Identify Disease-Causal Genes from GWAS Loci by 3D Genome Structure, Regulatory Landscapes & Deep Learning

Yi-Hsiang Hsu, MD, ScD

YongSheng Huang, Ph.D

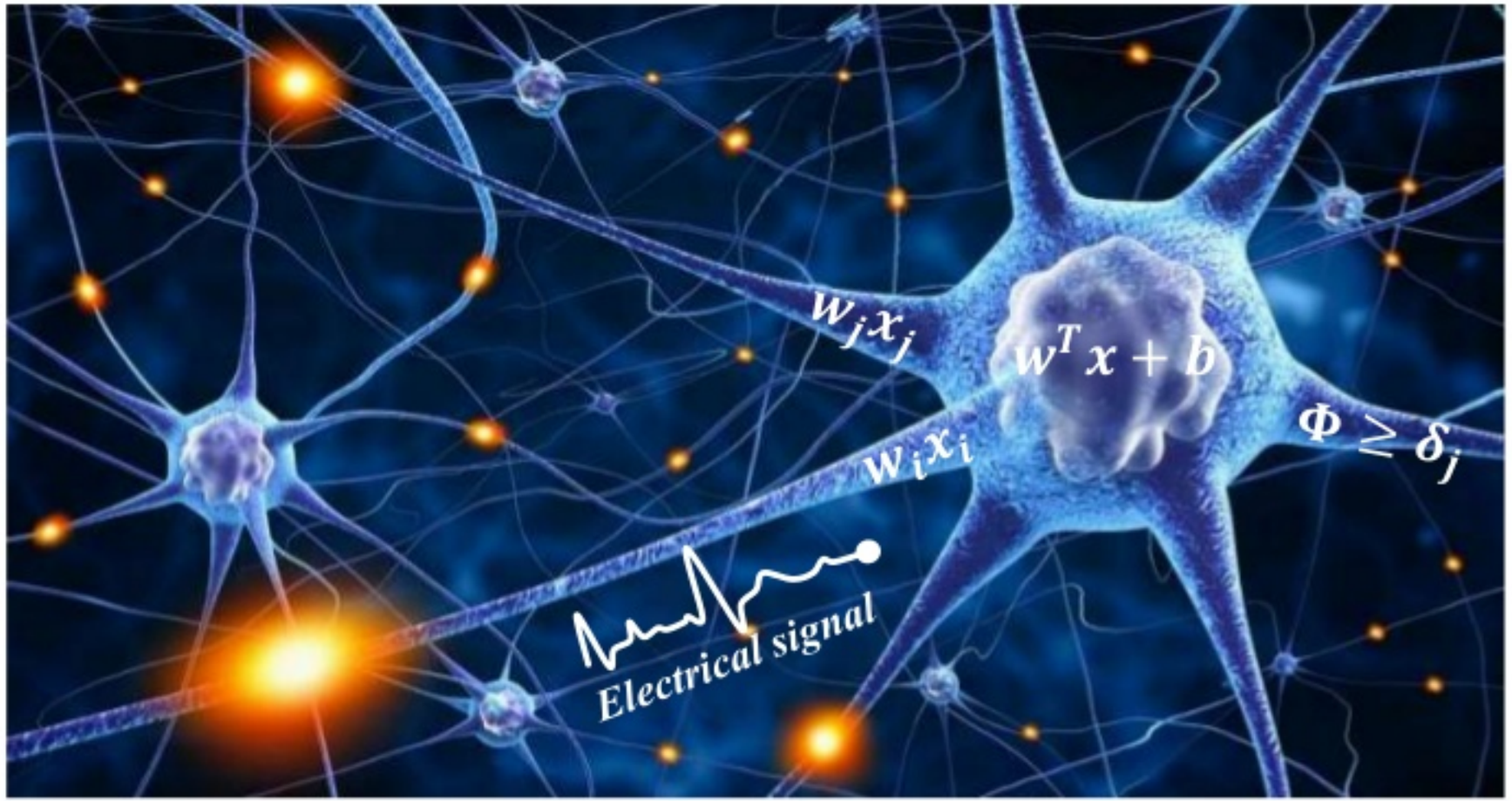
BROAD INSTITUTE

HARVARD MEDICAL SCHOOL

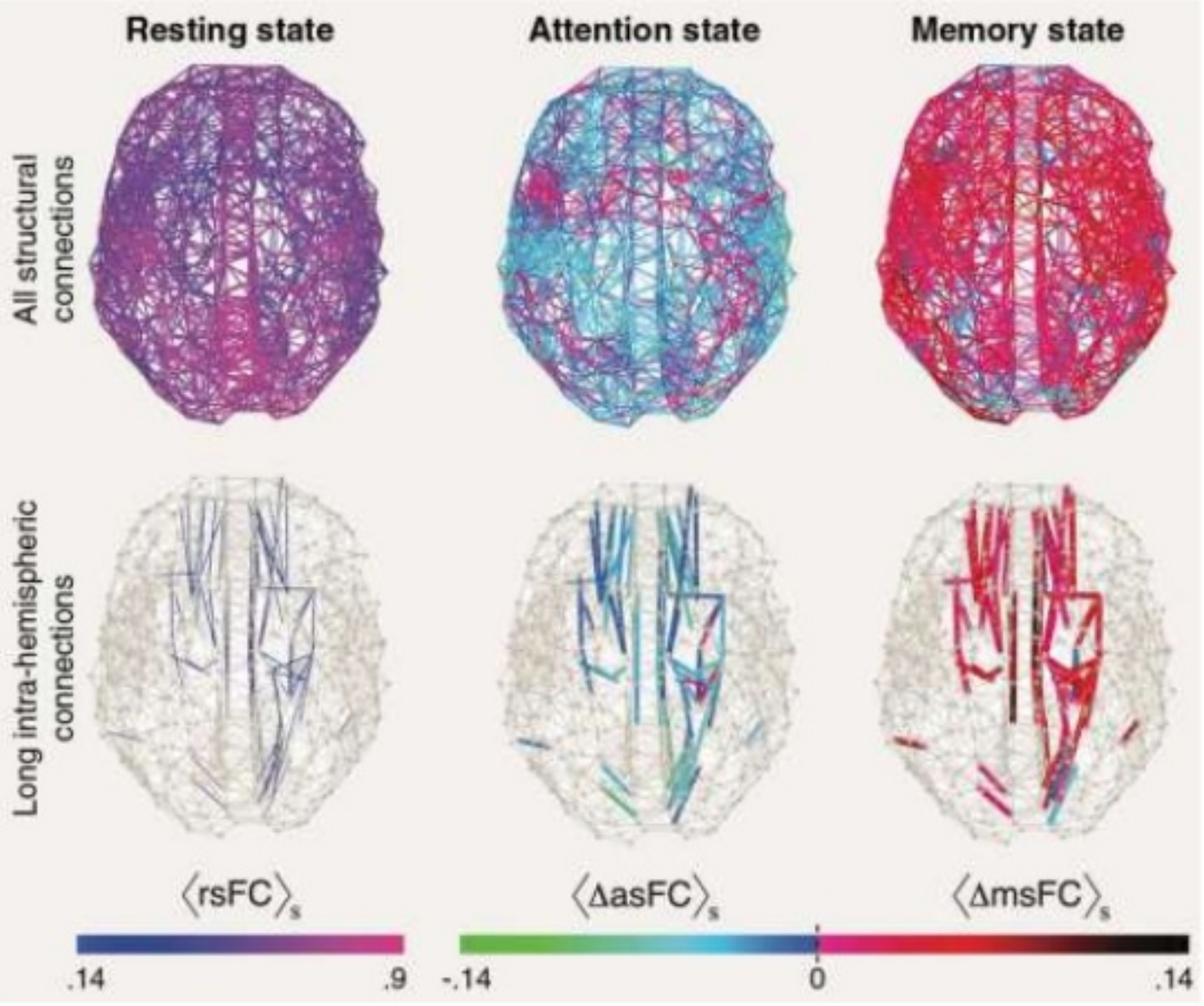
databricks

Deep Learning: The Inspiration

" the deepest concepts in mathematics are those which link one world of ideas with another"
---- Freeman Dyson



Deep Learning: The Natural Form



Science 2013 Nov

Deep-Learning: The Renaissance

The New York Times

Scientists See Promise in Deep-Learning Programs

by **JOHN MARKOFF** Nov. 23, 2012

In the 1960s, ….. believed that a workable artificial intelligence system was just 10 years away. In the 1980s, a wave of commercial start-ups collapsed, leading to what some people called the "**A.I. winter.**"

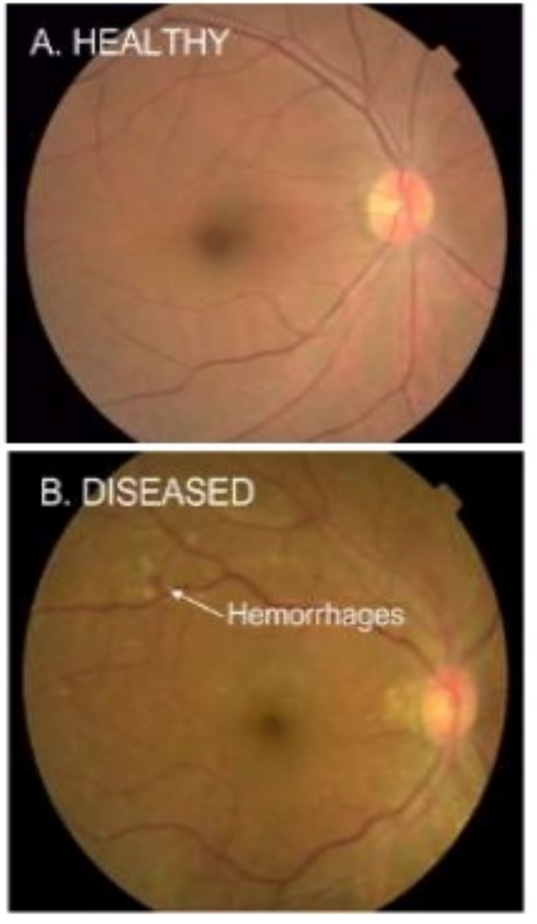
But recent achievements have impressed ….. In October, for example, a team of graduate students studying with the University of Toronto computer scientist Geoffrey E. Hinton won the top prize in a contest sponsored by Merck to design software to **help find molecules that might lead to new drugs**.

Deep Learning: Impact on Medicine

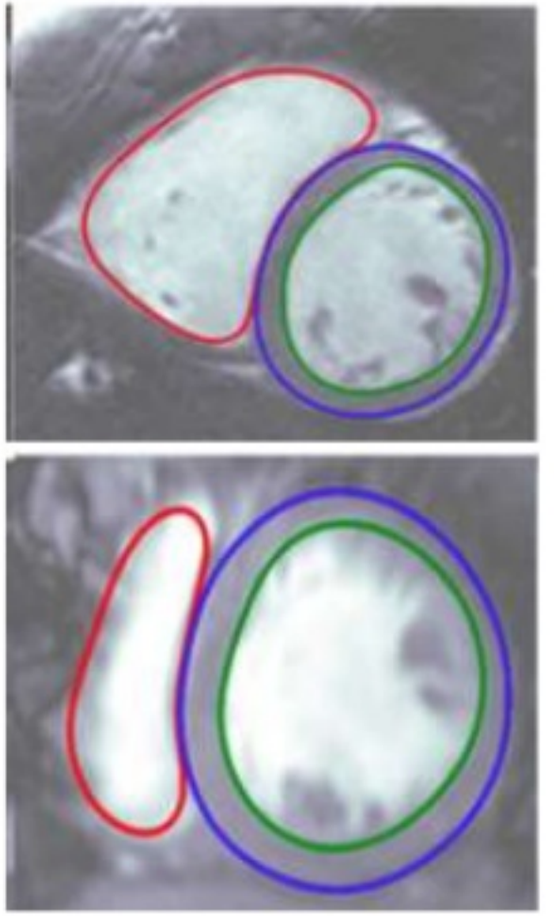
On par performance as 21 board-certified pathologists

Nature 2017 Feb

>90% specificity and sensitivity as board-certified ophthalmologists



Artery's Cardio DL wins FDA approval for clinical diagnosis (10-sec vs. 1hr)



databricks

Deep Learning: The New Disruption

Can we leverage DL to identify genetic variants that are disease causal, so that we can treat diseases at its root level per individual patient ?

