Abstracts presented at the

ENCODE 2015: Research Applications and Users Meeting

Bolger Center
Potomac, Maryland

June 29 – July 1, 2015

Table of Contents

Dynamic Enhancer Landscapes during Pancreatic differentiation of human ES cells
Feng Yue[1], Allen Wang[2], Yan Li[3], Maike Sander[2], Bing Ren[3] 33

Re-annotating histone modification data at a single nucleosome resolution

Yongbing Zhao, Yan Asmann, and Zhaoyu Li

Department of Cancer Biology, Mayo Clinic

We have developed a novel software, NUCLIZE, to annotate histone modification data in the ENCODE database in a single nucleosome resolution with our own sequenced nucleosome mapping data in human normal and cancer cell lines. The ENCODE project has generated substantial genome-wide histone modification maps in a large number of human normal and cancer cells using the chromatin immunoprecipitation and sequencing (ChIP-Seq) approach. The comparison of the reference epigenomic profiles with those in human cancer cells will greatly advance our understanding of the mechanisms of cancer initiation and progression and will guide therapeutic studies on human cancers. However, there are substantial analytic challenges for such comparisons: (1) the "peaks" called from the ChIP-Seq data are usually broad regions, ranging from several hundred to several million base pairs for a single histone marker; (2) the peaks from ChIP-Seq are limited in resolution due to the sonication process, so the precise locations and sizes of the peaks for the same marker vary from experiment to experiment; (3) the cross-linking protocol by formaldehyde introduces noise or non-specific signals of histone markers. Thus, we developed the NUCLIZE to re-annotate histone modification at a single nucleosome resolution. The NUCLIZE software requires high-resolution nucleosome mapping data from native MNase-Seq without crosslinking as we have established before. We have generated a series of MNase-Seq data in human normal and cancer cell lines. With these nucleosome mapping data, we have re-annotated histone modification data in the ENCODE database and found many novel information about histone modification in individual cell lines and between normal and cancer cell lines, which is critical for accurate understanding of epigenomic regulation during neoplastic transformation of human cancers.

Discovering Gene Regulatory Elements Using Coverage-based Heuristics

Rami Al-Ouran[1], Robert Schmidt[1], Ashwini Naik[2], Jeffrey Jones[3], Frank Drews[1], David Juedes[1], Laura Elnitski[4], and Lonnie Welch[1]

[1] School of Electrical Engineering and Computer Science, Ohio University, Athens, OH 45701
[2] Nationwide Children's Hospital, Columbus, OH 43110
[3] Computer Science and Engineering, The Ohio State University, Columbus, OH 43210
[4] The National Human Genome Research Institute, Bethesda, MD 20892

Data mining algorithms and sequencing methods (such as RNA-seq and ChIP-seq) are being combined to discover genomic regulatory motifs for a variety of phenotypes. However, motif discovery algorithms often produce unwieldy lists of putative transcription factor binding sites, hindering the discovery of important phenotype-related regulatory elements by making it difficult to select important motifs for experimental validation. The authors introduce coverage-based search heuristics to solve the motif selection problem. Analysis of 203 ChIP-seq experiments from the ENCyclopedia of DNA Elements project (ENCODE) shows that the methods perform well in terms of both precision and sensitivity. Additionally, the authors report new biological insights from their analysis of the ENCODE data.

Statistical discovery of synergetic transcription factor regulations with ENCODE data

M. Moria[1], A. Terada[2], J. Sese[3]
[1]AIST, [2]The university of Tokyo, [3]AIST

Genes are often regulated by multiple transcription factors (TFs). The discovery of combinatorial regulations was, however, computationally and statistically challenging. Recently, the introduction of a limitless arity multiple-testing procedure (LAMP) has made this problem tractable. We here present an integrated pipeline that takes in as input ChIP-seq and RNA-seq data and utilizes LAMP to detect statistically significant combinations of TFs regulating gene expressions. We here demonstrate on the application of the software on ES-cell expression analysis with ENCODE data. This software will contribute to find hidden associations between TFs by combining ENCODE data with ChIP-seq and RNA-seq data observed in your laboratory.

Spectacle: fast chromatin state annotation using spectral learning

Jimin Song, Kevin C Chen

Department of Genetics and BioMaPS Institute for Quantitative Biology, Rutgers University

A common use of ENCODE data is to associate combinations of epigenetic marks with regulatory elements, for example, using programs such as ChromHMM and Segway. We have developed a new program, Spectacle, that uses spectral learning instead of the commonly-used expectation-maximization algorithm and showed that Spectacle is much faster and can produce more biologically relevant results for some cell lines, as shown by enrichment of GWAS SNPs. The speed of Spectacle also allows us to jointly analyze epigenomic data from multiple cell lines while incorporating cell lineage or evolutionary relationships and in ongoing work we find that this integrated approach can improve our predictions of some genomic features, such as promoters. Our annotations of the ENCODE Tier 1 and 2 cell lines and our software are available at https://genfaculty.rutgers.edu/kcchen/home.

Genome-wide characterization of chromatin state plasticity

Luca Pinello[1,2], Alexander Gusev[2], Hilary Finucane[2], Alkes Price[2] and Guo-Cheng Yuan[1,2]

[1] Dana-Farber Cancer Institute, Boston MA, US
[2] Harvard T.H. Chan School of Public Health, Boston MA, US

With the increasing amount of epigenomic data, a pressing challenge is to understand how the chromatin states are regulated; in particular the mechanisms for their cell-type specific establishment and maintenance.

