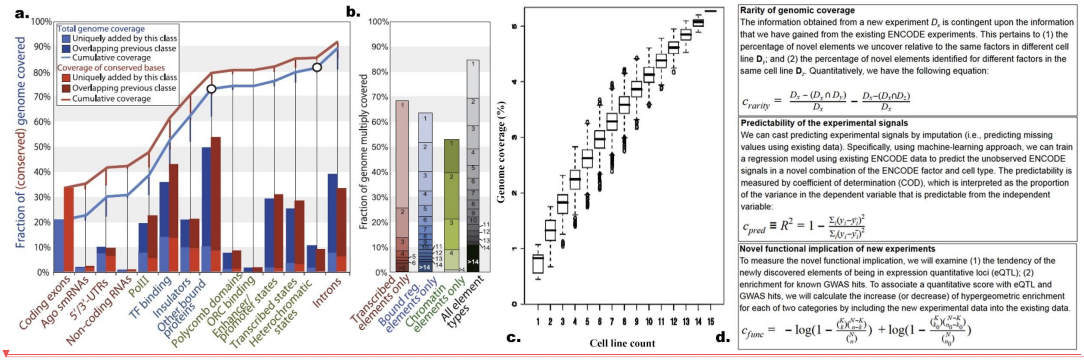
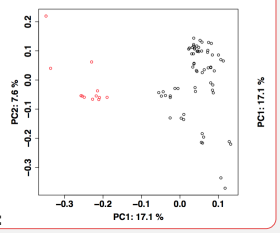


Figure 4.4. Genome coverage by ENCODE data types. **(a)** Unique (bars) and cumulative (lines) coverage of nonrepetitive (blue line) and conserved (red line) genome from *Drosophila* modENCODE. **(b)** Multiple coverage for datasets grouped into transcribed elements (red), bound regulators (blue), and chromatin domains (green) in *Drosophila* modENCODE. Across all three classes (black), 10.8% of the genome is covered 15 or more times, and 69.5% is covered at least twice. **(c)** Genomic coverage in human ENCODE with randomly reordered datasets as additional cell lines are added. Saturation plot for UW DNase1 over 15 cell lines, with all possible paths of random addition orders for cell lines. **(d)** Information-based measures of the contribution of individual assays towards completion of the ENCODE matrix covering DNA elements across diverse human cell types and diverse assays.

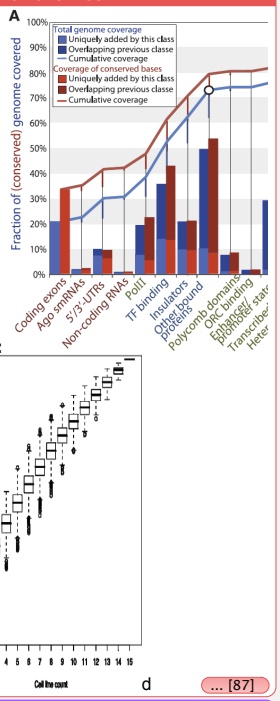
For each data type, we will develop and apply unbiased methods for evaluating the added per-nucleotide information content in a given cell type (**Fig. 4.4d**), based on: (i) genome-wide coverage, (ii) resolution, (iii) reproducibility. We will quantify the unique information each experiment provide in the context of the compendium using information-theoretic approaches that incorporate multiple factors: (1) the reproducibility of an assay between replicate experiments; (2) the resolution of the assay; (3) the robustness of the experiment

Author 3/20/2016 4:55 PM



Deleted:

Author 3/20/2016 4:55 PM



Deleted:

... [87]

Zhiping Weng 3/21/2016 2:25 AM

Comment [35]: +purcaro@gmail.com
Same thing for this figure

Michael Purcaro 3/21/2016 2:13 AM

Comment [36]: changed!

Michael Purcaro 3/21/2016 2:25 AM

Comment [37]: also bumped to 600 dpi

to variation in experimental conditions; (4) the rarity of the element type; (5) the ability to predict of a given assay from other assays in the same cell type; (6) the ability to predict a given assay from same/other assays in other cell types; (7) the increase in enrichments for independent datasets e.g. GWAS variants, regulatory motif matches, evolutionary conservation. resulting from the incorporation of a given experiment to an existing compendium; (8) the increased ability to predict known regulatory motifs by incorporation of the additional experiments; (9) the increase in the ability to predict the activity pattern of a given element resulting from incorporation of the additional experiment in an existing data compendium. Each of these metrics is influenced by: (a) the type of assay; (b) the specific cell type selected; (c) the experimental conditions used; (d) the quality of antibodies (when applicable); (e) the cell type heterogeneity of the sample; (f) the sequencing depth at which the experiment is carried out; (g) the amount of DNA extracted (and thus effective depth of the library).

4.4. Assay prioritization. To guide the ENCODE Consortium in terms of prioritizing particular datasets for usage, we will work closely with other members of the AWG to define metrics for evaluating the uniqueness or redundancy of different types of datasets. As of February, 2016, the ENCODE and Roadmap Epigenomics consortia have performed a total of 270 types of assays on at least one cell type, and at least one assay on a total of 320 cell types (encodeproject.org). Applying all these assay types to all these cell types would require 86,500 assays; however, the two consortia have performed only 2342 assays, 3% of the possible number.

Assay prioritization by imputation.