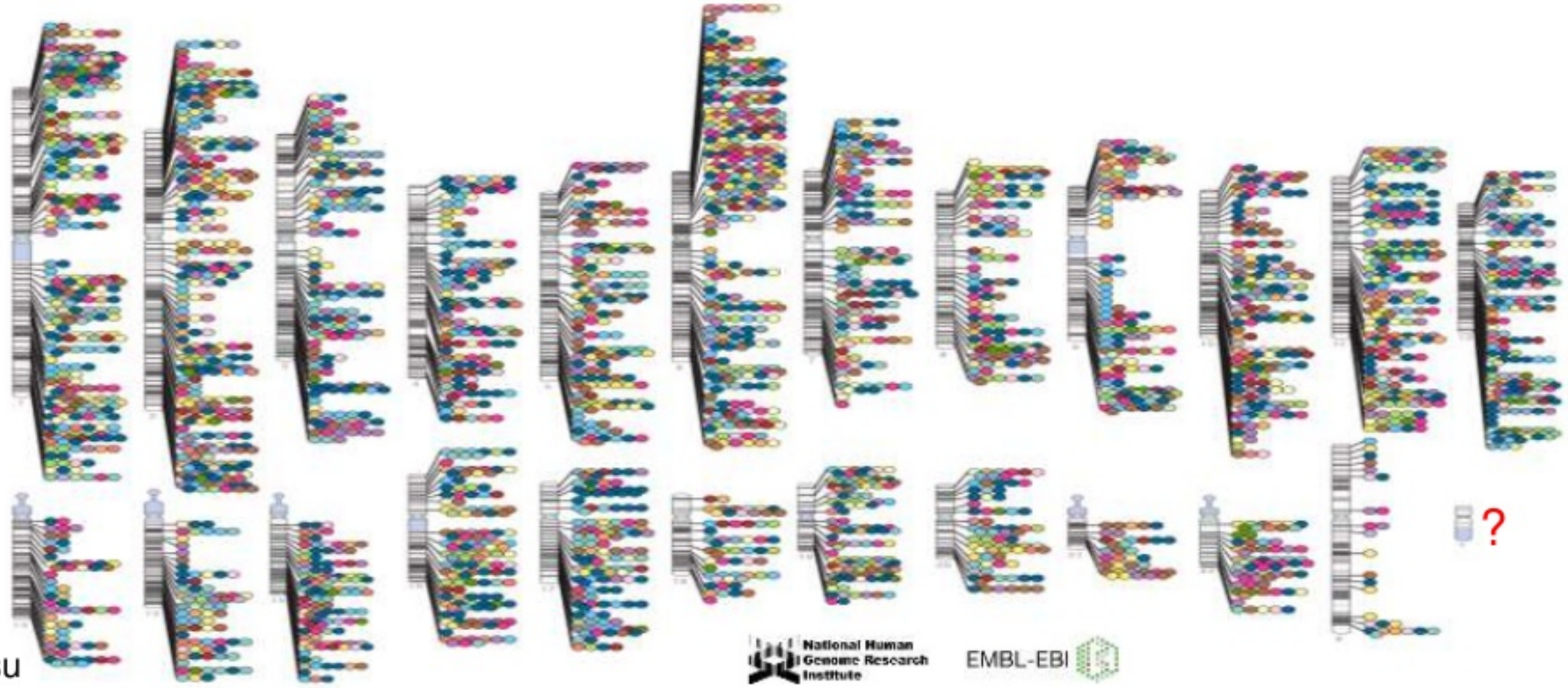
Yi-Hsiang Hsu, MD, ScD

yihsianghsu@hsl.harvard.edu
yihsiang@broadinstitute.org

Director & Associate Professor, HSL GeriOmics Center, Harvard Medical Sch
Program for Quantitative Genomics, Harvard School of Public Health
Associate Member, BROAD Institute of MIT and Harvard
NHLBI Framingham Heart Study Investigator

Genome-Wide Association Studies (GWAS) Catalog

- Identified ~13,000 genetic variants (single nucleotide mutations/ polymorphisms) to be associated with ~2,000 diseases/phenotypes