We have recently developed a computational method called HAYSTACK to systematically quantify the variability of a single histone modification mark and identify transcription factors that are likely to play an important role in mediating its cell-type specific patterns; and their association with gene expression variation.

Here we extend this method by investigating the variability of combinatorial states of multiple histone marks using 9 human cell lines from the ENCODE project. We find that the highly plastic regions are highly enriched in important regulatory regions, such as super-enhancers and Polycomb repressed regions. We also identify large-scale regions of co-variability, and show that such regions may span multiple co-regulated genes, such as those associated with the well-studied b-globin locus control regions in K562 cells. In addition, the co-variability measure shows a clear depletion in correspondence of the boundaries of topological associated domains (TAD), obtained by HI-C assays; suggesting that this measure could be helpful to highlight and refine functional modules within the TAD.

We are also able to recover specific chromatin state transitions within each functional group, in fact different contexts such as enhancer or Polycomb rich states can transition to completely different set of chromatin states; suggesting some constraint on the chromatin state usage during development. Moreover we find that the highly plastic regions are highly enriched for GWAS-associated non-coding variants and, in some cases, are more powerful for explaining the heritability compared to existing annotations of regulatory regions. We are currently using the chromatin state variation as a guide to search for functionally important genetic variants.

Our analysis provides new insights into the organization and dynamic change of cell-type specific chromatin structure during development and a valuable tool for investigating the mechanisms of chromatin state establishment and usage.

Modeling Reproducibility of High Throughput Sequencing Data with Tail Dependences when Pearson and Spearman Correlations Fail

Tao Yang[1], Qunhua Li[1,2]

[1] Bioinformatics and Genomics program, The Pennsylvania State University
[2] Department of Statistics, The Pennsylvania State University

The quality and reproducibility of sequencing-based experiments is essential to the reliability of downstream analysis and biological interpretation. Though Pearson and Spearman correlation coefficients are often used to assess the reproducibility of replicate sequencing experiments, they can be easily misled by highly repetitive regions or excessive amount of low count regions on the genome. Here we developed a novel reproducibility measure based on tail dependence that can overcome the drawbacks of correlation coefficients. We evaluate our methods on different sequencing experiments. Our measure is robust and can effectively distinguish experiments with different levels of reproducibility and quality. It is applicable to various sequencing based genetic and epigenetic data in which substantial amount of noise is often present. This measure will help practitioners identify suboptimal experiments and the causes of suboptimality.

A Supervised Learning Approach to the Prediction of Hi-C Data

Tyler Derr, Yanli Wang, Feng Yue

The Department of Biochemistry & Molecular Biology, Pennsylvania State University College of Medicine, Hershey, PA 17033

The state of the art method for studying genome-wide chromatin structures is Hi-C, which is a high-throughput chromosome conformation capture (3C) based technology. However, due to the cost of performing Hi-C experiments, only a small subset of all possible species, tissue, and cell type data are currently available. We propose a supervised learning method for the prediction of the entire intra-chromosomal Hi-C interaction data using a Random Forest (RF) approach. The learning algorithm is trained using a known Hi-C matrix and a set of organized features for each region in the chromosome, which includes the GC content, mappability, number of restriction enzyme cut sites, histone modifications (H3K4me3, H3K36me3, etc.), and transcription factor binding sites (TFBSs) (Pol2, Ctcf, etc.), which are freely available thanks to the recent efforts of the ENCODE and Roadmap Epigenomics projects. We have used roughly 10 histone modifications and 10 TFBSs, but these numbers vary based on the intersection of available data between the training chromosome and the one for which we are attempting to predict. The results we have gathered thus far on Hi-C predictions in human and mouse have not only proven to be highly correlated to the real Hi-C, but they also preserved many of the Topologically Associating Domains (TADs). An analysis on the feature importances returned by the RF has allowed us to determine which of the features are more meaningful for the interactions between regions at varying distances. Finally, a comparison is given between the inferred 3D structures of the real Hi-C and its corresponding predicted Hi-C using the methods available in the Pastis package.

SNPGenie: a software platform for detecting natural selection in pooled next-generation sequencing samples

Chase W. Nelson , Austin L. Hughes
Department of Biological Sciences, University of South Carolina, 715 Sumter St., Columbia, SC 29208

A new approach to population genetics analysis involves using next-generation sequencing (NGS) technologies to characterize genetic variants in pooled (i.e., multiple individuals) samples. SNPGenie is a software platform, written in PERL, that aids in the detection of natural selection using such pooled NGS data. By calculating and comparing nucleotide and gene diversities at nonsynonymous, synonymous, and ambiguous sites, SNPGenie allows competing hypotheses of molecular evolution to be tested. Importantly, these statistics do not rely on linkage information or extremely low-frequency variants, bypassing the main limitations of pooled NGS approaches. Additionally, data from any SNP caller may be analyzed using SNPGenie, making it the most versatile tool for such analyses. Use of the software is demonstrated with H5N1 serial infection data.