Assay prioritization by submodular optimization. We have developed a principled method to identify the subset of maximally-informative assays for completing the ENCODE experiment matrix, borrowing methods from the field of submodular optimization {Wei:2016hu}. Submodular functions {Fujishige:2005vx} have the property that the incremental gain of adding an item to the an existing set decreases as the set grows, and are widely used in economics{X. Vives, *Oligopoly Pricing: Old Ideas and New Tools*. MIT Press, 2001; M. Carter, *Foundations of Mathematical Economics*. MIT Press, 2001}, game theory{D. M. Topkis, *Supermodularity and complementarity* Princeton UP 1998; L. S. Shapley. "Cores of conex games" *Int. J. of Game Theory* 1(1):11-26, 1971}, combinatorial optimization{Lovasz:1983jc}{50–52}, electrical networks{53}, operations research{54}, and more recently, machine learning{55–58}, but they are not yet widely used for problems in biology. Selecting a panel of genomics assays that maximize a submodular function subject to a constraint on the size of the reported set is NP-hard, but can be approximately solved by a simple greedy algorithm with a worst-case approximation factor $(1-e^{-1})$){59}. Based on extensive empirical investigation, we selected the facility location function [54] as our submodular objective, which intuitively takes high values when every assay has at least one similar representative in *S,* and corresponds to the objective function of the k-medoids clustering problem{63}. This function has been previously applied in many fields, including document summarization[61], feature selection [57], and exemplar based clustering [62].

Zhiping Weng 3/21/2016 3:52 AM
Comment [38]: {Fujishige Submodula functions and optimization. Vol. 58. Elsevier Science, 2005}

Author 3/20/2016 4:55 PM
Deleted: 2015 http://dx.doi.org/10.1101/036137}.

Author 3/20/2016 4:55 PM
Deleted: *Submodula functions and optimization*. Vol. 58. Elsevier Science, 2005}

Author 3/20/2016 4:55 PM
Formatted: Highlight

Author 3/20/2016 4:55 PM
Deleted: {

William Stafford N..., 3/17/2016 10:50 PM
Comment [39]: [50] J. Edmonds. Matroids, submodular functions, and certain polyhedra. Combinatorial Structures and Their Applications, pages 69–87, 1970.

[51] L. Lovasz. Submodular functions and convexity. In M. Grotchel A. Bachem and B. Korte, editors, Mathemat-ical Programming – The State of the Art, pages 235–257. Springer-Verlag, 1983.
[52] A. Schrijver. Combinatorial Optimization. Springer, 2004.
[53] H. Narayanan. Submodular functions and electrical networks. Annals of Discrete Mathematics, 54, 1997.

[54] G. Cornunejols, G. L. Nemhauser, and L. A. Wolsey. The uncapacitated facility location problem. In P.B. Mirchandani and R.L. Franci, editors, Discrete Location Theory, chapter 3. Wiley/Interscience, New York, 1990.
[55] M. Narasimhan and J. Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. In Uncertainty in Artificial Intelligence (UAI), Edinburgh, Scotland, July 2005. Morgan ... [88]

William Stafford N..., 3/17/2016 10:49 PM
Comment [40]: Here are these cites: [57] Y. Liu, K. Wei, K. Kirchhoff, Y. Song, and J. Bilmes. Submodular feature selection for high-dimensional acoustic score spaces. In Acoustic ... [89]

Author 3/20/2016 4:55 PM
Deleted: a. ... [90]

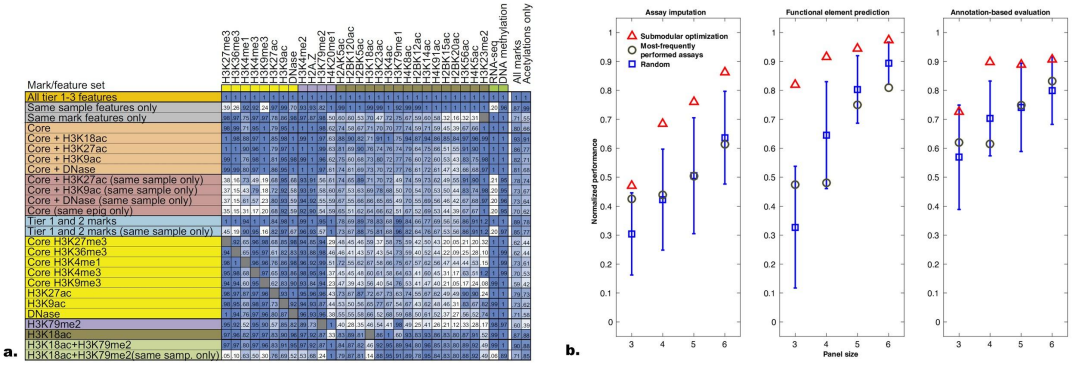


Figure 4.4: Assay prioritization. **a.** Imputation-based prioritization of assays based on their ability to impute other datasets, and the difficulty based on which they can be themselves imputed by other datasets, in the same cell type or across cell types. Each row denotes a feature set used as predictors. Each column denotes the dataset being imputed. Each entry denotes the imputation quality (blue=greater). Values denote fraction of signal correctly imputed compared to if that dataset was present. Core=H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3. Same-sample features are most important for acetylation marks, and same-mark features are most important for H3K27me3, H3K36me3, H3K9me3 and RNA-seq. Profiling of only H3K18ac and H3K79me2 allows higher relative imputation agreement than all five core marks, assuming a compendium with uniform coverage of marks. The last two columns show the average performance of the feature subset over all target marks and specifically for acetylations. **b.** Submodular optimization outperforms current sampling approaches in GM12878. Boxplots show the distribution of evaluation metrics over 40 random subsets (blue). The most-frequently performed assays (grey circle) is composed of the top k most frequent assays in our dataset.

