6.874, 6.802, 20.390, 20.490, HST.506
Computational Systems Biology
Deep Learning in the Life Sciences

Lecture 12: Human genetics, GWAS, and disease circuitry

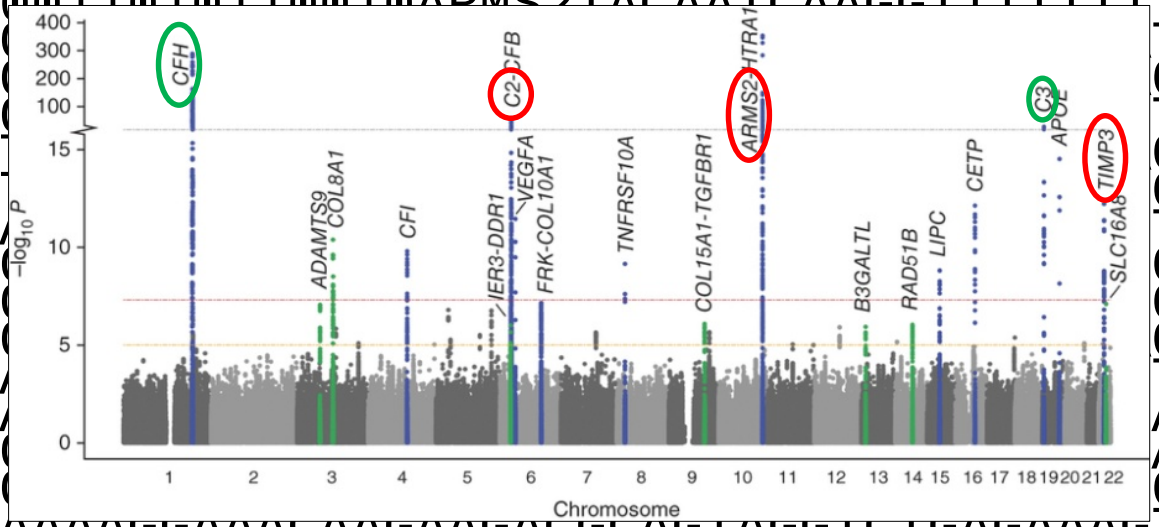
Prof. Manolis Kellis

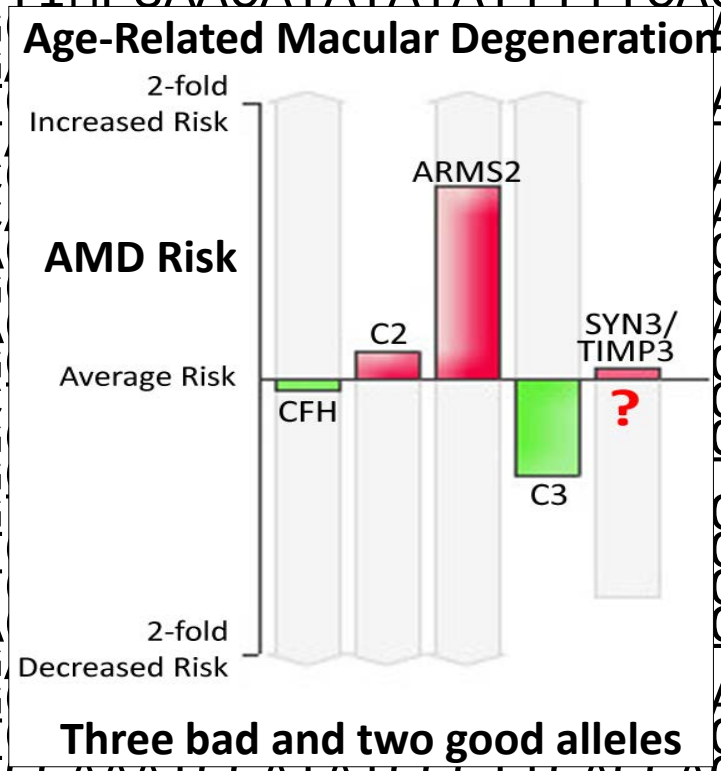
Guest lecture: Jian Zhu (DeepSEA)