Integrative functional annotation in the human genome and its application in prioritizing results in genome-wide association studies

Qiongshi Lu[1], Xinwei Yao[2], Yiming Hu[1], Hongyu Zhao[1,3]
[1] Department of Biostatistics, Yale School of Public Health, New Haven, CT;
[2] Yale College, New Haven, CT;
[3] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT

Identifying functional regions in the human genome is a major goal in human genetics. Great efforts have been made to functionally annotate the human genome either through computational predictions, such as genomic conservation, or high-throughput experiments, such as the ENCODE project. These efforts have resulted in a rich collection of functional annotation data of diverse types that need to be jointly analyzed for integrated interpretation and annotation. We have developed GenoCanyon (http://genocanyon.med.yale.edu), a whole-genome annotation method that performs unsupervised statistical learning using 22 computational and experimental annotations thereby inferring the functional potential of each nucleotide in the human genome. The ability of predicting functional regions as well as its generalizable statistical framework makes GenoCanyon a unique and powerful tool for whole-genome annotation. Based on GenoCanyon, we further developed GenoWAP (Genome Wide Association Prioritizer, available at http://genocanyon.med.yale.edu/GenoWAP), a post-GWAS prioritization method that integrates genomic functional annotation and GWAS test statistics. The effectiveness of GenoWAP is demonstrated through its applications to GWAS results for Crohn's disease and schizophrenia using the largest studies available. After prioritization based on a subset of all the available samples, highly ranked loci show substantially stronger signals in the whole dataset than the top loci before prioritization. At the single nucleotide polymorphism (SNP) level, top ranked SNPs after prioritization have both higher replication rates and consistently stronger enrichment of eQTLs. Within each risk locus, GenoWAP is able to distinguish real signal sources from groups of correlated SNPs. In summary, our functional annotations can guide genetic studies at multiple resolutions and provide valuable insights in post-GWAS prioritization.

Advances in ChIP-based Technologies for Profiling Epigenomic Landscapes and Gene Regulatory Networks

Adam Blattler, Bryce Alves, Mary Anne Jelinek, Johanna Samuelsson, Brian Egan, Terry Kelly

Active Motif, Inc. Carlsbad CA

Efforts by national consortia such as ENCODE have resulted in a comprehensive characterization of epigenetic marks and transcription factor binding events across several cell lines, providing the scientific community with standards for validating and cross-checking new datasets. It has also provided a reference point for the development of technologies that aim to improve on the shortcomings of existing epigenomic assays. Here we present several assays that improve on the limitations of ChIP or allow for an expansion of our abilities to understand underlying mechanisms of the epigenetic regulation of our genomes. 1. TAM-ChIP takes advantage of transposase-conjugated antibodies to integrate barcoded sequencing adapters in a site-directed manner, enabling the potential to multiplex ChIP reactions for several factors in the same sample. 2. Tag-ChIP uses a novel epitope tag ("AM-Tag") specifically designed for ChIP that can be used to study proteins for which ChIP grade antibodies are not possible or available. 3. Spike-In takes advantage of a Drosophila-specific histone variant to control for technical variation that arises between ChIP-seq experiments. 4. enChIP allows for the pull-down of a single genomic locus in order to identify cis- and trans-chromosomal looping events. Along with these developments, Active Motif is constantly producing and validating new ChIP-seq-grade antibodies to complement the 78 ChIP-grade antibodies we currently offer, and continues to collaborate with several ENCODE groups to expand our ChIP validated antibody offerings.

MUC5B Promoter Methylation as a Transcriptional Regulator and Risk Factor for Idiopathic Pulmonary Fibrosis

Britney A. Helling[1], Brent Pedersen[1], Mark P. Steele[4], Kevin K. Brown[1,3], James E. Loyd[4], Gregory Cosgrove[3], Marvin I. Schwarz[1,3], Steve D. Groshong[1,3], Elissa Murphy[1], Tasha E. Fingerlin[2], Ivana V. Yang[1,2,3], David A. Schwartz[1,3]

[1] The University of Colorado–Denver, School of Medicine, [2]University of Colorado–Denver, School of Public Health, [3]National Jewish Health, [4]Vanderbilt University School of Medicine, Nashville

Rationale: A common MUC5B promoter polymorphism (rs35705950) has been associated with increased risk of idiopathic pulmonary fibrosis (IPF) and increased expression of MUC5B both in normal and diseased lung tissue. However, this polymorphism does not account for all the variability in MUC5B expression. Given that IPF is associated with global changes in DNA methylation and the previous association between MUC5B expression and global changes in DNA methylation, I hypothesized that methylation of the MUC5B promoter contributes to changes in MUC5B expression through altering transcription factor binding and is associated with disease risk.

Methods: Human lung tissue derived RNA and DNA were obtained from the NHLBI-Lung Tissue Research Consortium (NHLBI-LTRC) and National Jewish Health-Interstitial Lung Disease Program, including IPF (N=250) and unaffected control (N=140) samples. Lung MUC5B expression levels were determined using a pre-validated MUC5B Taqman assay normalized to a pre-validated GAPDH assay. DNA methylation in the 4 kb promoter region of MUC5B was measured using a custom designed EpiTyper methylation panel by Sequenom. The comb-p statistical method was used to determine differentially methylated regions (DMRs).
Results: In the 4 kb promoter of MUC5B I identified 2 DMRs associated with disease status($p=8.77 \times 10^{-59}$, $p=2.54 \times 10^{-7}$), one associated with MUC5B expression ($p= 4.19 \times 10^{-6}$) and one DMR associated with the common MUC5B promoter polymorphism ($p=1.08 \times 10^{-39}$). In each of these associations, the risk phenotype–IPF, increased MUC5B expression, and the risk allele–were all associated with hypermethylation of the MUC5B promoter in the associated region.

Conclusions: Our findings suggest that the common MUC5B promoter polymorphism may result in differential methylation of the MUC5B promoter which appears to be a regulator of MUC5B expression and a risk factor for IPF. This conclusion is further supported by the ENCODE data which indicates this area is likely regulatory. The DMRs are in an area of open chromatin with 18 transcription factor binding sites indicated by ChIP-Seq.