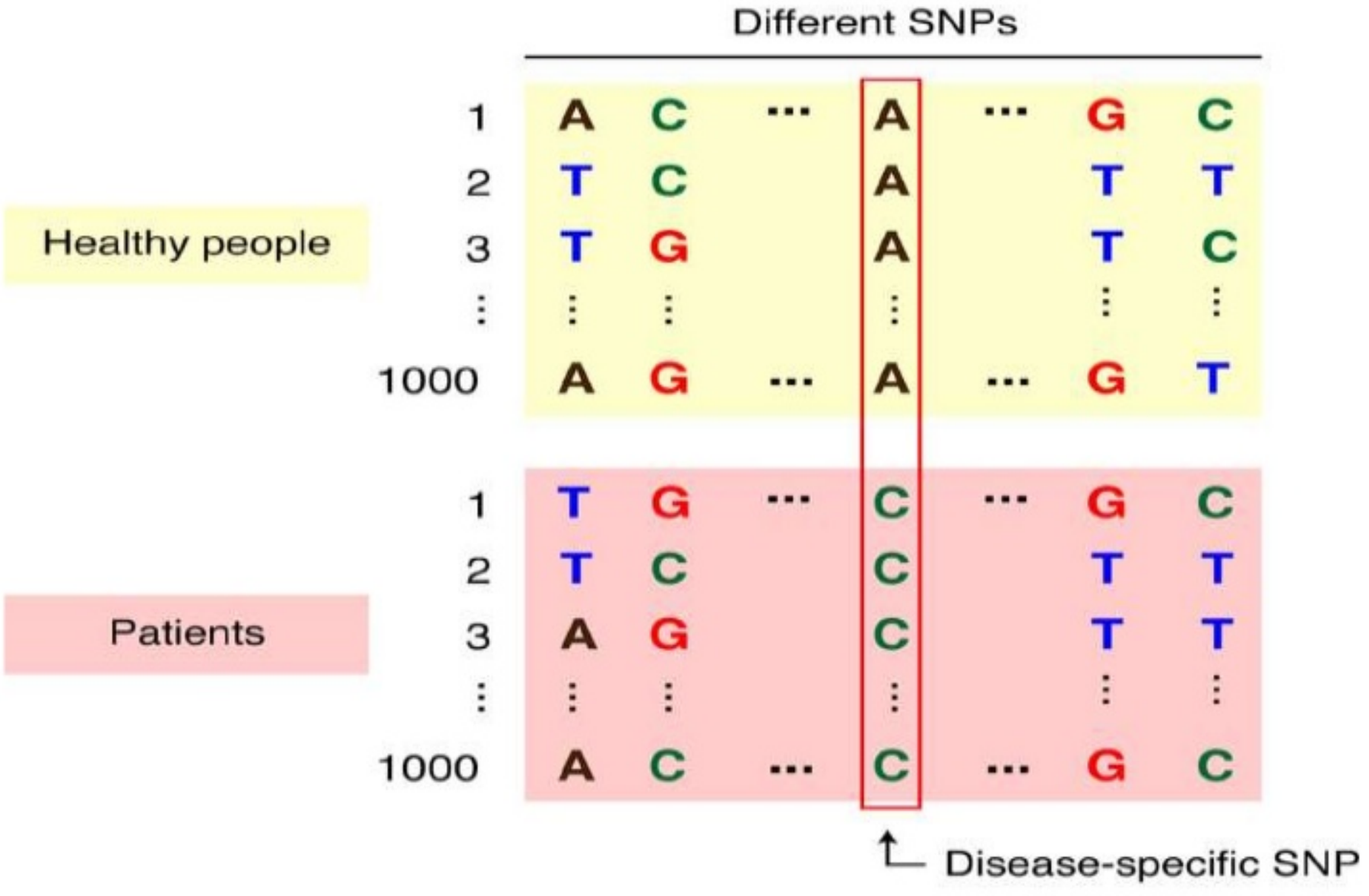


Y-H Hsu

National Human Genome Research Institute

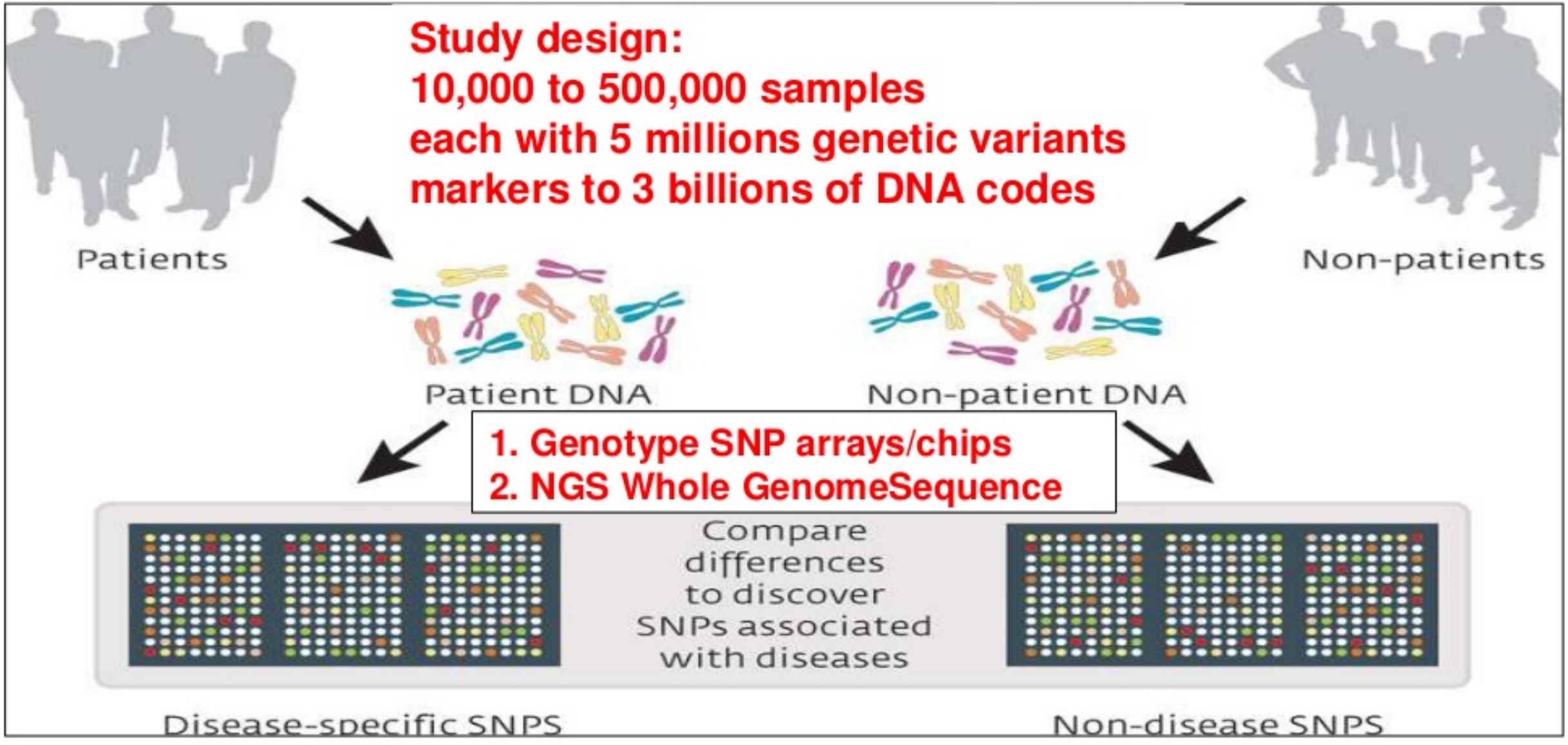
EMBL-EBI

Genome-Wide Association Scans



Y-H Hsu

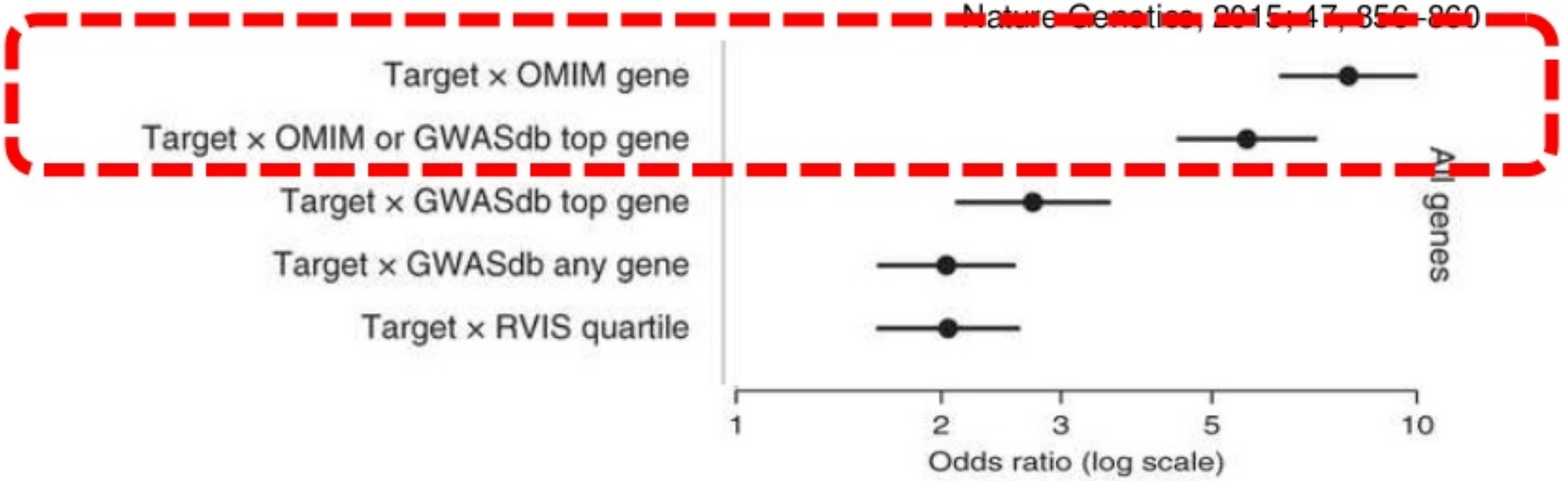
GWAS (Whole Genome Association) Scans



Y-H Hsu

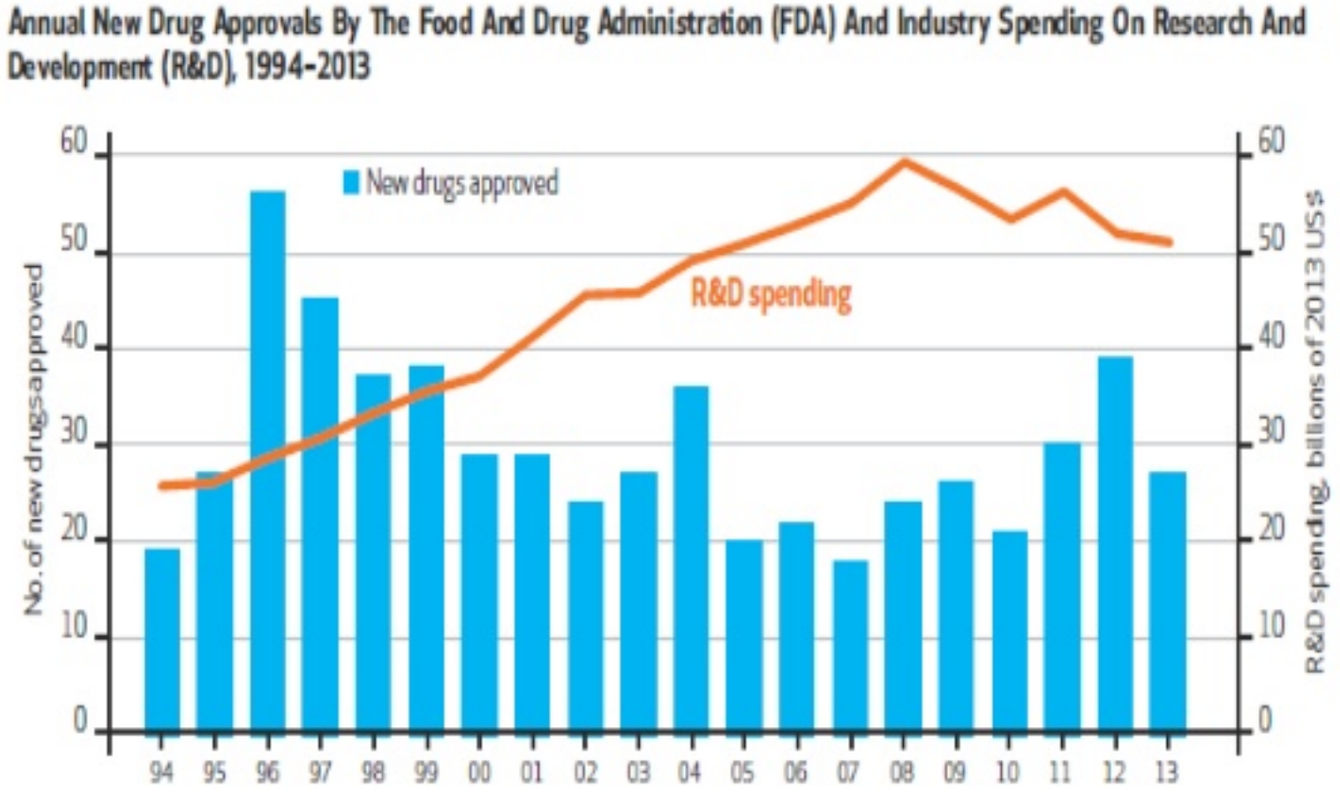
% Successfully Approved Drugs & Human Genetics

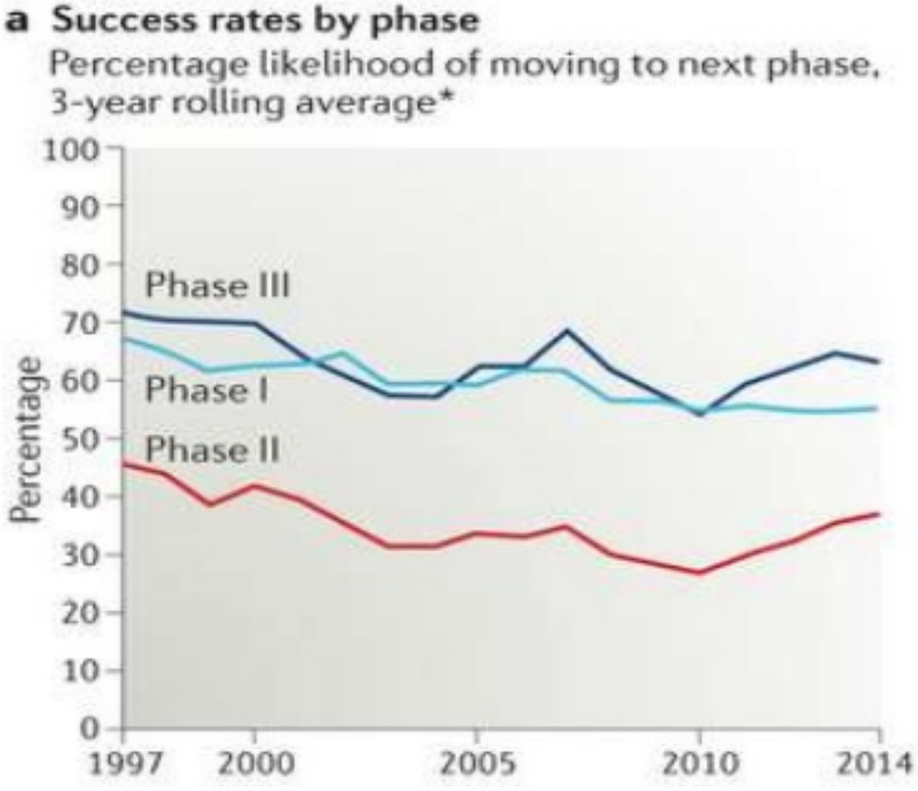
- The impact on medical care from GWAS could potentially be substantial



- FDA approved drugs with human genetic information are 5~10X more likely to be successful
- Failure targets at each drug development stage (pre-clinical, phase I, II, III) are more likely to be those targets without genetic validation

R&D Spending on New Drugs ≠ Drug Approvals

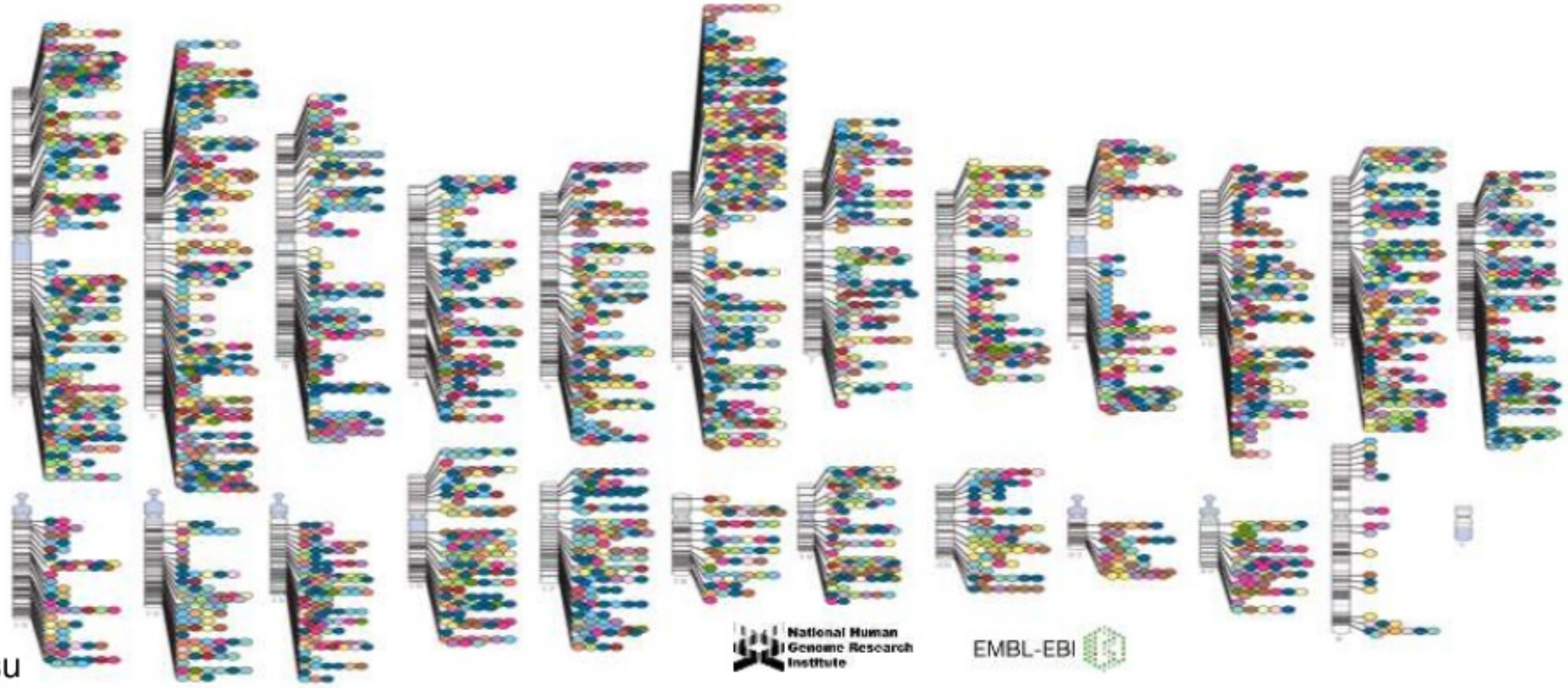




- New a better drug development pipeline
- Utilizing human genetic information/validation is the key