Massachusetts Institute of Technology

http://mit6874.github.io

Slides credit: Mark Daly, David Gifford, et al





ARMS2

TIMP3/SYN3

C2

Today: Deep Learning for Human Genetics and Disease

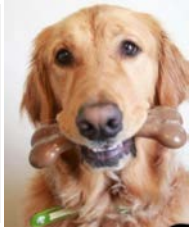
1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting: Common/rare alleles, Linkage vs. GWAS
3. LD, Haplotypes, Co-inheritance, and the challenge of fine-mapping
4. From locus to mechanism - Case study: FTO and Obesity
5. Epigenomics-GWAS integration: ENCODE, Roadmap, EpiMap
6. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA
7. Interpreting non-coding variation: DeepSEA (Jian Zhu guest lecture)

1. Intro to Human Genetics

Inheritance, Mendel, Fisher, SNPs, STRs, alleles

Inheritance and Genetics: Ancient foreshadowings

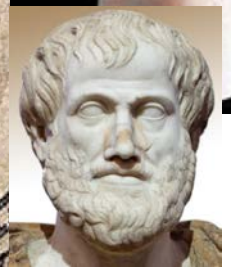
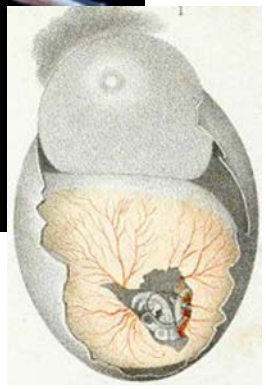


9000BC: Selective breeding of animals/plants

Inheritance: Eye/hair color long understood

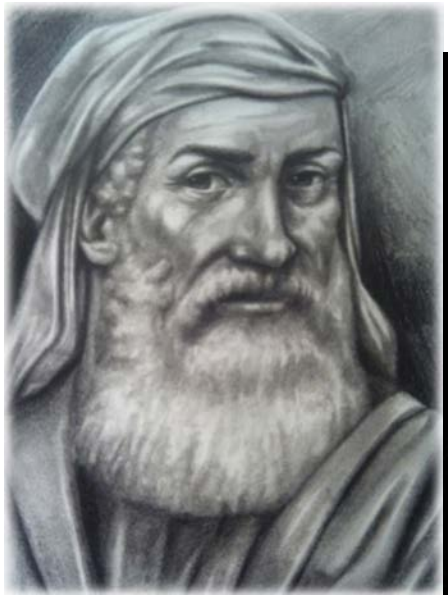


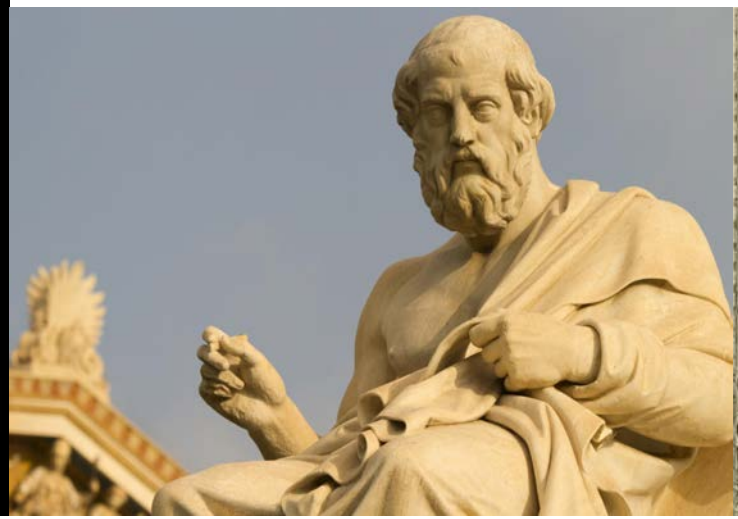



550BC: Anaximander: first human was born from non-human relative, fish origin of land animals