Distinct DNA methylation profiles distinguish the disease progression stages of low-risk MDS patients at the time of diagnosis

Tingting Qin[1], Jason Sotzen[1], Stephen Nimer[2], Virginia Klimek[2], Maria E. Figueroa[1]

[1] University of michigan
[2] Memorial Sloan Kettering Cancer Center

Myelodysplastic syndromes (MDS) are a heterogeneous group of clonal cancers characterized by the presence of peripheral cytopenias, bone marrow hypercellularity and dysplastic changes in the bone marrow and an increased risk of progression to acute leukemia. However, not all patients evolve to an aggressive leukemic phase and clinically it has been observed that some of the low-risk MDS patients (low or Int-1 by International Prognostic Scoring System, IPSS) will remain in a relatively stable cytopenic phase (stable MDS) while other low-risk patients will progress to an aggressive leukemic phase (progressive MDS). Limited understanding of the molecular mechanisms underlying this progression makes it currently impossible to distinguish at diagnosis patients who will later progress and hinders the development of therapies capable of preventing this transformation. Previous studies based on DNA promoter methylation array have revealed the presence of aberrant DNA methylation in MDS and this abnormality is more pronounced in the patients with more aggressive stages. However, it is unclear whether these more extensive abnormalities in progressive MDS are present from the early stages of the disease or simply a consequence of the disease progression. Therefore, it is critical that we elucidate the DNA methylation profiles that distinguish stable MDS patients from progressive ones at time of diagnosis and follow-up separately in order to (i) improve patient risk-stratification at diagnosis and (ii) better understand molecular mechanisms behind the different phenotype of MDS. For this purpose, we studied the DNA methylation status at ~3M CpG sites across the genome of 20 baseline and follow-up (~18 months after diagnosis) paired MDS patients (13 stable and 7 progressive MDS) using Enhanced Reduced Representation Bisulfite Sequencing (ERRBS). We identified 292 statistically significant differentially methylated regions (DMRs) (FDR<0.1 and methylation difference ≥25%) between the progressive and stable MDS at diagnosis (baseline) and not surprisingly the number increased up to and 593 at follow-up, out of which 67 DMRs were overlapped with the baseline DMRs. The DMRs were depleted at promoters (baseline DMRs 11% vs. Background [BG] 25%, p-value: 1.78×10-10; follow-up DMRs 15% vs. BG 24%, p-value: 7.58×10-9) and CpG islands (baseline DMRs 12% vs. BG 29%, p-value: 8.98×10-12; follow-up DMRs 17% vs. BG 28%, p-value: 1.05×10-9). Moreover, by correlating our DMRs with ENCODE defined enhancers we demonstrated that hypomethylated DMRs at follow-up were significantly enriched for hematopoietic enhancers, and in particular, enhancers located within gene bodies (hypo DMRs 34% vs. BG 19%, p-value: 6.61×10-14). The motif analysis showed an enrichment of the follow-up DMRs in the p53 genomic binging sites (Benjamini p-value =

0.08). Our findings demonstrated that specific DNA methylation profiles are associated with the progression stages of low-risk MDS and the distinct profiles are already present at the time of diagnosis, indicating that the DNA methylation profiles can be used to distinguish at diagnosis the more progressive forms of low-risk MDS.

The Battle of the Signals: an enhancer jungle competes with an imprinted lncRNA for regulation of the Kcnq1 domain

Bryant Schultz, Gwendolyn Gallicio & Nora Engel

Fels Institute for Cancer Research, Temple University School of Medicine

The imprinted Kcnq1 domain contains a differentially methylated region (KvDMR) in intron 11 of the Kcnq1 gene. The 92 kb Kcnq1ot1 non-coding RNA emerges from the unmethylated paternal KvDMR in antisense direction to Kcnq1, resulting in cis-repression of several neighboring genes. The KvDMR encompasses the Kcnq1ot1 promoter, CTCF sites and other DNA elements, but their individual contribution to the endogenous regulation of the domain is unknown. Paternal inheritance of a deletion of the Kcnq1ot1 promoter alone leads to derepression of the upstream Cdkn1c. Surprisingly, Kcnq1ot1 transcripts continue to emerge from alternative sites. During cardiac development, substantial chromatin reorganization results in discontinuous RNA production throughout the Kcnq1ot1 region in both wild-type and mutant mice, accompanied by loss of imprinting of Kcnq1. We report that there are a multitude of enhancers within the Kcnq1ot1 region, and present detailed mechanisms for a novel heart enhancer engaged in Kcnq1 expression and its conformational dynamics. Our results have important implications on tissue-specific imprinting patterns and how transcriptional mechanisms compete to maximize the expression of vital tissue-specific genes. We also report on stage-specific CTCF binding events that are involved in allelic chromosome configurations.

Functionally characterizing noncoding variants associated with psychiatric disorders

Jill E. Moore, Zhiping Weng
Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School

Schizophrenia, bipolar disorder, and major depressive disorder affect a significant percentage of the population. While the etiologies of these psychiatric disorders are unknown, each have strong hereditary components. Genome wide association studies have associated over 400 single nucleotide polymorphisms (SNPs) with these disorders. A majority of the associated SNPs lie in non-coding regions of the genome and our goal was to functionally characterize these non-coding SNPs. We used epigenomic data to define regulatory regions overlapping each of the SNPs. The SNPs are enriched in active enhancers in brain tissues as well as T-cells. Additionally, we analyzed transcription factor (TF) binding data and observed the SNPs are enriched in SP1, SP4, NFKB1, and EGR1 binding sites as well as these TF motifs. To determine the effect of the SNPs on TF binding, we remapped reads from ChIP-seq data and analyzed heterozygous loci. We identified 10 SNPs that are predicted to disrupt motifs for transcription factors with evidence of allele specific binding. Finally, we performed pathways analysis on predicted target genes. Immune pathways as well neural development pathways are enriched, demonstrating the complexity of psychiatric disorders and supporting the growing evidence of immune system dysregulation in psychiatric disorders.