Genome-Wide Association Studies (GWAS) Catalog

- Identified ~13,000 genetic variants (single nucleotide mutations/ polymorphisms) to be associated with ~2,000 diseases/phenotypes

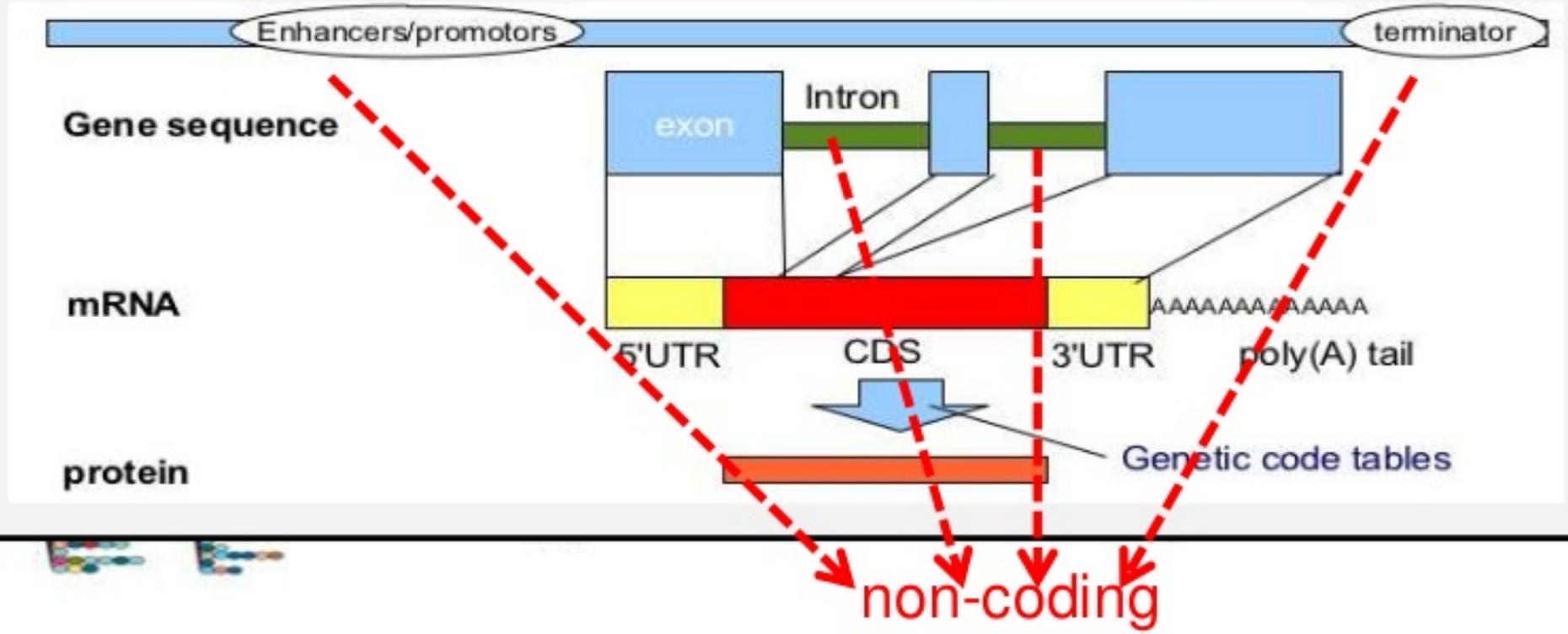


Y-H Hsu

Genome-Wide Association Studies (GWAS) Catalog

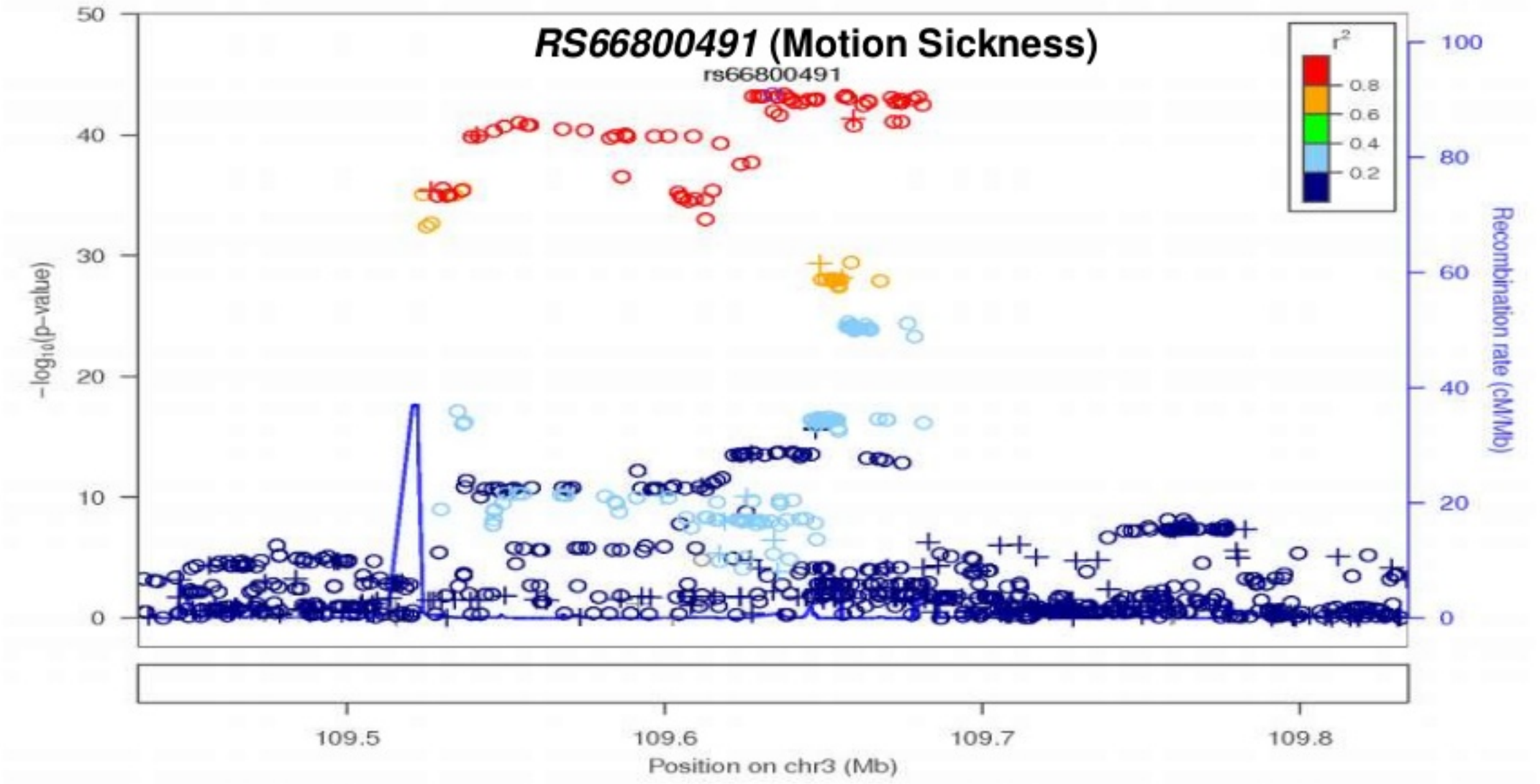
- Identified ~13,000 genetic variants (single nucleotide mutations/ polymorphisms) to be associated with ~2,000 diseases/phenotypes

- **91% of** disease-associated **genetic variants are located in non-protein-coding regions; used to call "junk DNA"**
- **Unknown function, difficult to translate findings into clinical use**



Y-H Hsu

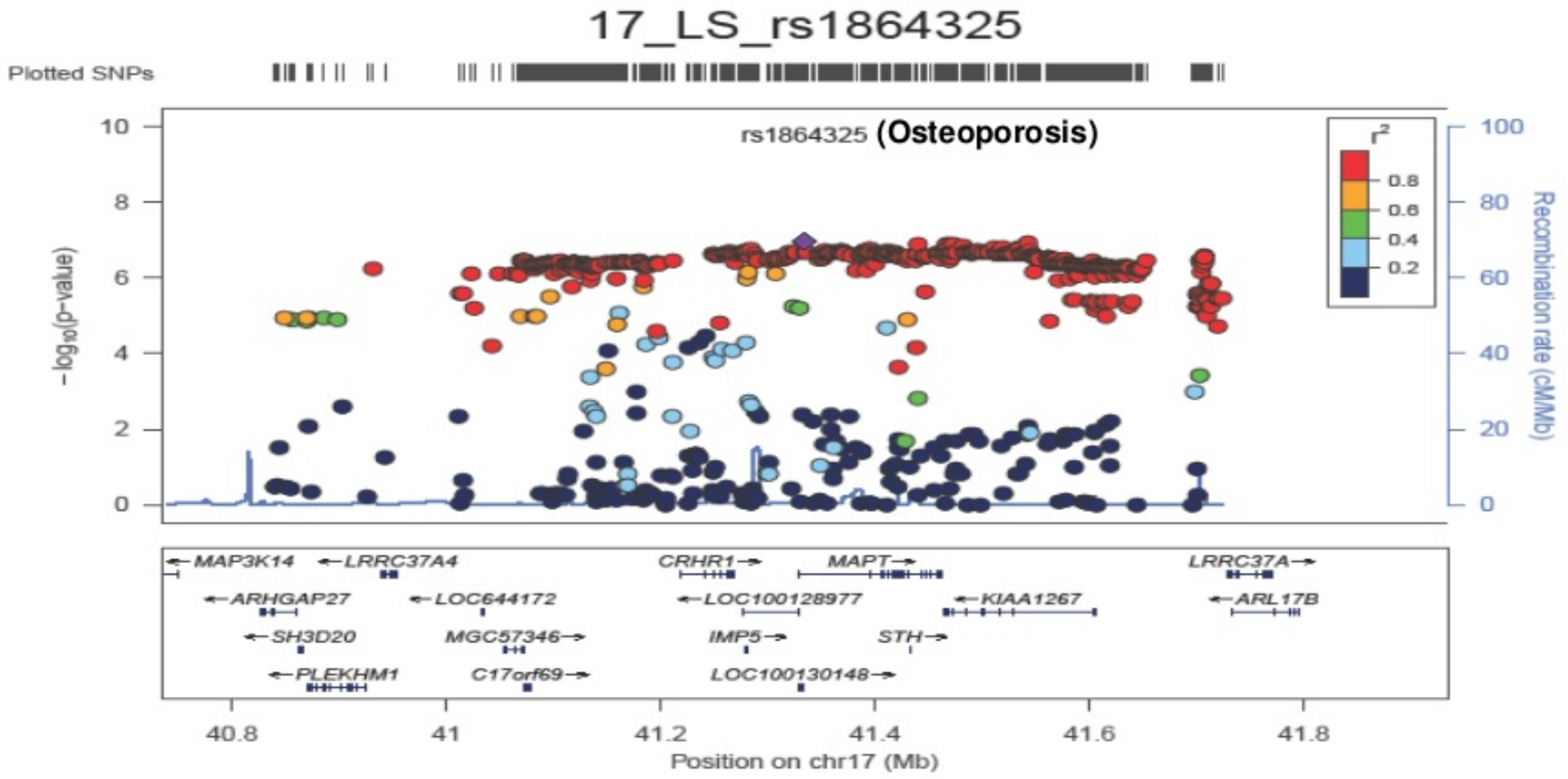
Associated Variants Located in Gene Desert



