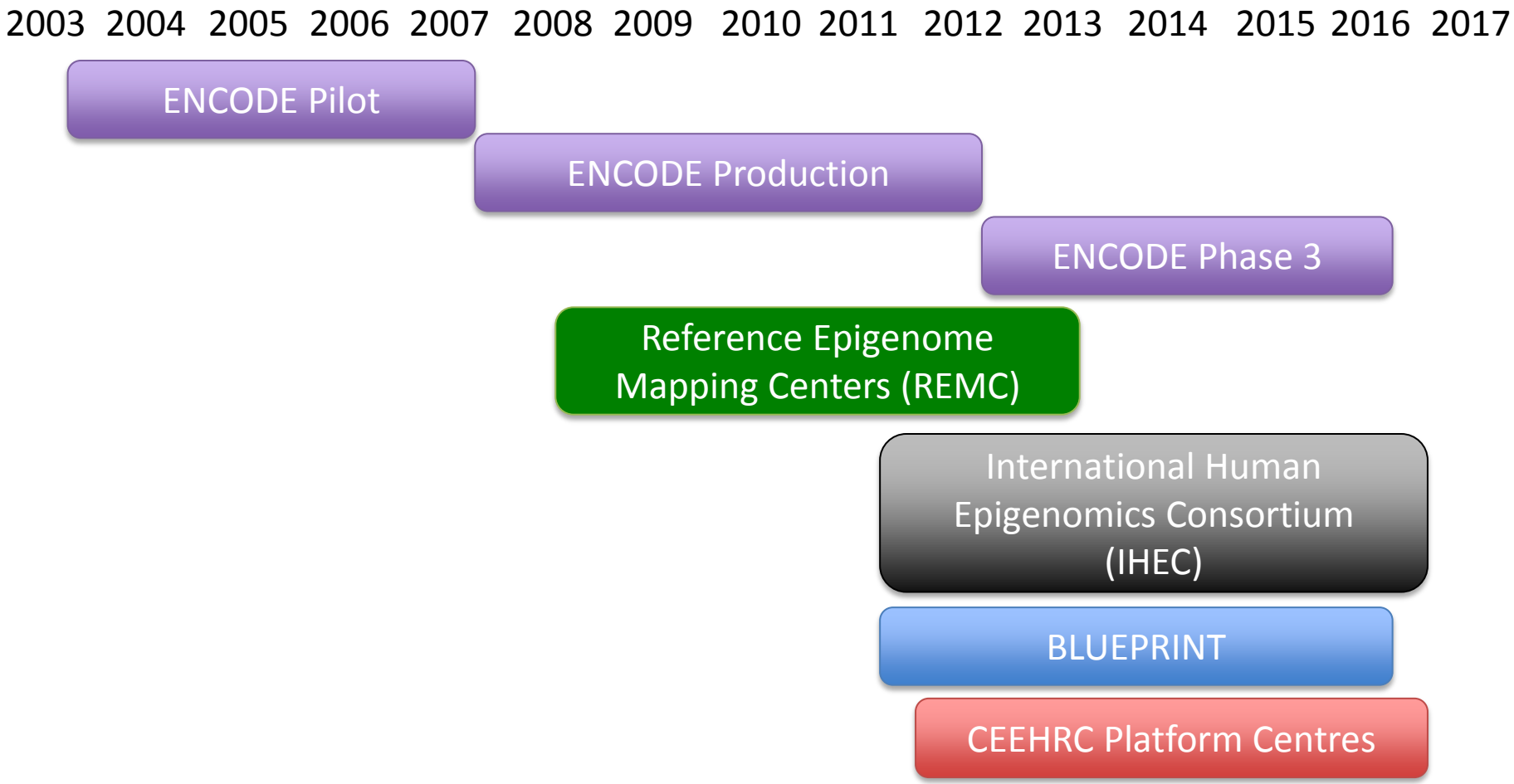
Genome Function Circa 2016: Updates from ENCODE and Related Projects

NHGRI Workshop
From Genome Function to Biomedical Insight:
ENCODE and Beyond
10-11 March 2015

Mike Snyder
Mike Pazin
Dan Gilchrist

National Human Genome Research Institute

Timeline For ENCODE, REMC, IHEC, BLUEPRINT, And CEEHRC



REMC

- Reference Epigenome Mapping Centers
- Goal is to generate a public community resource of human epigenomic data
- Funded by NIH Common Fund
- Publication package 18 February 2015 in Nature
- NIH Common Fund Epigenomics includes REMC, as well as disease-focused projects, analysis projects, and technology development projects

REMC

- Built on years of gene regulation studies
- Data production effort completed
- High assay diversity
  - DNA methylation
  - Histone modification ChIP-seq
  - DNase I hypersensitivity
  - Gene expression

- High sample diversity
  - Broad distribution of organs and tissues
  - Fetal and adult samples

IHEC

- International Human Epigenome Consortium
- Goal is understand the role of the epigenome in health, disease, environment/lifestyle, and aging
- Support prevention, diagnosis, and treatment of disease
- Working Groups include bioethics, assay standards, metadata standards, and data ecosystems
- Data collection:
  - Transcriptome (mRNA-seq)
  - DNA methylation (WGBS)
  - Histone modification (6 marks + control)
- IHEC Data Portal http://epigenomesportal.ca/ihec/