Evaluating subtelomeric methylation from Illumina 450K data to identify age-related cognitive decline due to environmental exposures

Diddier Prada, Marc G Weisskopf, Avron Spiro III, Pantel Vokonas, Elena Colicino, Jia Zhong, Letizia Trevisi, Allan Just, Golareh Agha, Joel Schwartz, Lifang Hou, Nadereh Jafari, Andrea A. Baccarelli.

Harvard T.H. Chan School of Public Health

Background: Cognitive decline in older individuals is a public health concern. Environmental exposures to ambient pollution and lead (Pb) may accelerate age-related cognitive decline, probably due to increase oxidative damage and inflammation[1,2]. Also, previous studies have reported that subtelomeric regions are more methylated in Alzheimer's Disease (AD) individuals than in controls, and it has been proposed as a characteristic of AD[3,4]. AD is characterized by high rates of oxidative damage, which could be modifying telomere length through demonstrated changes in subtelomeric methylation[5]. Using data from 450K Illumina arrays, recent studies have recently described an association between subtelomeric methylation and telomere length in mid-life (44-45 years old) in a small subset of individuals (n=38)[6].

Objective: 1) Determine whether subtelomeric methylation is associated with cognitive decline in older individuals. 2) Determine the role of subtelomeric methylation in the association between black carbon (BC) and Lead (Pb) on age-related cognitive decline. 3) Determine the impact of socioeconomic disparities (income and race) in the association between subtelomeric methylation and cognitive function.

Methods: We evaluated individuals from the Normative Aging Study (NAS, n=645, Mean age=68 years old), on whom several cognitive tests are available from 1995. Subtelomeric methylation will be determined using the distal 4Mbp to each chromosome from the 450K Illumina data. Multivariate-adjusted linear mixed models will be used for statistical analysis.

Results: After multiple-comparison adjustment, 24 CpG subtelomeric sites have been found statistically significant associated to global cognitive function, and 11 for executive function. Other cognitive domains, as well as telomere length were not associated to changes in subtelomeric methylation. Median income was associated with methylation in 1 subtelomeric CpG site. Location and impact of those CpG sites as predictors of cognitive function will be discussed.

Funding: NIH - NIA (R01AG020727), NIH- NIEHS (R01ES02173; P30ES00002; R01ES01572), U.S.A. EPA (RD-83479801), Harvard Catalyst Award for Research in

Disparities.

References:

1. Power, M. C. et al. Traffic-Related Air Pollution and Cognitive Function in a Cohort of Older Men. Environ Health Perspect 119, 682–687 (2010).
2. Weisskopf, M. G. et al. Cumulative lead exposure and cognitive performance among elderly men. … 18, 59–66 (2007).
3. Guan, J.-Z., Guan, W.-P., Maeda, T. & Makino, N. Analysis of telomere length and subtelomeric methylation of circulating leukocytes in women with Alzheimer's disease. Aging Clin Exp Res 25, 17–23 (2013).
4. Guan, J.-Z., Guan, W.-P., Maeda, T. & Makino, N. The Subtelomere of Short Telomeres is Hypermethylated in Alzheimer's Disease. Aging Dis 3, 164–170 (2012).
5. Guan, J.-Z., Guan, W.-P., Maeda, T. & Makino, N. Effect of vitamin E administration on the elevated oxygen stress and the telomeric and subtelomeric status in Alzheimer's disease. Gerontology 58, 62–69 (2012).
6. Buxton, J. L. et al. Human leukocyte telomere length is associated with DNA methylation levels in multiple subtelomeric and imprinted loci. Sci Rep 4, 4954 (2014).

HLA-D regulatory haplotypes that potentiate systemic autoimmunity

Prithvi Raj[1], Ekta Rai[1], Ran Song[1], Benjamin E. Wakeland[1], Kasthuribai Viswanathan[1], Carlos Arana[1], Chaoying Liang[1], Bo Zhang[1], Ferdicia Carr-Johnson[1], Mitja Mitrovic[2], Graham B. Wiley[3], Jennifer A. Kelly[3], Bernard R. Lauwerys[4], Nancy J. Olsen[5], Chris Cotsapas[2], Christine K. Garcia[6], Carol Wise[7], John Harley[8], Swapan Nath[3], Judith A. James[3], Chaim O. Jacobs[9], Betty P. Tsao[10], David R. Karp[11], Quan-Zhen Li[1], Patrick M. Gaffney[3], and Edward K. Wakeland[1]

[1]Department of Immunology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA, [2]Department of Neurology, Yale School of Medicine, New Haven, CT, USA
[3]Arthritis and Clinical Immunology Program, Oklahoma Medical Research Foundation, Oklahoma City, OK 73104, USA, [4]Pôle de pathologies rhumatismales, Institut de Recherche Expérimentale et Clinique, Université catholique de Louvain, Brussels, Belgium, [5] Division of Rheumatology, Department of Medicine, Penn State Medical School, PA, USA, [6]Eugene McDermott Center for Human Growth & Development, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX 75335 USA, [7] Eugene McDermott Center for Human Growth & Development, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX 75335 USA, [8]Cincinnati Children's Hospital Medical Center and Cincinnati VA Medical Center, Cincinnati, OH, USA, [9]Department of Medicine, University of Southern California, Los Angeles, CA 90089, USA, [10]Department of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA, [11]Rheumatic Diseases Division, Department of Medicine, University of Texas Southwestern Medical Center, TX 75335, USA