Zhiping Weng 3/21/2016 2:30 AM
Comment [41]: +purcaro@gmail.com This figure too!
Michael Purcaro 3/21/2016 2:30 AM
Comment [42]: changed; also 600dpi

We have demonstrated empirically that this submodular optimization approach yields high quality panels of assays. To do so, we developed three evaluation schemes based on evaluating the panel's ability to impute missing data, identify functional elements, and accurately annotate the genome in a cross-validated setting (**Fig. 4.4b**). The quality of the selected panel is higher than randomly selected panels, as well as panels selected based on their frequency in the ENCODE+Roadmap data matrix (Figure 3). We have produced similar results for panels of transcription factor ChIP-seq assays (not shown). In addition, we note that the panel of histone modifications selected by our method closely overlaps with the panel selected by the Roadmap Epigenomics Consortium (our panels of size 6 or greater contain the Roadmap panel of size 5). Moving forward, we aim to extend our methods in several directions. First, we will explore application of the submodular approach to important variants of the problem. We have already investigated selection of assays in three scenarios: (1) "I want to select a set of x assays to perform in a new cell type." (2) "I have carried out x assays in a given cell type and want to select an additional set of y assays to perform." (3) "I have carried out x assays in a given cell type and want to select a representative subset of y assays from among them, for use in training a statistical model." We plan to use a similar approach to address the following scenarios: (1) "I have carried out x assays across a variety of cell types and assay types, and I would like to select a set of y additional assays in any combination of cell and assay types." (2) "Based on orthogonal data such as gene expression profiles, which new cell types should I perform assays in?". In addition, we plan to use our cross-validated evaluation strategy to investigate variants of our approach. For example, we will substitute alternative similarity measures in place of the Pearson correlation, to account for the non-uniform nature of the data across the genome or to allow for a more principled combination of similarities within the (additive) facility location function. We will also investigate weighting strategies to take into account, for example, the level of noise in individual experiments as measured by replicate experiments.

4.5. Cell type prioritization. For analyses whose goal is the definition of a particular type of element, we will use the predictive value of each data type as a way to prioritize datasets and assays. For unsupervised

learning analyses, we will use information theory metrics to evaluate the information content of each type of dataset per nucleotide, based on the reproducibility of the dataset, the resolution with which elements are defined for that data type, and the genomic coverage of that data type. We will also compare datasets to each other, to ask how predictable a given dataset is based on the combination of other datasets, and to evaluate and prioritize its added value in the context of the data generated by the entire Consortium. Intuitively, this is like determining if a particular analysis can potentially be reproduced without that dataset or if it is reproduced how inaccurate or how much it changes if that particular dataset is left out. This last type of calculation is particularly useful when one is thinking of repeating the same experiment, say a particular ChIP-seq experiment, over many different cell lines. That is, one wonders how much incremental information one gets from each additional cell line. We will also study the saturation of coverage for ChIP-seq and RNA-seq datasets to provide recommendations on the sequencing depth that should be achieved for different types of chromatin features and different types of transcripts. In addition to these goal-agnostic measures of dataset utility, we will develop and apply methods for prioritizing datasets based on specific tasks of interest. Specifically in the context of disease, we will provide a ranking of cell types and assays based on their predictive ability for genetic variants (from GWAS) or epigenetic marks (e.g. from MWAS) are associated with specific diseases and traits.

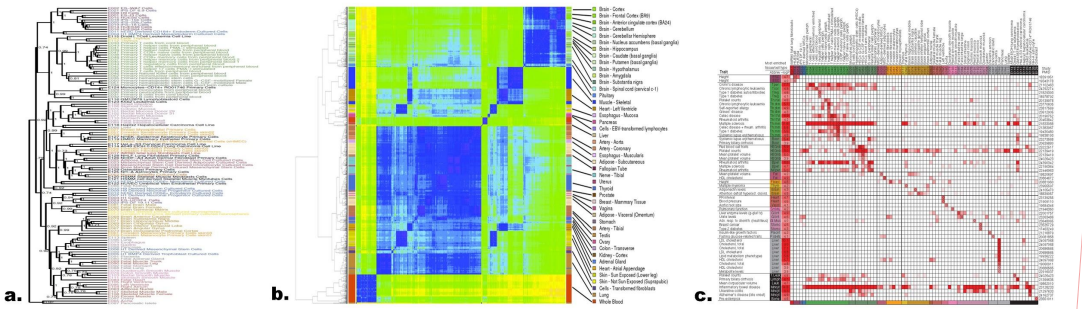
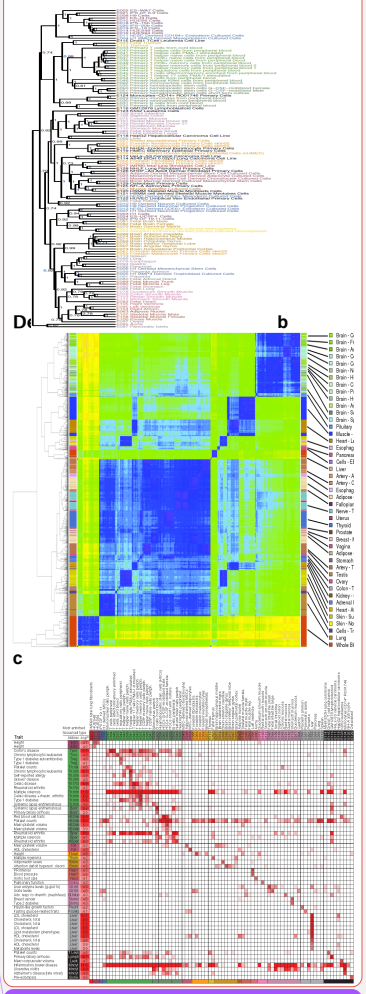


Figure 4.5. Cell type prioritization. (**a**) Cell type prioritization based on phylogenetic relationship of epigenomic information. H3K4me1 signal for existing cell types vs. all tissues and cell types profiled by related projects. **(b)** Cell type prioritization based on phylogenetic relationship of RNA-seq gene expression patterns. **(c) Prioritizing cell types based on disease enrichment.** Tissue-specific H3K4me1 peak enrichment significance (−log10 *P* value) for genetic variants associated with diverse traits. Circles denote reference epigenome (column) of most significant enrichment for SNPs reported by a given study (row), defined by trait and publication (PubMed identifier, PMID). Tissue (Abbrev) and *P* value (−log10) of most significant enrichment are shown. Only rows and columns containing a value meeting a FDR of 2% are shown.