Genetic Coordination: 1D Physical Location on Linear DNA Sequences

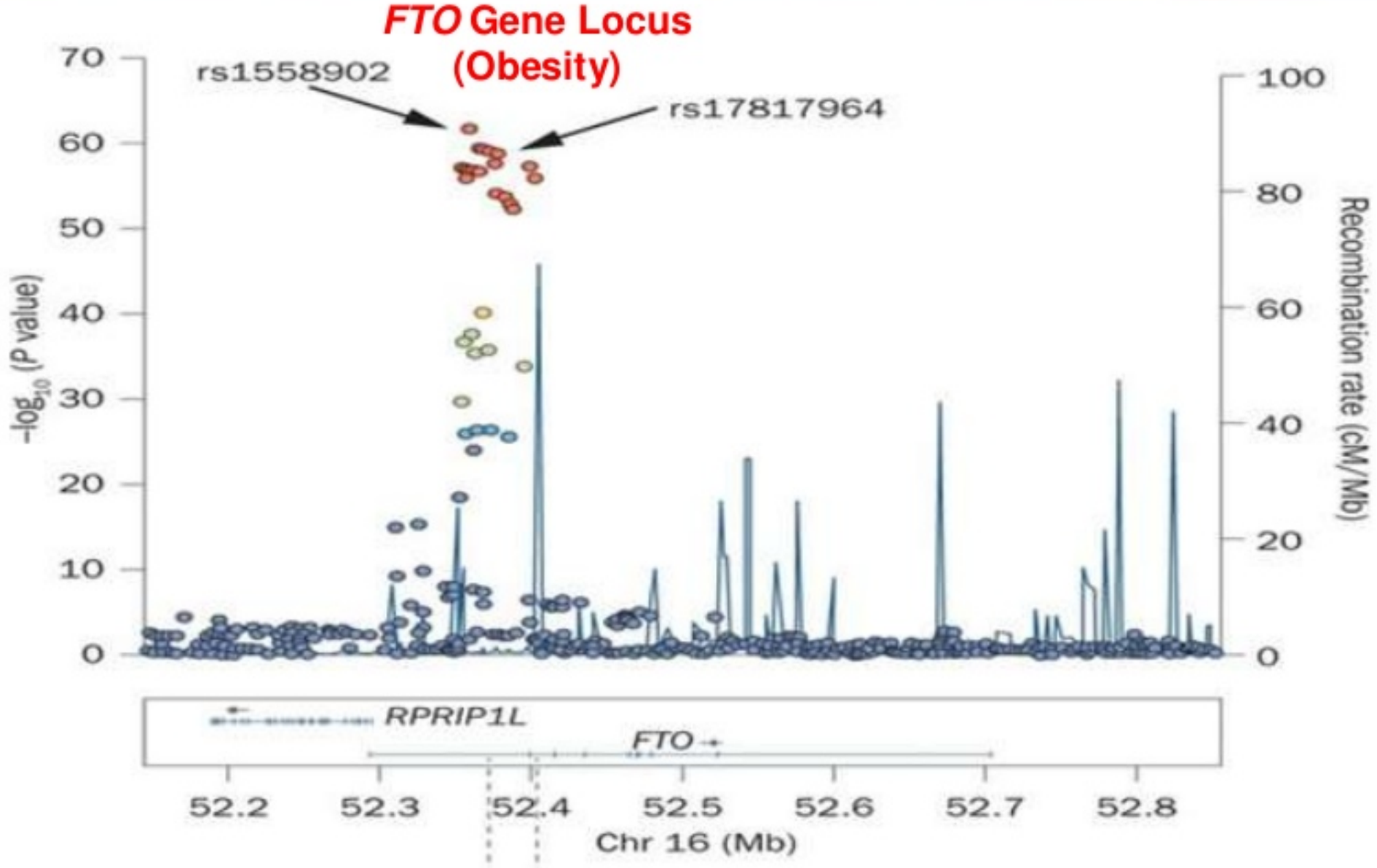
Y-H Hsu

Too Many Genes: Which Gene(s)?



Y-H Hsu

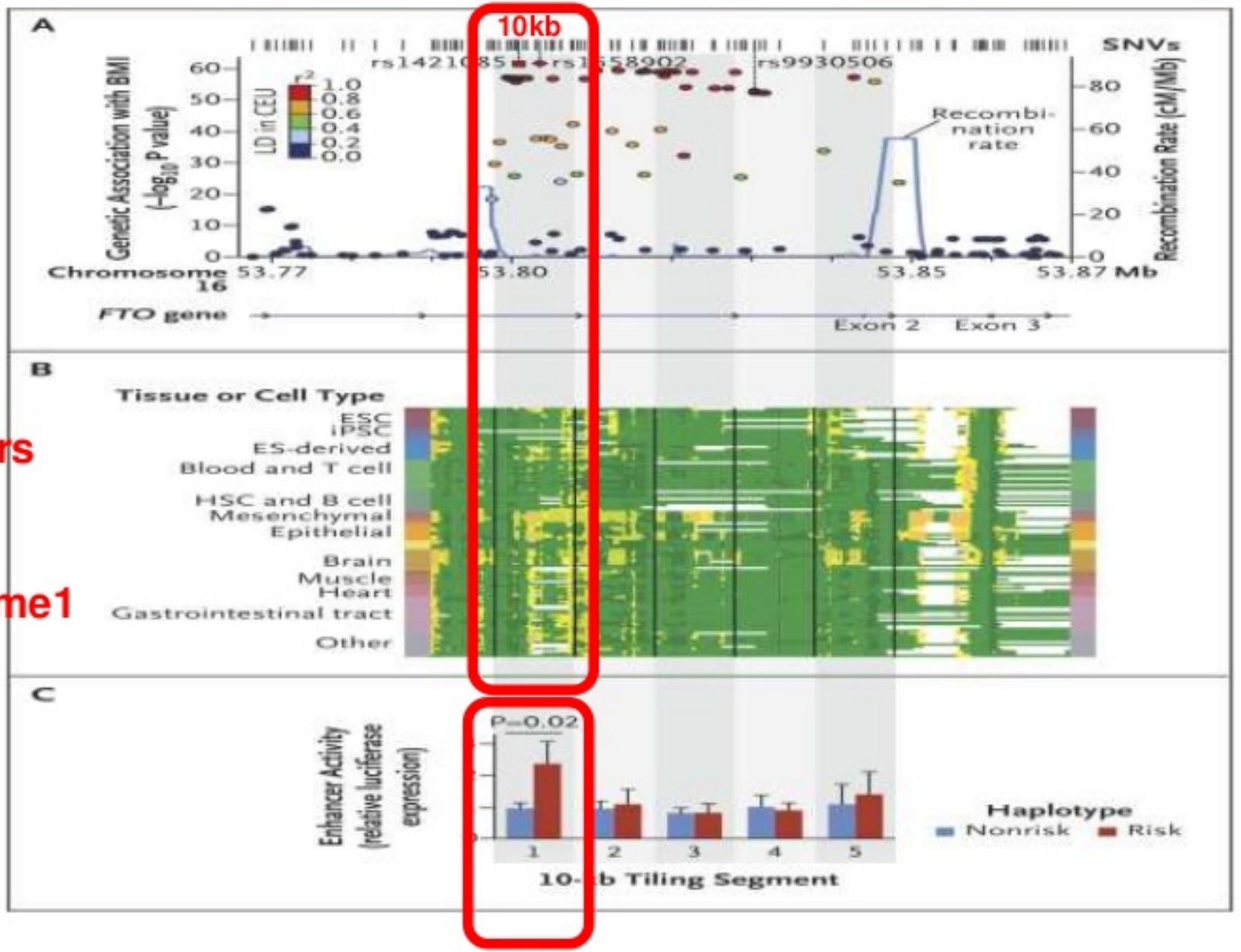
Associated Variants Located in Introns: Looks Promising?



Y-H Hsu

Functional Genomics Approaches

Tissue-Specific Active Enhancers predicted by Histone Marks H3K27ac, H3K4me1 P300

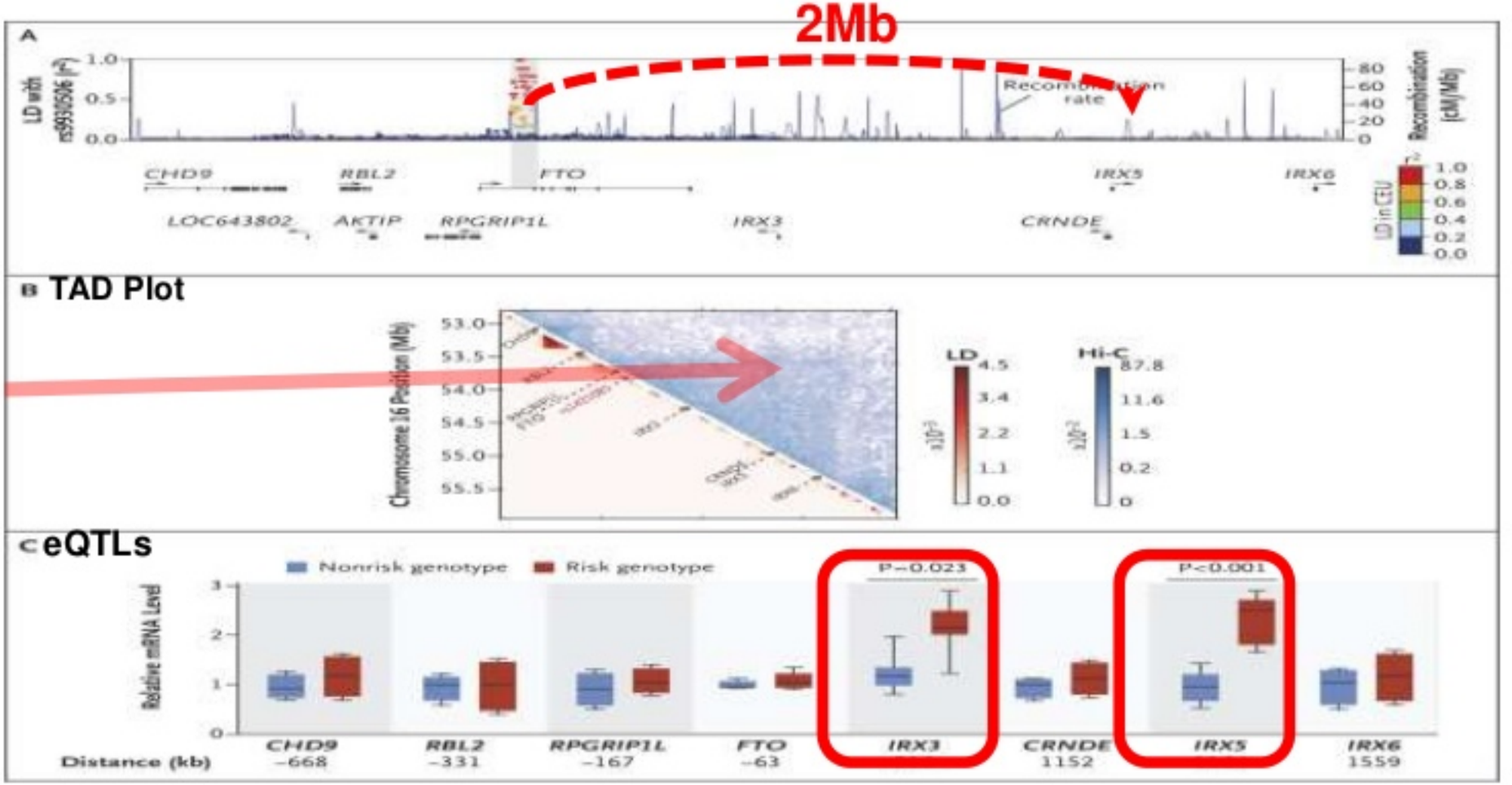


NEJM, 2016

3D Genome Interaction Structure with *IRX5* Gene

- Tissue-Specific Chromatin Confirmation Capture (3C Tech)
- eQTLs (associations between variants and gene expression)
- Allele-specific expression