Systemic lupus erythematosus (SLE) is an autoimmune disease caused by loss of humoral immune tolerance leading to the production of autoantibodies to a spectrum of self-antigens. Genetic predisposition is key for SLE susceptibility, however little is known about the nature or functional properties of causal genetic variants. We used targeted population sequencing to comprehensively characterize genetic variability at HLA-D in a panel of 1349 Caucasian SLE cases (773) and controls (576). The HLA-D region contains the strongest risk loci identified for SLE, with multiple alleles of both HLA-DR and -DQ showing strong associations. Sequence analysis of the 380 Kb segment spanning the BTNL2-DR-DQB2 region identified 15,261 common (MAF >0.05) genetic variants. Analyses of these sequence-defined HLA-D variations identified three independent risk-associated signals reaching genome wide significance. Subsequent analyses demonstrated that these disease-associated variations are imbedded in a series of stable haplotypes formed by multiple, ENCODE and eQTL-defined functional variations impacting the transcription of more than 20 genes that encode components of the antigen processing and presentation (APP) pathways of HLA class I and class II genes. Median neighbor joining analyses identified three HLA-D region regulatory

haplotypes forming a risk clade strongly associated with SLE, all of which contained eQTL variants that increased the transcription of HLA-DR, DQ, DP, and other elements of the APP pathway in multiple myeloid and lymphoid cell lineages. This risk clade contains all of the classical HLA-D class II alleles previously associated with SLE, indicating that the systemic upregulation of the APP pathway is a consistent feature of all SLE-associated HLA-D alleles. Our analyses demonstrate that such regulatory haplotypes have increased disease-associated odds ratios in comparison to the disease odds for maximal GWAS tagging SNPs in these loci. These findings are consistent with the hypothesis that the functional variations that underlie many common disease alleles form regulatory haplotypes that modulate the transcription of multiple cis and trans genes in immune system pathways and that their functional phenotypes are potent and complex.

Identification of cis-regulatory elements in the zebrafish genome

Hongbo Yang, Tingting liu, Yanli Wang, Feng Yue

The Department of Biochemistry & Molecular Biology, Pennsylvania State University College of Medicine, Hershey, PA 17033

Zebrafish is a good vertebrate model system in animal development and biomedical research. A thorough annotation of the zebrafish genome will be valuable to understand the function of the human genome because of sequence conservation. So far, most of the functional elements in the zebrafish genome have not been found and validated, especially for the cis-regulatory elements. It is now apparent that epigenetic modifications of both DNA and histone tails, and chromatin open status are equally important in the regulation of gene expression during development. Here, we apply the ATAC-seq, Chip-seq, DNA methylation assays to map functional elements of the zebrafish genome, including DNA and chromatin regulation in the zebrafish adult brain and heart tissue.

Visualizing three-dimensional organization and long-range interactions of the mammalian genome with the 3D Genome Browser

Yanli Wang, Gal Yaroslavsky, Tyler Derr, Lijun Zhang, Feng Yue

Department of Biochemistry and Molecular Biology, Penn State College of Medicine, Hershey, PA, 17033

The mammalian genome subscribes to a complex spatial organization that defines the three-dimensional interactions of potentially distant functional elements that control the regulation of transcription and replication. Recent advancements in sequencing and analysis techniques –specifically Hi-C, or high-throughput chromosome conformation capture– have revealed these interactions genome-wide at unprecedented resolutions. Unfortunately, navigating the Hi-C data remains a daunting feat for many biologists, as its O(n2) complexity for the already big data intrinsic to mammalian genomes poses a challenge to its analysis (time and memory usage), storage and transfer.
Our laboratory has developed and extended the functionality of the 3D Genome Browser (http://3dgenome.org), a web-based, intuitive and accessible browser of Hi-C data. The browser adopts a gene-centric approach: given the user input of gene symbol or genomic coordinates, it queries the Hi-C intra-chromosomal contact matrix for interactions from the regions in vicinity and display those values as a heatmap. Furthermore, our browser contextualizes the region by directly aligning it to the corresponding region as displayed by the established and familiar University of California Santa Cruz (UCSC) Genomic Browser while retaining its flexibility to customize genome tracks and load personalized UCSC sessions. While our browser contains several existing high-quality Hi-C datasets for a variety of human and mouse tissues for viewing, it also supports the browsing of user-generated Hi-C data. By converting Hi-C contact matrices into an indexed, binary format file and hosting it on any HTTP accessible server, the 3D Genome Browser could directly query and display the specified region without requiring the upload of entire files onto the server. In addition to the Hi-C heatmap, the contact matrix could also be visualized as virtual 4C, a linear plot detailing the number of interactions between a single genomic site of interest (bait or anchor locus) with other loci. Given the user input of gene or rsid, the virtual 4C plot with the TSS(s) or SNP as anchor locus would facilitate the identification of potential cis-regulatory elements. This feature would be supplemented with the inclusion of DNase I Hypersensitive Site (DHS)-linkage and ChIA-PET data, both currently under development. Currently, our browser boasts ~3000 views every month and receives frequent improvements in its user interface.
With our gene-centric, binary-file browser approach, the 3D Genome Browser improves the accessibility in browsing Hi-C data. With the visualization of the spatial organization and long-range interactions of particular genomic regions along with their genetic and epigenetic context, our browser seeks to drive hypothesis-generation about and enrich

the understanding of the intrinsic link between genomic organization and genetic regulation.