Author 3/20/2016 4:55 PM



Zhiping Weng 3/21/2016 2:41 AM

Comment [43]: +purcaro@gmail.com
This one also!

Michael Purcaro 3/21/2016 2:41 AM

Comment [44]: changed; 600 dpi file

REFERENCES CITED

1. Feingold, E., Good, P. & Guyer, M. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* (2004).
2. ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
3. ENCODE Project Consortium An Integrated Encyclopedia of DNA Elements in the Human Genome. *ENCODE Companion submitted to Nature*
4. Wang, J. et al. Genome-wide mapping of the binding sites of 119 human transcription factors. *ENCODE Companion submitted to Nature*
5. Dong, X. et al. Correlating histone modifications and gene expression. *ENCODE Companion submitted to Genome Res*
6. Xi, H.S., Wang, J., Xu, J. & Weng, Z. Identification and Characterization of Allele-Specific Transcriptional Regulation in the Human Genome. *ENCODE Companion submitted to Genome Res*
7. Whitfield, T.W. et al. Functional analysis of transcription factor binding sites in human promoters. *ENCODE Companion submitted to Genome Biol*
8. Gerstein, M.B. et al. Architecture of the human regulatory network derived from ENCODE data. *ENCODE Companion submitted to Nature*
9. Gerstein, M.B. et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**, 1775–1787 (2010).
10. Gerstein, M.B. et al. What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17**, 669–681 (2007).
11. Trinklein, N.D. et al. Integrated analysis of experimental datasets reveals many novel promoters in 1% of the human genome. *Genome Res* **17**, 720–731 (2007).
12. Rozowsky, J.S. et al. The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. *Genome Res* **17**, 732–745 (2007).
13. Zhang, Z.D. et al. Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res* **17**, 787–797 (2007).
14. Zheng, D. et al. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res* **17**, 839–851 (2007).
15. Washietl, S. et al. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* **17**, 852–864 (2007).
16. Euskirchen, G.M. et al. Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res* **17**, 898–909 (2007).
17. Lu, Z.J. et al. Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res* **21**, 276–285 (2011).
18. Cheng, C. et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* **12**, R15 (2011).
19. Nègre, N. et al. A cis-regulatory map of the *Drosophila* genome. *Nature* **471**, 527–531 (2011).
20. Habegger, L. et al. RSEQtools: a modular framework to analyze RNA-seq data using compact, anonymized data summaries. *Bioinformatics* **27**, 281–283 (2011).
21. Sboner, A. et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol* **11**, R104 (2010).
22. Montgomery, S.B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
23. Li, Q., Brown, J.B., Huang, H. & Bickel, P.J. Measuring reproducibility of high-throughput experiments.

The Annals of Applied Statistics **5**, 1752–1779 (2011).
24. Gonzalez-Porta, M., Calvo, M., Sammeth, M. & Guigó, R. Estimation of alternative splicing variability in human populations. *Genome Res* (2011).doi:10.1101/gr.121947.111
25. Djebali, S. et al. Landscape of transcription in human cells. *ENCODE Companion submitted to Nature*
26. Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**, 996–1001 (2009).
27. Luco, R.F. et al. Regulation of alternative splicing by histone modifications. *Science* **327**, 996–1000 (2010).
28. Tilgner, H. et al. Chromatin mediated regulation of alternative splicing. *ENCODE Companion submitted to Genome Res*
29. Tilgner, H. et al. Deep sequencing of RNA from distinct subcellular fractions shows that splicing in the human genome occurs predominantly during transcription. *ENCODE Companion submitted to Genome Res*
30. Ørom, U.A. et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58 (2010).
31. Derrien, T. et al. The GENCODE v7 catalogue of human long non-coding RNAs: Analysis of their gene structure, evolution and expression. *ENCODE Companion submitted to Genome Res*
32. Kharchenko, P.V., Tolstorukov, M.Y. & Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**, 1351–1359 (2008).
33. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol* **9**, R137 (2008).
34. Rozowsky, J. et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**, 66–75 (2009).
35. Kheradpour, P. & Manolis, K. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *ENCODE Companion submitted to Genome Res*
36. Giardine, B. et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451–1455 (2005).
37. Liu, T. et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* **12**, R83 (2011).
38. Niranjan, T.S. et al. Effective detection of rare variants in pooled DNA samples using Cross-pool tailcurve analysis. *Genome Biol* **12**, R93 (2011).
39. Wu, H., Irizarry, R.A. & Bravo, H.C. Intensity normalization improves color calling in SOLiD sequencing. *Nat Methods* **7**, 336–337 (2010).
40. Bravo, H.C. & Irizarry, R.A. Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics* **66**, 665–674 (2010).
41. Hansen, K.D., Wu, Z., Irizarry, R.A. & Leek, J.T. Sequencing technology does not eliminate biological variability. *Nat Biotechnol* **29**, 572–573 (2011).
42. McCall, M.N., Murakami, P.N., Lukk, M., Huber, W. & Irizarry, R.A. Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics* **12**, 137 (2011).
43. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read datasets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**, e105 (2008).
44. Kao, W.-C., Stevens, K. & Song, Y.S. BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Res* **19**, 1884–1895 (2009).
45. Yang, X., Dorman, K.S. & Aluru, S. Reptile: representative tiling for short read error correction. *Bioinformatics* **26**, 2526–2533 (2010).
46. Zhao, X. et al. EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *J. Comput. Biol.* **17**, 1549–1560 (2010).
47. Schröder, J., Schröder, H., Puglisi, S.J., Sinha, R. & Schmidt, B. SHREC: a short-read error correction