Intensity of 3D Physical Interaction by Hi-C seq

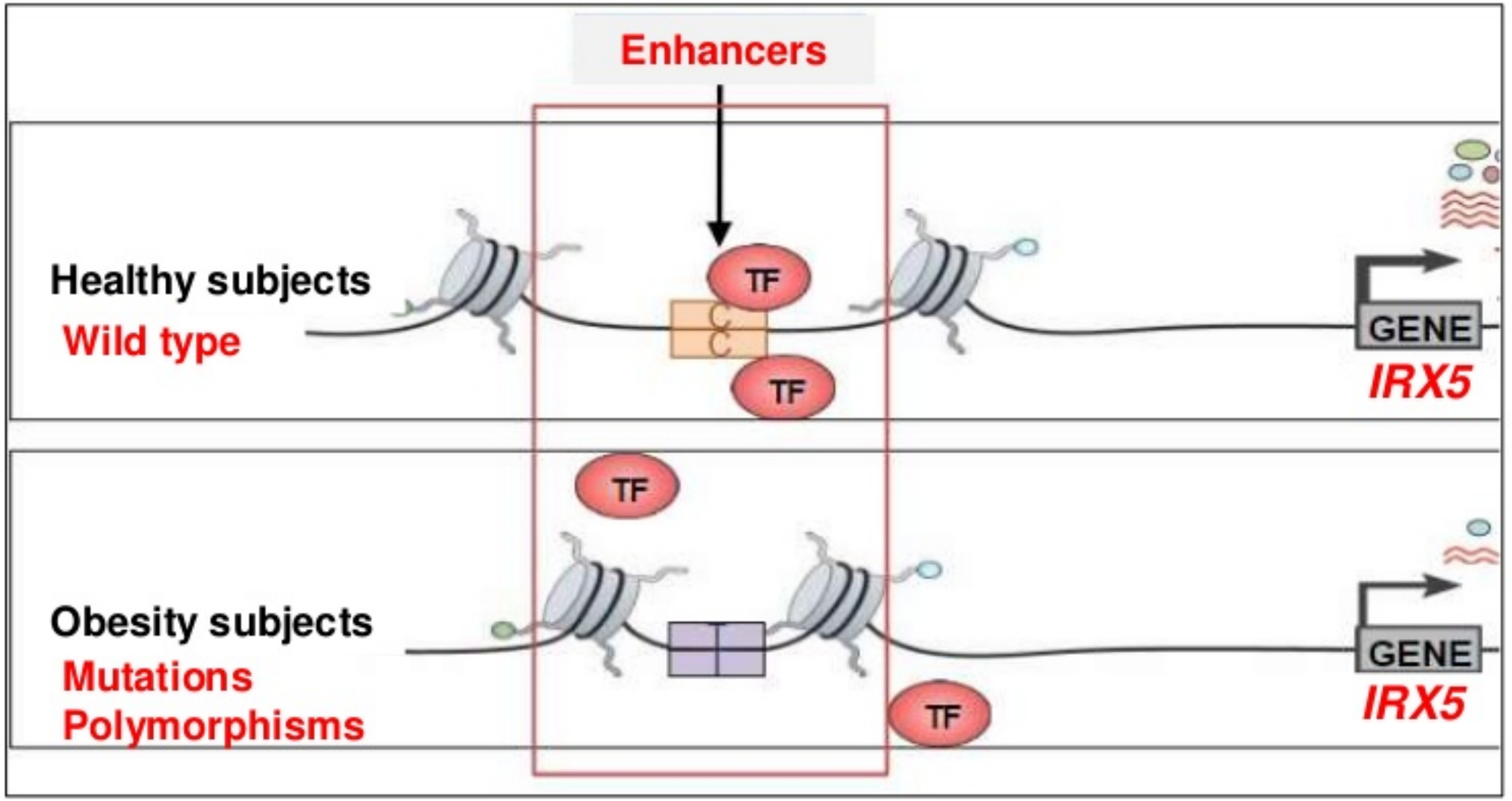


Y-H Hsu

NEJM, 2016

FTO Genetic Variants and *IRX5* Gene Regulation

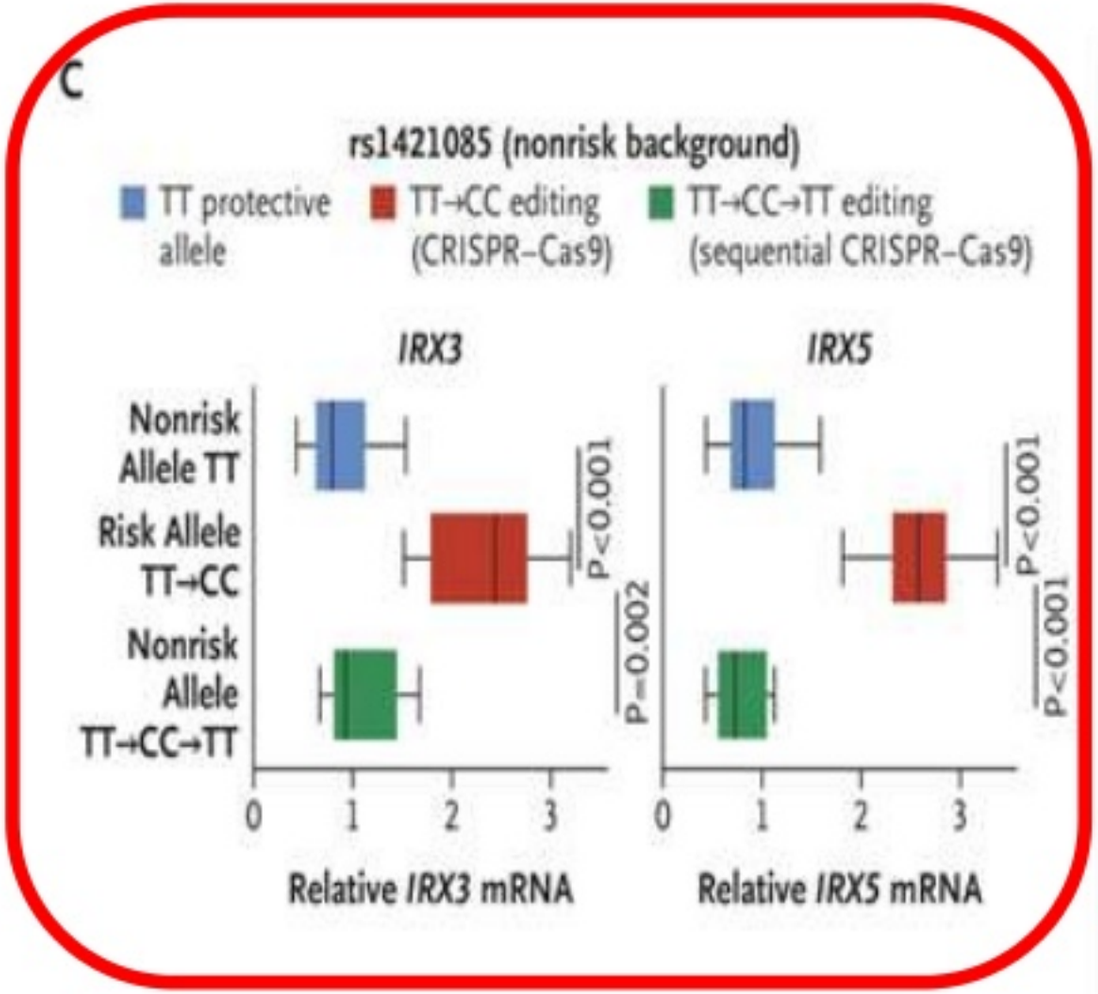
- Obesity associated genetic variants disrupt TF binding and then reduce *IRX5* gene expression

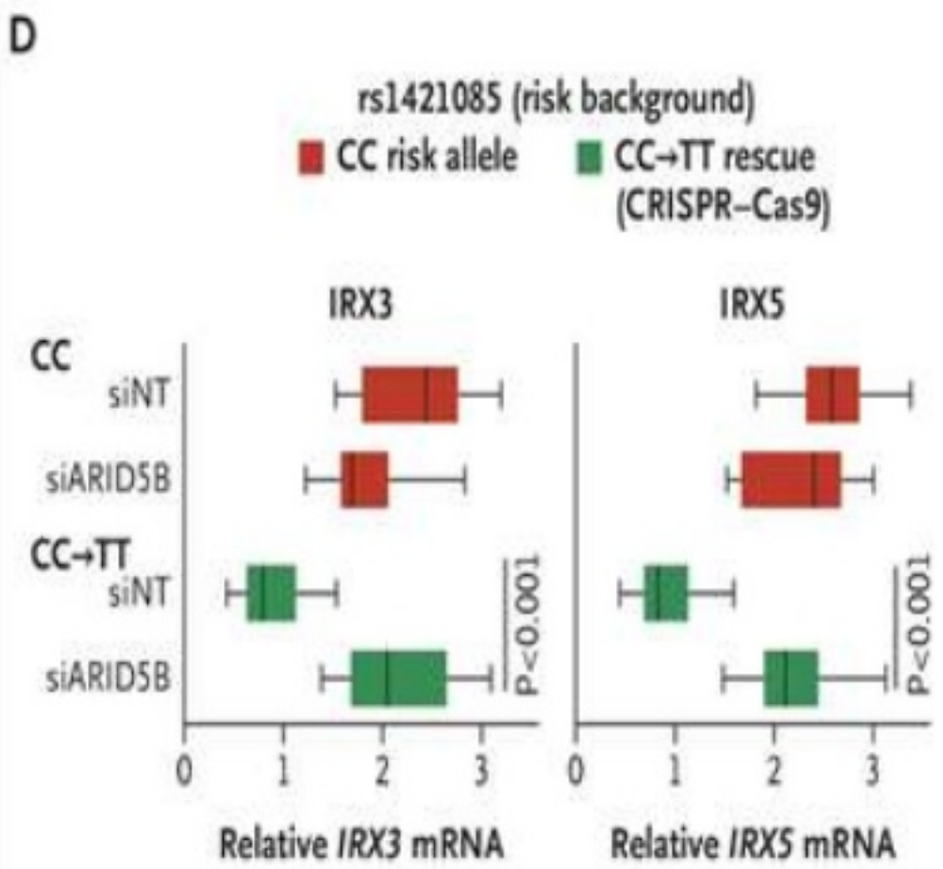


Y-H Hsu

Gene-Editing: Functional Validation

- **Gene Editing by CRISPR/Cas9 in Human adipocytes from subjects carried "risk allele" and subjects carried "protective allele"**
- **The Risk Allele C: Gain-of-function**