BLUEPRINT

- Goal is to generate Epigenome resources using a clearly defined set of primarily human samples from healthy and diseased individuals
- Funded by European Union
- Diseases include leukemias/lymphomas and autoimmune disease (Type 1 Diabetes)
- Functional genomics analysis
- Publication package 26 September 2014 in Science
- Discovery and validation of epigenetic markers for diagnostic use and by epigenetic target identification

CEEHRC Platform Centres

- Canadian Epigenetics, Environment and Health Research Consortium overall goal is to translate epigenetic discoveries into health benefits
- Platform Centres (Vancouver and Montreal) are key to building epigenomic capacity in Canada
  - Blood, breast, brain, thyroid
  - Cancer
  - Developing and hosting IHEC Data Portal http://epigenomesportal.ca/ihec/
  - Multi-agency, pan-Canadian initiative led by the Canadian Institutes of Health Research (CIHR)

IHEC

- High assay diversity
  - Transcriptome (mRNA-seq)
  - DNA methylation (WGBS)
  - Histone modification (6 marks + control)
  - Additional assays are typical

- High sample diversity
  - Broad distribution of organs and tissues for the consortium as a whole
  - Focused distribution of samples within the biological areas of each project

- Some potential to detect individual variation

PsychENCODE

www.psychENCODE.org

- PsychENCODE Consortium includes 11 NIMH-funded projects, five of which are funded through RFA-MH-14-020
- Goal is to identify non-coding functional genomic elements in human brain and elucidate their role in the etiology of mental disorder.
- Large-Scale (> 1000 human brains), integrative omic analyses across psychiatric phenotypes and brain regions
- Schizophrenia, Autism, Bipolar Disorder, Brain development

PsychENCODE

www.psychENCODE.org

- High assay diversity:
  - Histone modification
  - Gene expression
  - DNase, ATAC-Seq
  - Proteome
  - eQTL
  - Functional assays
- High potential to detect variation across individuals
- Deep collection of biosamples within one particular organ

Genomics of Gene Regulation (GGR)

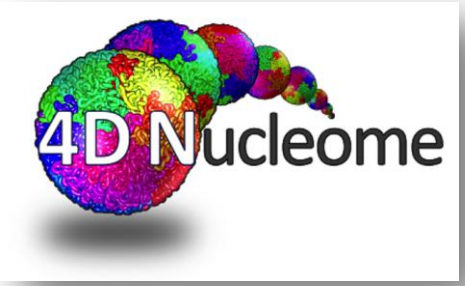
- Goal is to determine how to construct predictive, accurate gene regulatory network models from genomic data
- 5 projects funded by NHGRI, December 2014-2017
- Keratinocyte differentiation, T cell inflammatory response, innate immune response, steroid response
- Transcriptional regulation, post-transcriptional regulation
- Data Repository: ENCODE DCC

Genomics of Gene Regulation (GGR)

- Goal is to determine how to construct predictive, accurate gene regulatory network models from genomic data
- Within each project, closely related cell fates/states are compared
- High assay diversity (not standardized across projects)

Function Of Non-coding Variants

- Goal is to develop better computational tools to prioritize non-coding variants associated with disease
- About 90% of common variants associated with disease lie outside protein-coding regions
- Genetic variation in non-coding regions is known to cause and modify human disease
- Awards in process at NHGRI and NCI



4D Nucleome

- Goals are to understand:
  - the principles of nuclear organization
  - the role of nuclear organization in cellular function, development, and disease
- Technology development, reference maps, and predictive modeling of structure/function relationships
- Imaging and genomic assays
- Funded by NIH Common Fund
- Projects could start as early as the end of this fiscal year

FANTOM
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

- Functional Annotation of the Mammalian Genome
- Goal is identification of functional elements in mammalian genomes
- Human and Mouse Tissues, Primary Cells, and Cell Lines
- Data Collection
  - CAGE (Cap Analysis of Gene Expression)
  - Transcriptomic
- Data Repository (http://fantom.gsc.riken.jp/data/)
- cDNA Clone Bank
- Data Analysis and Integration
  - Functional Element Annotation (promoters, enhancers)
  - Examination of Cell-state transitions
  - Gene Expression Mechanisms

- Very focused range of assays – CAGE and transcriptomic
- Very high sample diversity – many tissue and cell types
- High cell state/fate diversity
  - Many biological stimuli
  - Time-resolved data collection

GTEx

- Genotype-Tissue Expression Program
- Goal is to provide an atlas of gene expression across human tissues; to provide resource for exploring how genetic variation modulates gene expression
- Numerous human tissues (n = ~30)
- Numerous donors (n = ~900)
- Data Collection
  - Transcriptomic
  - WGS/WES
  - Limited Epigenomic, Proteomic
- Data Portal (http://www.gtexportal.org), dbGaP
- Data Analysis: eQTL Browser

GTEx

- Large sample size provides unique resource to study inter-individual variation
- High sample/tissue diversity
- Focused set of assays – transcriptomic, WES, WGS