method. *Bioinformatics* **25**, 2157–2163 (2009).
48. Kelley, D.R., Schatz, M.C. & Salzberg, S.L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol* **11**, R116 (2010).
49. Kuan, P.F. et al. A Statistical Framework for the Analysis of ChIP-seq Data. *Journal of the American Statistical Association* **106**, 891–903 (2011).
50. Lee, W. et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
51. Hansen, K.D., Brenner, S.E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**, e131 (2010).
52. Li, J., Jiang, H. & Wong, W.H. Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biol* **11**, R50 (2010).
53. Leek, J.T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733–739 (2010).
54. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724–1735 (2007).
55. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
56. ENCODE Project Consortium et al. A user's guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
57. Kolasinska-Zwierz, P. et al. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics* **41**, 376–381 (2009).
58. Guttman, M. et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
59. Stark, A. et al. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).
60. modENCODE Consortium et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
61. Cheng, C., Weng, Z. & Gerstein, M. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *ENCODE Companion submitted to Genome Res*
62. Cheng, C., Shou, C., Yip, K.Y. & Gerstein, M.B. Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors. *Genome Biol* **12**, R111 (2011).
63. Cheng, C. & Gerstein, M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* (2011).doi:10.1093/nar/gkr752
64. Barash, Y. et al. Deciphering the splicing code. *Nature* **465**, 53–59 (2010).
65. Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R. & Misteli, T. Epigenetics in alternative pre-mRNA splicing. *Cell* **144**, 16–26 (2011).
66. He, H.H. et al. Nucleosome dynamics define transcriptional enhancers. *Nature genetics* **42**, 343–347 (2010).
67. Kheradpour, P., Stark, A. & Roy, S. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res* (2007).
68. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
69. Yip, K.Y. et al. Genome-wide analysis of the binding sites of more than 100 transcription-related factors defines different types of genomic regions with distinct biological properties. *ENCODE Companion submitted to Genome Res*

70. Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. & Burge, C.B. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* **320**, 1643–1647 (2008).
71. Thurman, R.E., Day, N., Noble, W.S. & Stamatoyannopoulos, J.A. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res* **17**, 917–927 (2007).
72. Filion, G.J. et al. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**, 212–224 (2010).
73. Kharchenko, P.V. et al. Comprehensive analysis of the chromatin landscape in *Drosophila* melanogaster. *Nature* **471**, 480–485 (2011).
74. Jaschek, R. Spatial clustering of multivariate genomic and epigenomic information. *Research in Computational Molecular Biology* (2009).
75. Lian, H. et al. Automated mapping of large-scale chromatin structure in ENCODE. *Bioinformatics* **24**, 1911–1916 (2008).
76. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817–825 (2010).
77. Hoffman, M.M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Submitted*
78. Hoffman, M.M. et al. Integrative annotation of chromatin elements from ENCODE data. *ENCODE Companion submitted to Genome Res*
79. Day, N., Hemmaplardh, A., Thurman, R.E., Stamatoyannopoulos, J.A. & Noble, W.S. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**, 1424–1426 (2007).
80. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
81. Wang, J., Lunyak, V.V. & Jordan, I.K. Genome-wide prediction and analysis of human chromatin boundary elements. *Nucleic Acids Res* (2011).doi:10.1093/nar/gkr750
82. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 552 (Morgan Kaufmann: 1988).
83. Reynolds, S.M., Käll, L., Riffle, M.E., Bilmes, J.A. & Noble, W.S. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol* **4**, e1000213 (2008).
84. Entropic Graph Regularization in Non-Parametric Semi-Supervised Classification. 1803–1811 (2009).
85. Subramanya, A. & Bilmes, J. Semi-Supervised Learning with Measure Propagation. *J. Mach. Learn. Res.* **12**, 3311–3370 (2011).
86. Parallel Graph-Based Semi-Supervised Learning. (2011).
87. Jee, J. et al. ACT: Aggregation and Correlation Toolbox for Analyses of Genome Tracks. *Bioinformatics* (2011).doi:10.1093/bioinformatics/btr092
88. Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M.B. Annotating non-coding regions of the genome. *Nat Rev Genet* **11**, 559–571 (2010).
89. Frey, B.J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
90. Reiss, D.J., Baliga, N.S. & Bonneau, R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* **7**, 280 (2006).
91. Bonneau, R. Learning biological networks: from modules to dynamics. *Nat. Chem. Biol.* **4**, 658–664 (2008).
92. Bonneau, R. et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology datasets de novo. *Genome Biol* **7**, R36 (2006).
93. Cheng, C. et al. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol* **7**, e1002190 (2011).

94. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**, 37–46 (2002).
95. Lu, L.J., Xia, Y., Paccanaro, A., Yu, H. & Gerstein, M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res* **15**, 945–953 (2005).
96. Yu, H. & Gerstein, M. Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci USA* **103**, 14724–14731 (2006).
97. Shou, C. et al. Measuring the evolutionary rewiring of biological networks. *PLoS Comput Biol* **7**, e1001050 (2011).
98. Friedman, R.C., Farh, K.K.-H., Burge, C.B. & Bartel, D.P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92–105 (2009).
99. Yip, K.Y., Yu, H., Kim, P.M., Schultz, M. & Gerstein, M. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics* **22**, 2968–2970 (2006).
100. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3**, e59 (2007).
101. Bhardwaj, N., Kim, P.M. & Gerstein, M.B. Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci Signal* **3**, ra79 (2010).
102. Bhardwaj, N. et al. Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets. *PLoS Comput Biol* **6**, e1000755 (2010).
103. Bhardwaj, N., Yan, K.-K. & Gerstein, M.B. Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. *Proc Natl Acad Sci U S A* **107**, 6841–6846 (2010).
104. Yan, K.-K., Fang, G., Bhardwaj, N., Alexander, R.P. & Gerstein, M. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc Natl Acad Sci U S A* **107**, 9186–9191 (2010).
105. Malmström, L. et al. Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *PLoS Biol.* **5**, e76 (2007).
106. Mu, X.J., Lu, Z.J., Kong, Y., Lam, H.Y.K. & Gerstein, M.B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* **39**, 7058–7076 (2011).
107. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
108. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
109. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
110. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* (2011).doi:10.1093/nar/gkr917
111. Rossin, E.J. et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273 (2011).
112. Raychaudhuri, S. et al. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nature genetics* **40**, 1216–1223 (2008).
113. Raychaudhuri, S. et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5**, e1000534 (2009).
114. Segrè, A.V. et al. Common inherited variation in mitochondrial genes is not enriched for associations