300BC: Aristotle: species taxonomy classification

Seedlings: Theophrastus,Hippocrates,Aeschylus

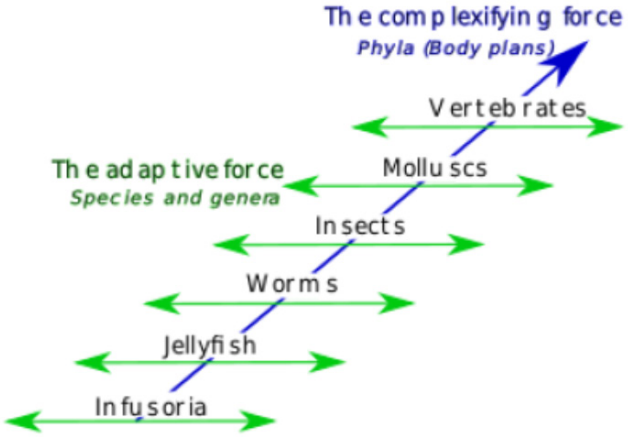


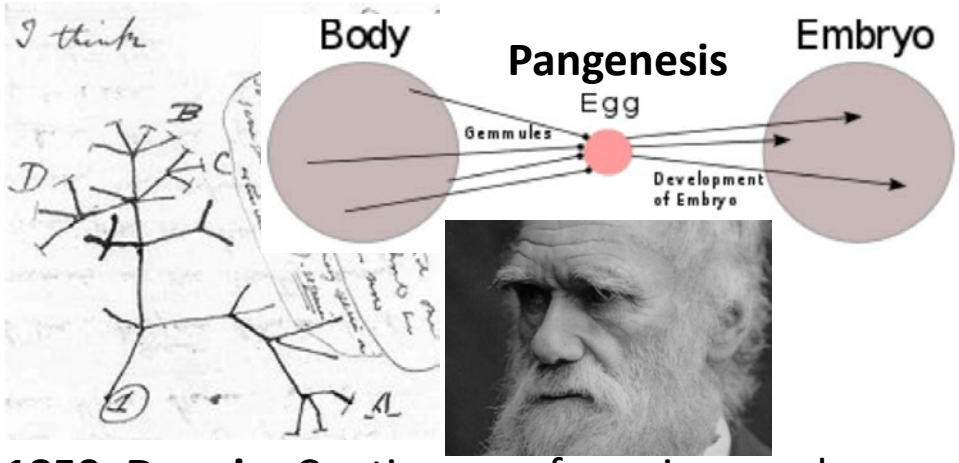

450BC: Empedocles: Random mixing of traits, natural variation, successful ones survive, giving semblance of 'purpose'

300BC: Epicurus: purely naturalistic generation of diversity, no supernatural intervention. (Contrast: Plato, Stoics, Religion, Christianity)

19th Century: Lamarck, Darwin, Mendel, Biometrics

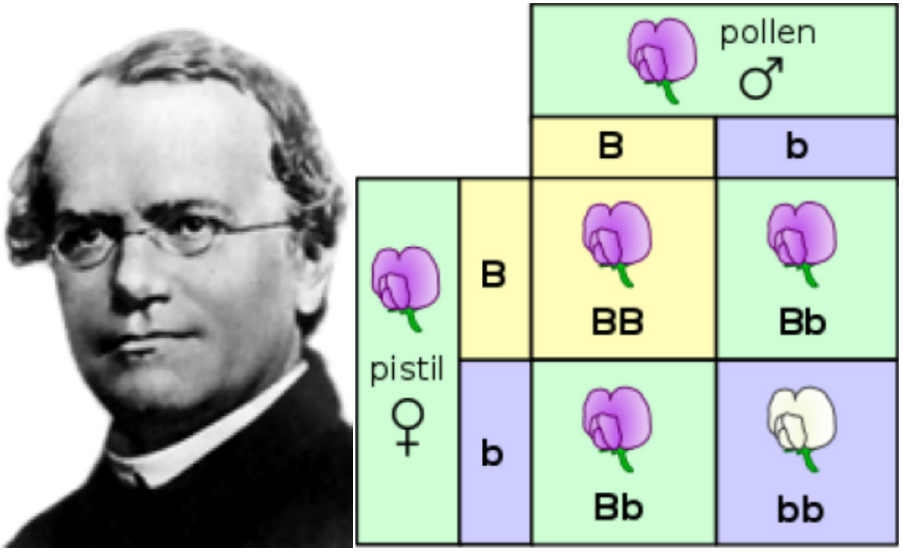


1809: Lamarck: Transmutation, adaptation spontaneous generation of simple life-forms, innate life force drives increased complexity