LINCS

- Library of Integrated Network-based Cellular Signatures
- Goal is to create network-based understanding of biology; elucidation of cellular signatures through systematic perturbation experiments and computational analysis
- Primary cells, cell lines, iPS, cardiomyocytes, neurons
- Data collection
  - Transcriptomic
  - Phosphoproteomic
  - Epigenomic
  - Imaging
- Data Portal (http://www.lincsproject.org/data/)
- Data Integration and Analysis
  - Reference set of query-able cellular signatures
  - Tools for generating cellular signatures

LINCS

- Very high cell state/fate diversity – many biological, chemical stimuli, some time-resolved measurements

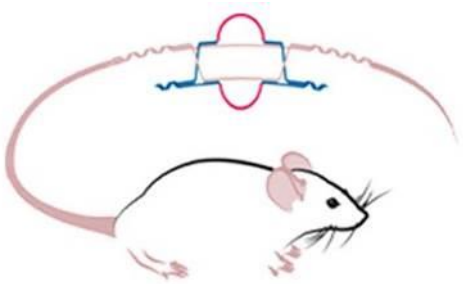
- Initial high assay diversity followed by data-guided focusing to most information-rich assays

- High sample diversity – many cell types, though limited number of cell types subjected to all assays

TCGA

- The Cancer Genome Atlas
- Goal is to improve cancer care, by accelerating the understanding of the molecular basis of cancer through genome analysis technologies
- Tumor and Normal Samples (n = ~10,000 tumor/normal pairs)
- Data collection
  - WGS and WES
  - Transcriptomic
  - Epigenomic
  - Proteomic
- Data Portal (https://tcga-data.nci.nih.gov/tcga/), CGHub (https://cghub.ucsc.edu/)

- International Cancer Genome Consortium
- Goal is to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 tumor types of clinical and societal importance across the globe
- Tumor and Normal Samples
- Data collection
  - WGS and WES
  - Transcriptomic
  - Epigenomic
- Data Portal (https://dcc.icgc.org/)



KOMP2

- Knockout Mouse Phenotyping Program
- Goal is to provide broad, standard phenotyping of genome-wide collection of mouse knockouts
- International Mouse Phenotyping Consortium (IMPC) Member
- Phenotypic Data Collection including
  - Morphological
  - Histopathological
  - Behavioral
- Data Repository (http://www.mousephenotype.org)

ENCODE and Related Projects – Different Strategies For Exploring Functional Genomics Space

Project	Assay Diversity	Sample Diversity	Number of Individuals	Cell Perturbation
ENCODE	++++	++++	+	+
REMC/IHEC	+++	++++	++	+
PsychENCODE	++++	+	+	+
GGR	+++	++	+	+++
4DN	TBD	TBD	TBD	TBD
FunVar				
FANTOM	+	++++	+	+++
GTEx	++	+++	+++	+
LINCS	++	+++	+	++++
TCGA	++	++++	+++	+
KOMP2	++	++++	+	+

Projects Grouped By Workshop Topics

Topic 1: Unbiased Mapping/Annotation	Topic 3: Basic Biological Studies	Topic 2: Disease-Focused Studies
ENCODE REMC GTEx KOMP2	LINCS GGR	TCGA/ICGC IHEC BLUEPRINT PsychENCODE FunVar
FANTOM 4D Nucleome		
	Common Fund Epigenomics	

Acknowledgments

NHGRI

- Lisa Brooks
- Julie Coursen
- Elise Feingold
- Adam Felsenfeld
- Colin Fletcher
- Peter Good
- Carolyn Hutter
- Hannah Naughton
- Ajay Pillai
- Jeff Schloss
- Heidi Sofia
- Jeff Struewing
- Simona Volpi

CIHR

- Eric Marcotte

NCI

- Judy Meitz
- Stefanie Nelson

NIDA

- John Satterlee

NIDDK

- Olivier Blondel

NIMH

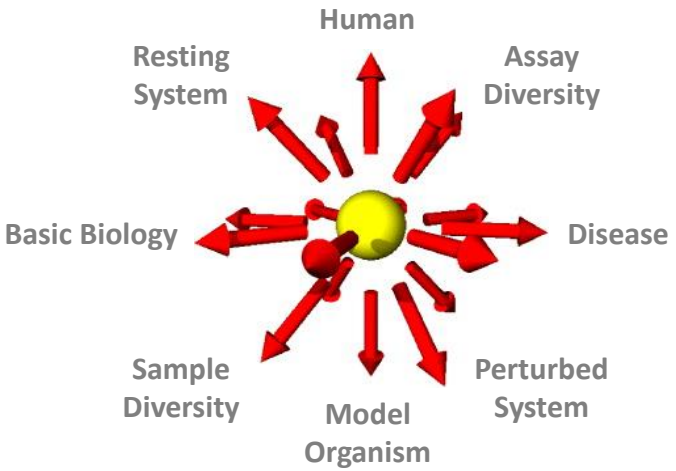
- Geetha Senthil

RIMLS

- Henk Stunnenberg

RIKEN

- Piero Carninci



National Human Genome Research Institute