with type 2 diabetes or related glycemic traits. *PLoS Genet* **6**, (2010).
115. Raychaudhuri, S. et al. Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nature genetics* **41**, 1313–1318 (2009).
116. Franke, A., McGovern, D., Barrett, J. & Wang, K. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics* (2010).
117. Gerstein, M. Annotation of the human genome. *Science* **288**, 1590 (2000).
118. Seringhaus, M.R. & Gerstein, M.B. Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinformatics* **8**, 17 (2007).
119. Gerstein, M., Seringhaus, M. & Fields, S. Structured digital abstract makes text mining easy. *Nature* **447**, 142 (2007).
120. Cheung, K.-H., Samwald, M., Auerbach, R.K. & Gerstein, M.B. Structured digital tables on the Semantic Web: toward a structured digital literature. *Mol. Syst. Biol.* **6**, 403 (2010).
121. Yale Law School Roundtable on Data and Code Sharing Reproducible Research: Addressing the Need for Data and Code Sharing in Computational Science. *Computing in Science and Engineering* **12**, 8–13 (2010).

Resource Sharing Plan, make sure to include text regarding software and analysis release document by the ENCODE

Multi-PI LEADERSHIP PLAN

The ENCODE DAC will be led by the contact PI, Prof. Zhiping Weng; 25% of her time will be spent on DAC/ENCODE issues. All major decisions, including hiring of personnel, budgeting, etc., will be made jointly by a three-member leadership committee consisting of Profs. Zhiping Weng, Mark Gerstein, and Manolis Kellis (they are multiple-PIs of the grant). This will be supported by a project manager (75% effort on DAC), to be hired at UMass. The remaining co-investigators (Guigo, Irizzary, Kundaje, Liu, and Noble) will report to the leadership committee. Finally, individual team members, including research scientists, postdocs, graduate and undergraduate students, will report to their respective co-investigators. The leadership committee as a whole will report to the AWG. All eight labs will participate in all four aims of the project, in response to the types and volumes of data as well as analysis needs. This management structure is depicted in **Figure 1**.

The project manager (supporting the whole leadership team) will be an individual with postdoctoral experience in bioinformatics coupled with good interpersonal skills, ideally with large-consortium experience. The project manager will spend 50% effort on managing the day-to-day activities of the DAC in the larger context of the AWG, communicating priorities, monitoring results and being accessible to all Consortium members and the NHGRI program officers at short notice. The other 25% of the project manager's effort will be spent on data analysis. The project manager will be hub of the communication network of the DAC.

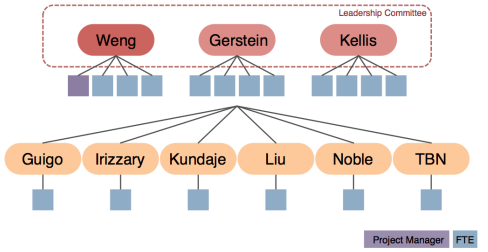


Figure 1. Management structure of the ENCODE DAC.

The Weng, Gerstein, and Kellis groups will each be staffed with the equivalent of four full time persons (FTEs in **Figure 1**, including the program manager for Dr. Weng, mostly postdocs with one or two graduate students), and the remaining five groups will each be staffed with one postdoc. Roughly 70% of the FTEs will perform directed analysis, handling both "routine" analysis and more development or pipelining tasks, rapidly changing priorities as set by the leadership committee and directed by the project manager. The remaining 30% of the personnel will develop or adapt methods of interest to the AWG. They will be assigned tasks in a longer time frame, with priorities jointly set by the leadership committee and co-investigators.

Currently the five participating groups (Guigo, Irizarry, Kundaje, Liu and Noble) are each staffed with one postdoc, which represents the baseline effort because we do not know yet which production groups are going

Author 3/20/2016 4:55 PM
Deleted: Leadership and reporting relationships.

Author 3/20/2016 4:55 PM
Deleted: around

Author 3/20/2016 4:55 PM
Deleted: .

Author 3/20/2016 4:55 PM
Deleted: ,

Author 3/20/2016 4:55 PM
Deleted: 11.

Author 3/20/2016 4:55 PM
Formatted: Space After: 0 pt

Author 3/20/2016 4:55 PM
Deleted: 50

Author 3/20/2016 4:55 PM
Deleted: 11

Author 3/20/2016 4:55 PM
Deleted: DAC

Author 3/20/2016 4:55 PM
Deleted: <*sp*>

to be funded in the next round of ENCODE. We have also reserved the funding for one more FTE and 5% of an investigator's effort (indicated as TBN in the figure). When we obtain the information of the funded production and characterization groups in ENCODE4, we will be able to assess the number of data types and the amount of data in each type, and provide the reserved funding to the groups with the matching expertise, or recruit a new group.