The role of the ENCODE Data Coordination Center

Jean M. Davidson[1], Esther T. Chan[1], Cricket A. Sloan[1], Eurie L. Hong[1], Venkat S. Malladi[1], Laurence D. Rowe[1], J. Seth Strattan[1], Marcus Ho[1], Nikhil R. Podduturi[1], Benjamin C. Hitz[1], Forrest Tanaka[1], Brian J. Lee[2], Matt Simison[1], W. James Kent[2], J. Michael Cherry[1]

[1] Stanford University School of Medicine, Department of Genetics, Stanford, CA
[2] University of California at Santa Cruz, Center for Biomolecular Science and Engineering, Santa Cruz, CA

In recent years there has been an influx in the quantity of publicly available, large genomic datasets from individual labs and large consortia. Here we describe the efforts of the Data Coordination Center (DCC) in improving the accessibility of data available for the Encyclopedia of DNA Elements (ENCODE) project. ENCODE is a collaborative effort to generate a comprehensive catalog of functional elements in human and mouse genomes. The ENCODE database currently includes more than 40 experimental techniques in over 400 tissue types and cell lines to analyze DNA and RNA-binding proteins, transcription and chromatin structure. All experimental data and computational analyses of these data are submitted to the DCC for validation, tracking, storage, and distribution to the scientific community. To ensure that the data generated by the production labs and the analysis performed on these data are accurately represented, the ENCODE DCC works closely with members of the Consortium groups to capture structured metadata related to experimental conditions, data quality metrics, and analysis methods. These experiments can be accessed via the ENCODE portal (http://www.encodeproject.org). Portal users query the database by searching for specific metadata terms, such as “p53”, or “K562” or by utilizing faceted searching of the structured metadata. The portal also supports the visualization of certain data files by launching a Genome Browser track hub. Data files can be downloaded either directly from the experiment pages at the portal or via bulk download by programmatically accessing the ENCODE REST API. By providing direct data downloads based on flexible and powerful search capabilities that rely on highly organized metadata, the DCC strives to expand the access of ENCODE data to the scientific community.

Use the UCSC Genome Browser to Visualize and Analyze your Genomic Data

Pauline A. Fujita, Matthew L. Speir, Angie S. Hinrichs, Maximilian Haeussler, Brian J. Raney, Hiram Clawson, Kate R. Rosenbloom, Galt P. Barber, Jonathan Casper, Steve Heitner, Luvina Guruvadoo, Brian T. Lee, Ann S. Zweig, Donna Karolchik, Robert M. Kuhn, David Haussler, W. James Kent

The UCSC Genome Browser (http://genome.ucsc.edu) is a free, web-based tool that integrates and displays genomic data from a wide variety of sources, including GenBank, ENCODE, UCSC and many others. We provide several tools to help users upload their own data and view it alongside this genomic information or export the data for analysis with other applications. Large genomic data sets and custom genome assemblies can be uploaded and displayed using the browser's data hub tools. If you need to view private data, such as protected patient data from a clinical trial, Genome Browser in a Box (GBiB) allows you to run your own private copy of the Genome Browser on your own computer. The new Data Integrator tool lets you quickly combine input from up to five genome-wide data sets, including your own data uploaded through custom tracks or track hubs, and then export a customized output set based on intersections with a primary track. We are continually working to extend our toolset to allow users to explore their data in new and unique ways.

Tracking data provenance at the ENCODE DCC

Eurie L Hong[1], Venkat S Malladi[1], Benjamin C Hitz[1], Esther T Chan[1], Jean M Davidson[1], Timothy R Dreszer[1], Marcus Ho[1], Brian T Lee[2], Nikhil R Podduturi[1], Laurence D Rowe[1], Cricket A Sloan[1], J. Seth Strattan[1], Forrest Tanaka[1], W. James Kent[2], J. Michael Cherry[1]

[1]Stanford University, Genetics, Stanford, CA
[2]University of California, Santa Cruz, Center for Biomolecular Science and Engineering, Santa Cruz, CA

The provenance of experimental reagents and transparency of computational analyses are essential to compare, reproduce, and interpret experimental data. The task of tracking this information consistently across diverse sequencing assays can be especially challenging in large projects like the ENCODE (ENcyclopedia Of DNA Elements) Consortium that perform 40+ genomic assays using 400+ cell and tissue types. The identification of a transcription factor binding site or the quantification of a transcript's expression level is dependent on the software versions, the parameters used when running that software version, which files were used, the library preparation methods, and how the biological samples were selected or obtained. To capture the provenance of experimental methods and computational results, the ENCODE DCC (Data Coordination Center) has created a
rich data model that represents how experiments were performed, what software and pipelines were used, and which files were analyzed. These details of the experimental and computational methods, known as metadata, can then be used to identify related data for further analysis, interpret the results of the assays, and allow reproducibility of pipelines that are run to generate the data. All metadata and data generated by the ENCODE Consortium are freely available at at the ENCODE Portal (https://www.encodeproject.org/).

Annotating non-genic regions in Ensembl

Emily Perry, Daniel R. Zerbino, Nathan Johnson, Thomas Juettemann, Steven Wilder, David Richardson, Laura Clarke and Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

Ensembl is one of the world's leading sources of information on the structure and function of the genome. It already provides an up-to-date, comprehensive and consistent database that brings together genome sequences, genes, non-coding RNAs, known variants, etc. As personal genomics develops and automated annotation becomes commonplace, there will be an increasing need for such infrastructure.