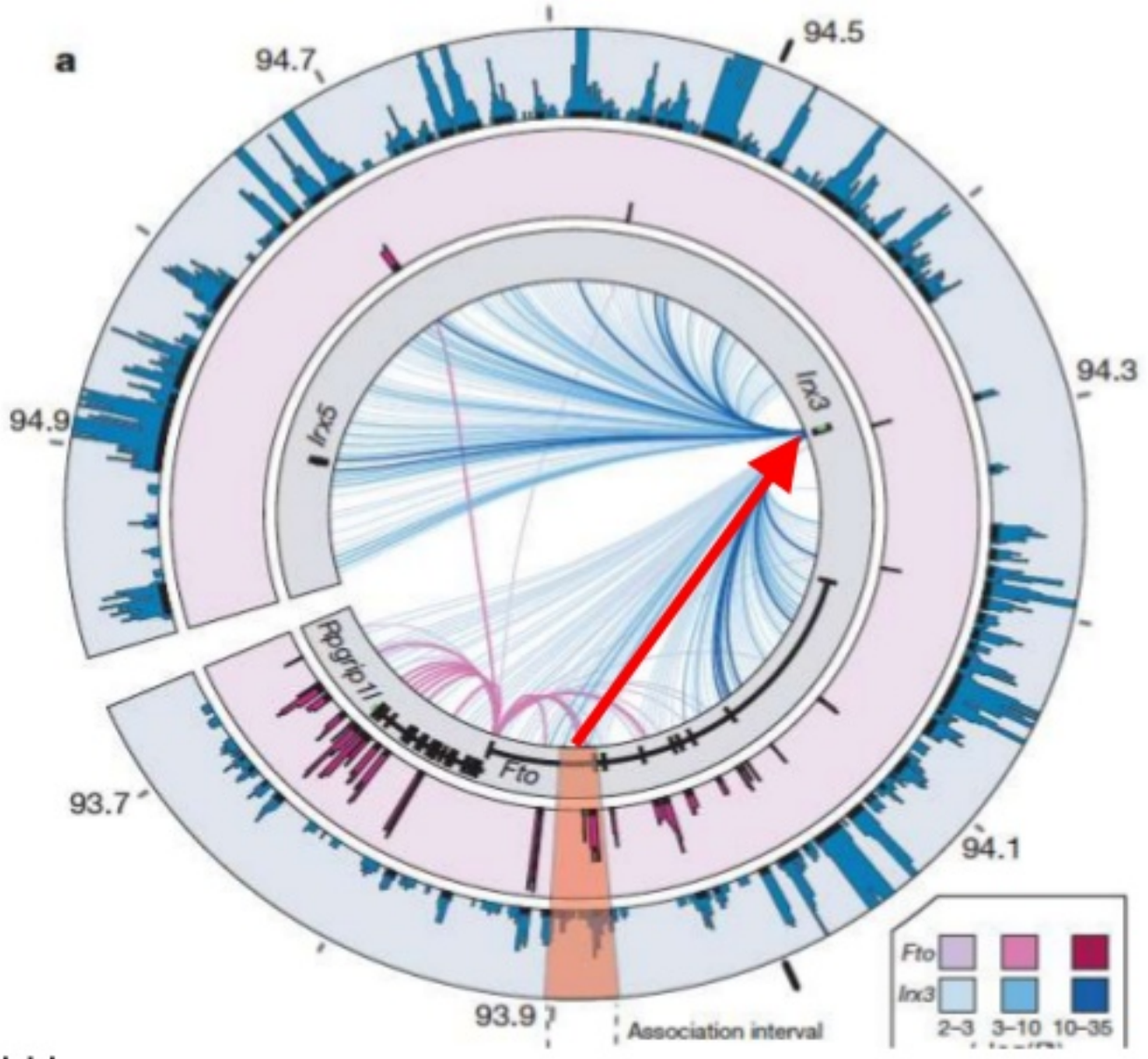




Y-H Hsu

NEJM, 2016

FTO Variants Link to *Irx3* Gene in Brain

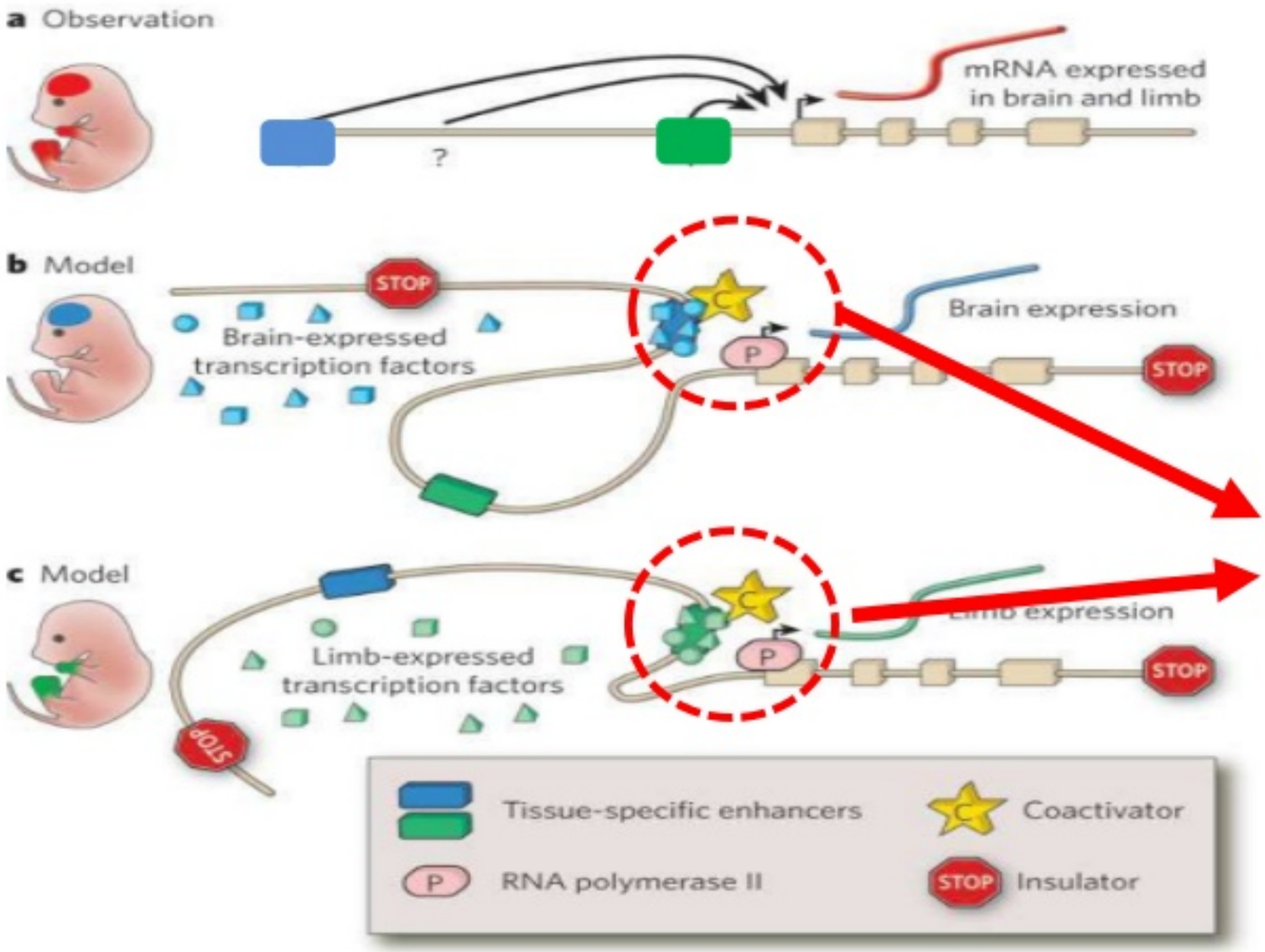


- The obesity associated variants physically interacts with promoter of *Irx3* gene, but not *Fto*, not *Irx5* in mouse brain by 4C-seq
- 4C-seq: Regional Chromatin Confirmation Capture (3C Tech)

Nature, 2014

Y-H Hsu

Gene Regulatory Models



Gene regulatory elements in physical proximity (3D space) with the gene promoters via looping mechanisms

Tissue (Cell)-Specific DNA Loops: Enhancer-Promoter Interactions

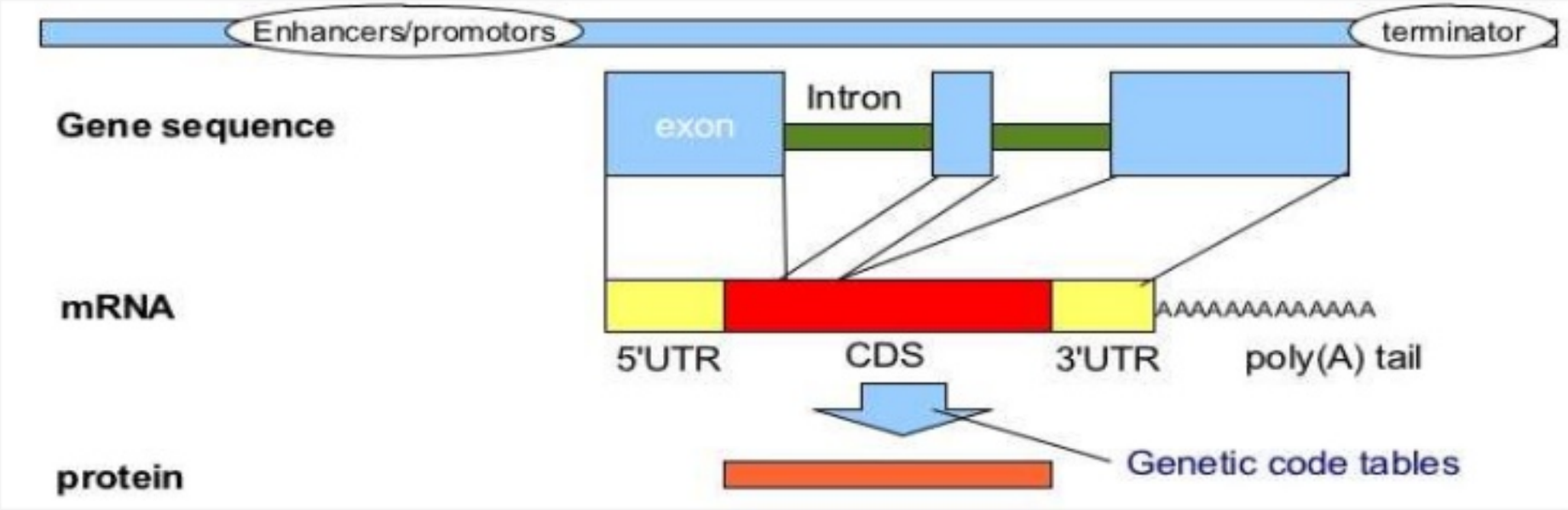
Y-H Hsu

Nature, 2009, 461, 199-205

Genome-Widely Identify/Predict Targeted Genes?

- Identified ~13,000 genetic variants (single nucleotide mutations/ polymorphisms) to be associated with ~2,000 diseases/phenotypes

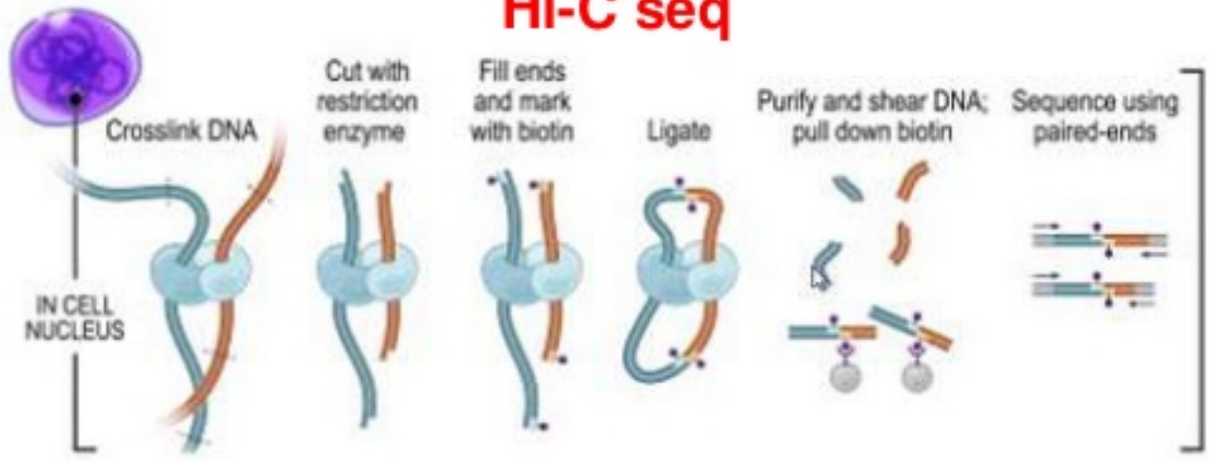
- **91% of disease-associated genetic variants are located in non-protein-coding regions; used to call "junk DNA"**
- **Unknown function, difficult to translate findings into clinical use**
- **May involve in tissue/cell type-specific gene regulation**