The physical proximity of Weng, Gerstein, and Kellis (~2 hours door-to-door) will greatly facilitate the interactions among the DAC members in these three groups. The groups will have monthly physical meetings, rotating among the three universities. The three groups will be in constant communication. This close interaction will be particularly beneficial for completing analyses and formulating integration manuscripts.

All members of the DAC are expected to participate in open discussion forums. These include:

1. Use of a single AWG mailing list for email discussion.
2. A semimonthly priority-setting phone conference with the AWG. We will adopt the current chairing structure of the AWG calls where the leaders are appointed by the NHGRI. Weng currently co-chairs the AWG with production PIs Mike Snyder and Ross Hardison.
3. At least two physical meetings per year of the entire group, often coordinated with Consortium-wide programs, but with a day before or after the Consortium meeting dedicated to analysis tasks. In some cases, these analysis meetings will be run in a mini-jamboree mode where intensive collaborative work occurs.
4. All postdocs, students, and the project manager will have Skype-based voice and instant messaging active during their working hours, allowing for quick and spontaneous interactions.

Coordination of DAC efforts

We will begin the ENCODE4 DAC activities with the project launch meeting that NHGRI will organize. All of the funded members of the DAC will attend this meeting and meet one another as well as meet other members of the Consortium. We will establish shared expectations among the co-investigators and the leadership team, and clarify individual roles within the larger AWG effort.

Thereafter, as in the current phase of ENCODE, the DAC members will interact with one another via the bimonthly main AWG calls. Attendance at this conference call will be mandatory for all funded members of the DAC. Scientific presentations on topics that are of broad interest to the entire Consortium will be made by each funded DAC lab on a rotating basis. The AWG calls are open to all Consortium members and other groups also make presentations. The project manager of the DAC will organize and take minutes for these AWG calls.

As mentioned in Section Section 2.3, there are four AWG subgroup calls that started in the third year of ENCODE3, two semimonthly and two monthly.

the project manager of the DAC will organize semi minuted conference call. Scientific presentations will be made by each funded lab on a rotating basis.

Project documentation will be managed via the ENCODE Wiki, which will include a copy of the chart shown in Figure 11, as well as a list of all funded personnel. The Wiki will include one page dedicated to each participating group. These team-specific pages will serve as the primary reporting mechanism for each subcontract. Each page will be organized chronologically, and will include links to presentations made on the biweekly conference calls, results of interim analyses, in-progress or submitted manuscripts, etc.

Author 3/20/2016 4:55 PM
Formatted: Space After: 0 pt

Author 3/20/2016 4:55 PM
Deleted: monthly

Author 3/20/2016 4:55 PM
Formatted: Space After: 0 pt

Author 3/20/2016 4:55 PM
Deleted: We will adopt the current chairing structure of the AWG calls (where the leaders are appointed by the N... [91]

Author 3/20/2016 4:55 PM
Deleted: involves all

Author 3/20/2016 4:55 PM
Formatted: Space After: 0 pt

Author 3/20/2016 4:55 PM
Deleted: ENCODE

Author 3/20/2016 4:55 PM
Deleted: a two-day

Author 3/20/2016 4:55 PM
Deleted: , with representation from the funded DCC. This

Author 3/20/2016 4:55 PM
Deleted: will provide scientific background for new

Author 3/20/2016 4:55 PM
Deleted: ,

Author 3/20/2016 4:55 PM
Moved (insertion) [1]

Author 3/20/2016 4:55 PM
Deleted: a biweekly,

Author 3/20/2016 4:55 PM
Formatted: Space After: 0 pt

Author 3/20/2016 4:55 PM
Moved up [1]: Attendance at this conference call will be mandatory for all funded members of the DAC.

Conflict Resolution

Most issues will be resolved by calls and meetings between the three members of the leadership committee. We will have a standing weekly 30-minute call where all issues arising will be brought up, discussed and resolved. Each member of the DAC will be able to contact any of the three members of the leadership committee, with the guarantee that any concerns will be brought up anonymously. Similarly, all members of the AWG will be able to address any concern they have anonymously to any member of the leadership committee. Similarly, the NHGRI will be able to raise its concerns with the leadership committee. Contentious issues about which complete agreement is not immediately reached among the three members of the leadership committee will be decided by majority vote.

Budget reassignments

Budget adjustments will be made yearly based on productivity of each member of the DAC and shifting priorities of the Consortium. This will only happen will full agreement of all three members of the leadership committee, and after notification of NHGRI staff.

Team expertise of the DAC

We have organized the DAC such that Weng, Gerstein, and Kellis are multiple-PIs and Guigo, Irizarry, Kundaje, Liu, and Noble are co-investigators. All eight of us have collaborated effectively, in particular as members of the current ENCODE DAC, and have highly complementary expertise. Each of us will participate in all four aims of the project, in response to the types and volumes of data as well as analysis needs. Nevertheless, due to the complexity of the Consortium, we have devised a management plan to maximize the effectiveness and responsiveness of the DAC.

Weng, Gerstein, and Kellis will jointly make decisions on all matters related to the DAC. They will jointly coordinate the integration of diverse data sources and oversee the running of the pipelines (Aim 1 & 3.2). Weng currently leads the development of the Encyclopedia and will continues to lead this effort (Aim 2). Gerstein has made important contributions to all aspects of the current DAC and he will oversee the DAC activities in support of the AWG (Aim 3, excluding 3.2). Kellis has perform extensive analysis on the quality and utility of ENCODE data and will continue to lead this effort (Aim 4).