However, it is very common for GWAS to return statistically significant hits that are hard to interpret because they fall outside of coding regions. Although Ensembl has provided functional annotation of non-coding regions for a few years, we are currently redesigning from the ground up the way we define regulatory regions of the genome. Large projects such as ENCODE and Roadmap Epigenomics measure epigenomic marks through a wide spectrum of experimental assays across a diversity of cell types. They paved the way for more specialized endeavors such as BLUEPRINT, which focuses on hematopoietic differentiation.

Ensembl provides a summary of the available public datasets produced by these projects by processing them through a unified pipeline and making them available through a single interface. The data is synthesised into the Regulatory Build, which defines functionally active regions across the human genome (on both the GRCh37 and GRCh38 assemblies) and mouse cell lines, assigning them a function wherever possible. We are currently collaborating with a number of data producing teams to broaden our coverage of cell types. Having defined a set of active regions along the genome, their activity levels can then be reliably determined in a new sample with a reduced set of assays.

Our goal is to progressively develop an annotation of the genome into regulatory elements, akin to the gene annotations that Ensembl already produces. To enrich this annotation we are looking at an array of technologies and assays to determine the links between enhancers and their target genes, such as eQTLs or Hi-C data. This will ultimately provide medical researchers greater resolution when prioritizing candidate variants for functional relevance.

In parallel, we are developing tools for basic research in epigenomics. For example, the WiggleTools browser allows users to remotely compute statistics on large collections of data, as produced for example by the BLUEPRINT project, without downloading data or software. Our simplified representation of epigenomes can also be used to quickly compute differences between cell types, and establish clear differentiation pathways. This opens the way for rapid identification of cell type based on epigenomic markers.

Demonstration of ENCODE Universal Pipelines for RNA Seq

Ben Hitz, ENCODE DCC, The ENCODE Consortium

We have developed universal analysis pipelines for several core ENCODE assays. We here demonstrate the pipeline for RNA-seq mapping and quantification using STAR, RSEM, and tophat. The pipelines have been implemented in the cloud at DNANexus, which provides a handy user interface for processing RNA-Seq and other experiments.

The ENCODE ChIP-seq Pipeline: Shareable, Scalable, Replicable, Cloud-Based Analysis

J Seth Strattan, Esther T Chan, Jean M Davidson, Tim Dreszer, Ben C Hitz, Venkat S Malladi, Nikhil R Podduturi, Laurence D Rowe, Cricket A Sloan, Forrest Tanaka, Nathan Boley, Anshul Kundaje, Eurie L Hong, J Michael Cherry

Stanford University School of Medicine, Department of Genetics, Stanford, CA

From Ammon's horn to zone of skin, members of the ENCODE Consortium have measured RNA quantity, RNA-protein interactions, DNA-protein interactions, DNA methylation, replication timing, chromatin structure, and histone modifications in over 4,000 experiments on more than 400 cell or tissue types. The ENCODE Data Coordination Center (DCC) have built a new web resource, the ENCODE Portal, to distribute the results of these experiments. Web-based faceted browsing and search are supported, as is programmatic access through the ENCODE REST API. The ENCODE Data Analysis Center (DAC) have specified uniform processing pipelines for four ENCODE datatypes: ChIP-seq, RNA-seq, DNase-seq, and whole-genome bisulfite sequencing. The DCC have implemented these pipelines and deployed them to a cloudbased
platform. The results of these analyses and metadata describing them are distributed through the ENCODE Portal, and illustrate general methods of accessing and interpreting ENCODE data. The ENCODE Portal is https://www.encodeproject.org/. The DCC codebase is freely available at https://github.com/ENCODE-DCC/.

Genome-wide positioning of bivalent mononucleosomes

Subhojit Sen(1,2) Kirsten F. Block (1), Alice Passini (3), Stephen B. Baylin (1), Hariharan Easwaran (1)

(1) Department of Oncology and The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, The Johns Hopkins University School of Medicine, Baltimore, MD 21231, USA
(2) UM-DAE Centre for Excellence in Basic Sciences, University of Mumbai Kalina Campus, Santacruz (East), Mumbai 400098, India.
(3) Department of Electrical, Electronic and Information Engineering "G. Marconi" (DEI), via Venezia 52, 47521 - Cesena (FC), University of Bologna, Italy.

TBA

Dynamic Enhancer Landscapes during Pancreatic differentiation of human ES cells

Feng Yue[1], Allen Wang[2], Yan Li[3], Maike Sander2, Bing Ren[3]

[1] The Department of Biochemistry & Molecular Biology, Pennsylvania State University
[2] Ludwig Institute for Cancer Research, UC San Diego School of Medicine
[3] Department of Pediatrics, University of California, San Diego

Temporal and spatial-specific gene transcription is tightly controlled by *cis*-regulatory elements such as promoters and enhancers. Here we show that epigenetic priming of enhancers signifies developmental competence using during pancreatic differentiation of human ES cells. We performed RNA-Seq, GRO-Seq and ChIP-Seq for H3K4me3, H4K4me1 and H3K27Ac in each developmental stages, including hESCs, definitive endoderm (DE), primitive gut tube (GT), posterior forgut (FG), and pancreatic endoderm (PE). We observed that poised enhancer state could be used to predict the ability of developmental intermediates to respond to inductive signals. We further find that lineage-specific enhancers are first recognized by transcription factors involved in chromatin priming, while subsequent recruitment of lineage- inductive transcription factors leads to enhancer and target gene activation. Our results identify acquisition of a poised chromatin state at enhancers as a general mechanism by which progenitor cells gain the competence to rapidly activate lineage-specific genes in response to inductive signals.