Chromosome Conformation Capture To Identify DNA Loops

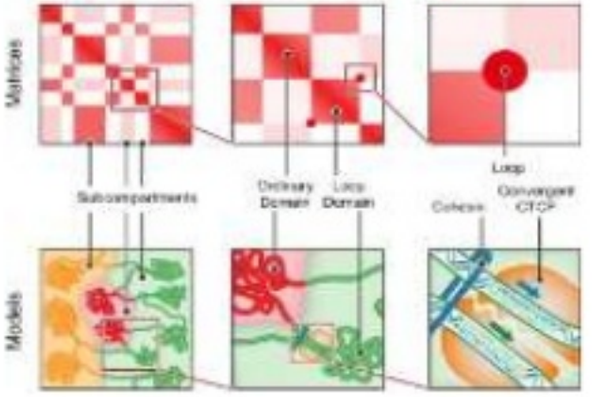
- **3C, 4C, 5C, HiC, capture-HiC, etc to estimate 3D interaction among genome**

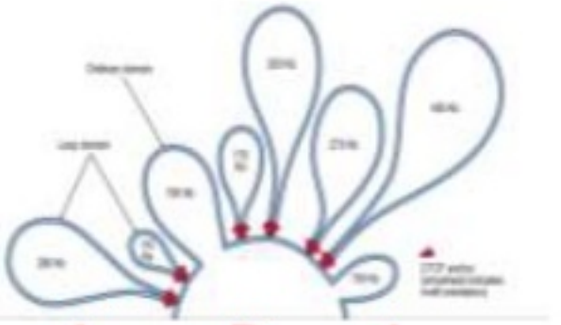
Hi-C seq



Science. 2009.; 326(5950): 289–293
Nat Rev Genet. 2010;11(6):439-46.
Cell. 2014;159(7):1665-80
Nature Genetics 2016; 48, 488–496

Contact Map





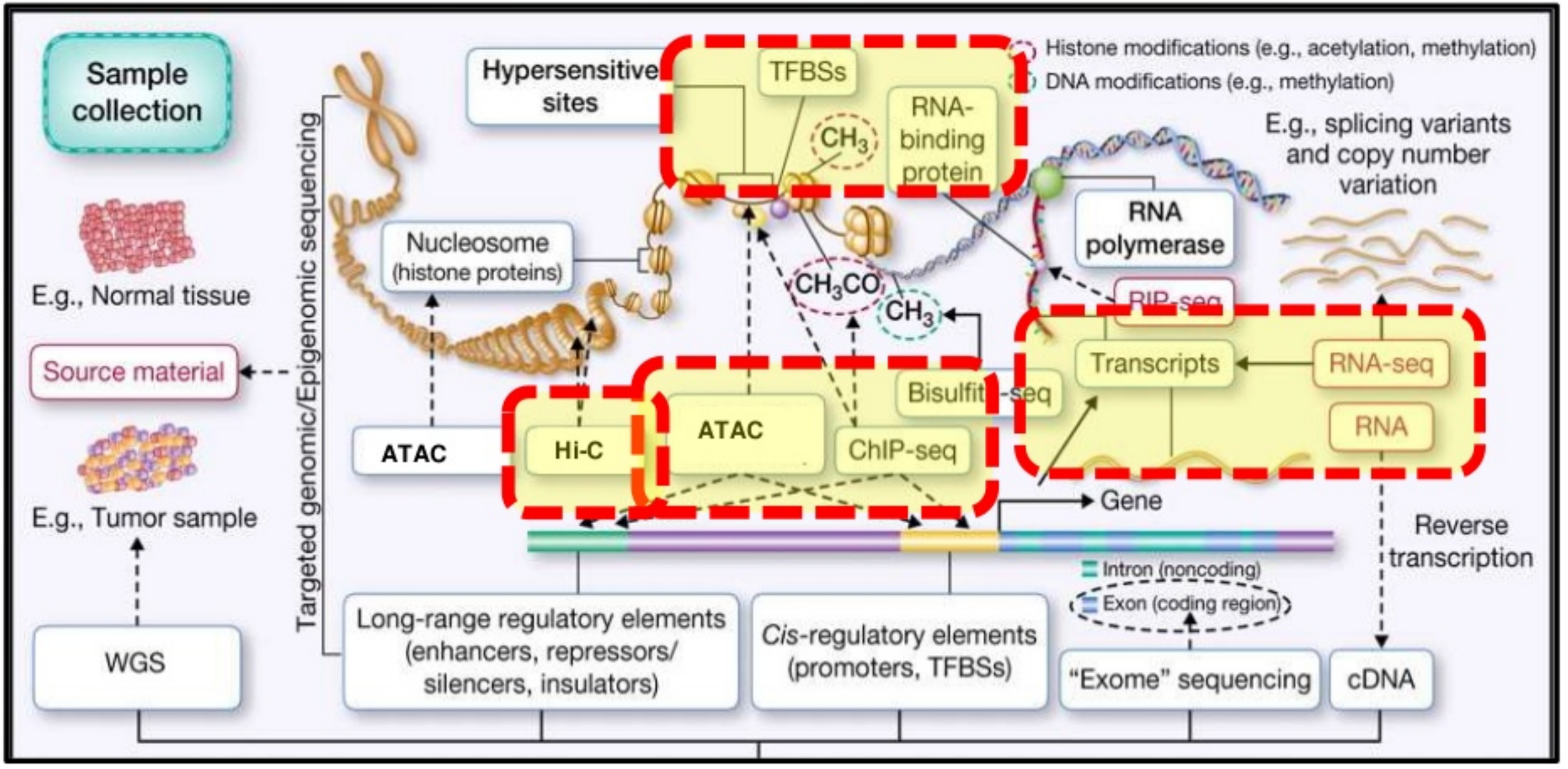
Loop Domains

Enhancer-Promoter
Enhancer-Enhancer
Promoter-Promoter
Physical Interactions
False-Pos (seq error, miss-matched cutting,…)

Prediction

Y-H Hsu

Building Tissue-Specific Gene Regulatory Circuits



Y-H Hsu

Building Gene Regulatory Circuits On Human Heart

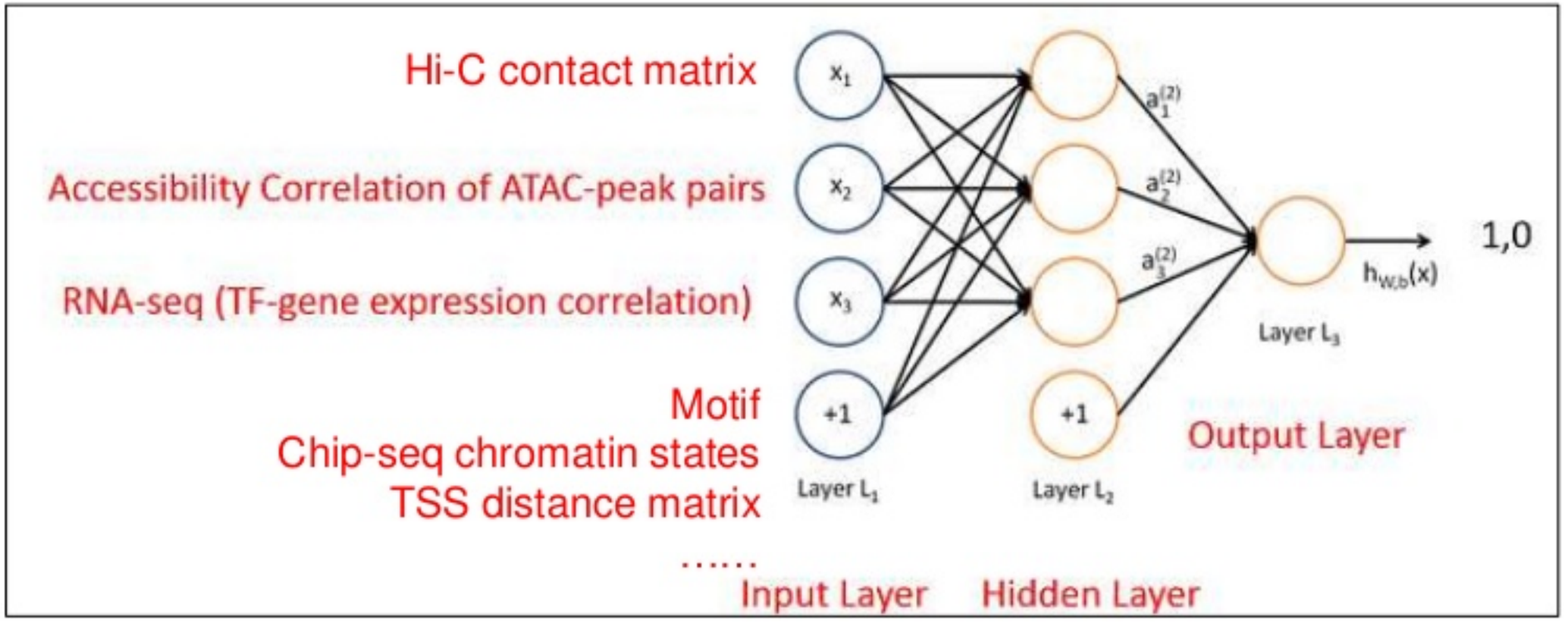
- Omics experiments on normal human primary cardiac fibroblasts and myocytes from atrium and ventricle; HMSC; skeletal muscle cells
- Publicly available (low resolution): Left ventricle, right ventricle and aorta tissues

Experiments	Functions	Notes
ATAC-seq	Active cis-regulatory region	Active TF binding
Hi-C	Chromatin confirmation capture	1.5 to 2 kb resoultion (DpnII, 2 Billions Reads, 2Tb)
ChIP-seq	H3K4me3: Aactive promoter	Predicted chromatin states by HMM
	H3K27ac: Active enhancer/promoter	
	CTCF: Insulator	
	Cohesin: Insulator-RAD21	
	Cohesin: Insulator-SMC3	
	H3K27me3: Polycomb repressed/bivalent promoter/enhancer	
	H3K9me3: Heterochromatin	
	H3K36me3: Transcribed region	
RNA-seq	mRNA (active and)	Isoforms; coexpression with TF
	microRNA and small RNA	Enhancer RNA

Y-H Hsu

Model Gene Regulation with Deep Neural Network (DNN)

- **DNN implemented in the TensorFlow to predict enhancer-promoter gene pairs**



- **Training sets** (VISTA: enhancer elements are in 100kb of genes):
  1,564 Enhancer-promoter gene pairs (the positive set) functionally validated to have regulatory relationships in mouse models
  1,207 EP pairs without regulatory relationships (the negative set)

Acknowledgements

Erasmus MC
GEFOS
GENOMOS
SOF
UK 10K
McGill
Mr. OS
TwinsUK
CIHR IRSC
Canadian Institutes of Health Research
Institute of Genetics
wellcome trust
The Jackson Laboratory
Leading the search for tomorrow's cures
BROAD INSTITUTE
Arima Genomics
University of BRISTOL
Avon Longitudinal Study of Parents and Children
MGH 1811
WOMEN'S HEALTH INITIATIVE
HARVARD MEDICAL SCHOOL
THE UNIVERSITY OF QUEENSLAND AUSTRALIA
DIAMANTINA INSTITUTE
KING'S College LONDON
wellcome trust sanger institute
databricks
NIH National Institutes of Health
PEOPLE SCIENCE NHLBI HEALTH

Yi-Hsiang Hsu, MD, ScD

yihsianghsu@hsl.harvard.edu

yihsiang@broadinstitute.org