Guigo has extensive expertise on analyzing RNA transcription data and will participate in all the DAC activities that involve RNA. Irizarry is a bio-statistician highly recognized for his work on identifying and minimizing batch effects. He will participate in the analysis whenever we compare datasets of the same type but generated by different labs. Kundaje has been instrumental for establishing the existing uniform analysis pipelines and quality assessment metrics for TF and histone mark ChIP-seq data, and will continue to update these and related analysis pipelines and data standards. Noble has built a number of algorithms for analyzing chromatin conformation data, and will play a key role in generating the third level catalogs of the Encyclopedia. Liu has built many methods and databases for integrating DNase-seq and ChIP-seq data (both TF and histone marks), and will work on expanding the coverage of the Encyclopedia by incorporating these and other relevant data from the public domain. For example, there are many H3K27ac ChIP-seq datasets in GEO and integrating these with ENCODE data can significantly increase the cell types for which we can make reliable enhancer predictions.

Risk assessment and Leadership

Author 3/20/2016 4:55 PM
Deleted: .

Author 3/20/2016 4:55 PM
Formatted: Underline

Author 3/20/2016 4:55 PM
Formatted: Space After: 0 pt

Author 3/20/2016 4:55 PM
Formatted: Underline

Author 3/20/2016 4:55 PM
Deleted: .

Author 3/20/2016 4:55 PM
Deleted: Allocation of reserved funds for existing and additional investigators. Second, the five participating groups (Guigo, Irizarry, Kundaje, Liu and Noble) are each staffed with one postdoc, which represents the baseline effort because we do not know yet which production groups are going to be funded in the next round of ENCODE. When we obtain the information of the funded production groups, we will be able to assess the number of data types and amount of data in each type, and provide additional FTEs to the groups with the matching expertise.

The proposed management structure is designed to achieve multiple aims simultaneously. Accordingly, the leadership committee brings together three co-leaders with complementary strengths and very broad expertise. Furthermore, this joint leadership plan will likely provide greater responsiveness to the many partners involved than would a single, overworked PI trying to lead the whole DAC. Indeed, one of the primary goals of the reporting mechanisms and conflict resolution procedures outlined above is to maintain clarity, efficiency, and agility about the changing priorities of the various Consortium members, including the ENCODE data production labs, the AWG, the DCC, the DAC and the NHGRI. In particular, it will be critical to balance the scientific expertise and interests of the various DAC co-investigators with the needs of the Consortium to ensure that we are achieving our aims.

There are a number of potential risks that a project of this size and complexity could encounter. We have discussed these in the group and feel that all the risks can be mitigated or handled appropriately if they occur.

1. Inefficient collaboration due to geographical separation. This problem is unlikely to occur, mainly because five of the eight groups are located within 100 miles and all eight groups (including the three distal groups Kundaje, Noble and Guigo) already work together well as part of the current ENCODE DAC. In addition, monthly phone calls (rising in tempo during publication drives) coupled with pervasive VoIP/instant messaging will provide a strong sense of virtual community. If one group is not responsive or integrating with others, we will discuss the issue at the PI level first, followed by personnel visits to encourage networking and integration. If the problem persists, the leadership committee will discuss it with the PIs involved together with the NHGRI program director. In the extremely unlikely scenario that there is a truly intransigent group, which is hard to imagine given the DAC membership, we would consider withdrawing funds from this group upon the NHGRI consent.
2. DAC groupthink excluding input from other groups. This problem is unlikely to occur because of our track record in working openly with all groups and the genuine collaborative nature we take to solving problems. The presence of the AWG to provide independent input and prioritization is a formal mechanism to ensure the DAC groups do not become a closed club. We are also happy to work with the NHGRI program directors to implement other approaches if desired. Equally important will be informal aspects of our openness to collaboration. All DAC discussions will be open to all members of the Consortium.
3. Too many tasks from the AWG or switching of priorities too rapidly. We believe there is an appreciable risk that there will be a far larger task list than the DAC can accommodate, requiring tough prioritization decisions. The transparency of our process is critical here, as the AWG will need to set sensible priorities with some month-to-month consistency. It will be critical to build a high level of trust between the AWG and the DAC early in the project, before complex prioritization becomes an issue. Again, given all eight of the groups are existing members of the ENCODE Consortium, there is already considerable trust between these groups and the AWG.
4. Too few tasks from the AWG. This is a potential issue at the start of the project, but there is a healthy list of pipelining tasks that will provide infrastructure later. ENCODE datasets are accumulating rapidly, so there are a lot of data to provide input into sensible biological questions early on.

Mark Gerstein 3/15/2016 12:26 AM
Comment [45]: +manoli@mit.edu ToDo4MG: describe as already ongoing. Might need to be updated by Manolis

Mark Gerstein 3/15/2016 12:26 AM
Comment [46]: please put something in ASAP

Mark Gerstein 3/15/2016 12:26 AM
Comment [47]: +zhipingweng@gmail.com pls update too

Manolis Kellis 3/15/2016 12:27 AM
Comment [48]: Manolis to continue

Mark Gerstein 3/15/2016 12:26 AM
Comment [49]: +manoli@mit.edu ToDo4MG: describe as already ongoing. Might need to be updated by Manolis

Mark Gerstein 3/15/2016 12:26 AM
Comment [50]: please put something in ASAP

Mark Gerstein 3/15/2016 12:26 AM
Comment [51]: +zhipingweng@gmail.com pls update too

Zhiping Weng 3/11/2016 5:47 AM
Comment [52]: +wnoble@uw.edu please insert reference in the {FirstAuthor, Year, #PMID} format.

Mark Gerstein 3/16/2016 2:59 AM
Comment [53]: +wnoble@uw.edu +zhipingweng@gmail.com Suggest shortening these & removing latex look

Mark Gerstein 3/16/2016 2:59 AM
Comment [54]: cut by 50%

Mark Gerstein 3/16/2016 2:51 AM
Comment [55]: +zhipingweng@gmail.com think we should cut this entire gray block

Author 3/20/2016 4:55 PM
Deleted: Old Section 3: ... [92]