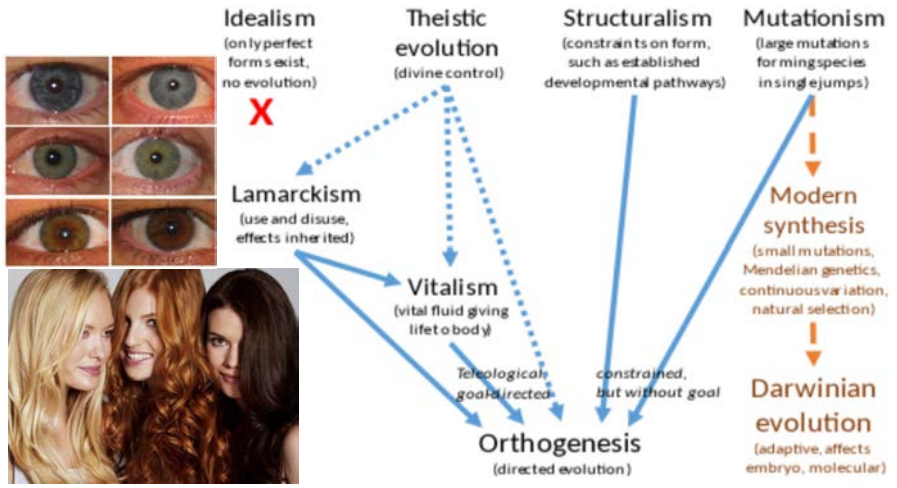


1859: Darwin: Continuum of species, random mutation→diversity, natural selection→fitness.

But: Blending inheritance, gemmules, Lamarckism

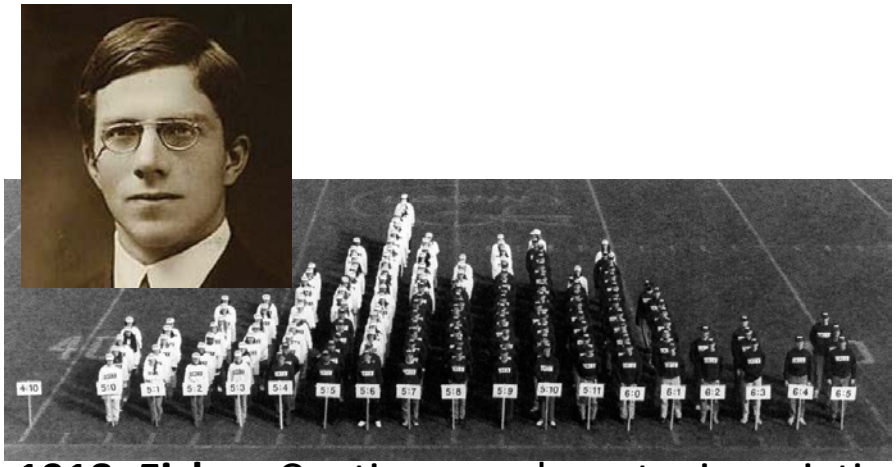


1866: Mendel: Particulate inheritance, no blend. Discrete units=genes. Dominant/recessive alleles Independent assortment. **Digital inheritance.**

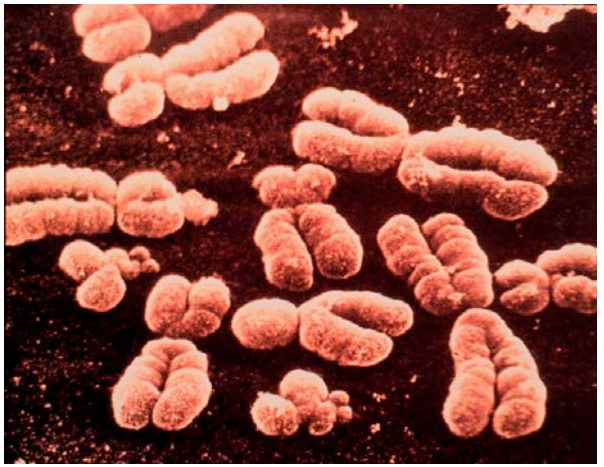


Biometrics: continuous phenotype variation.
Others: Saltationism, orthogenesis, vitalism, neo-Lamarckism, theistic evolution…

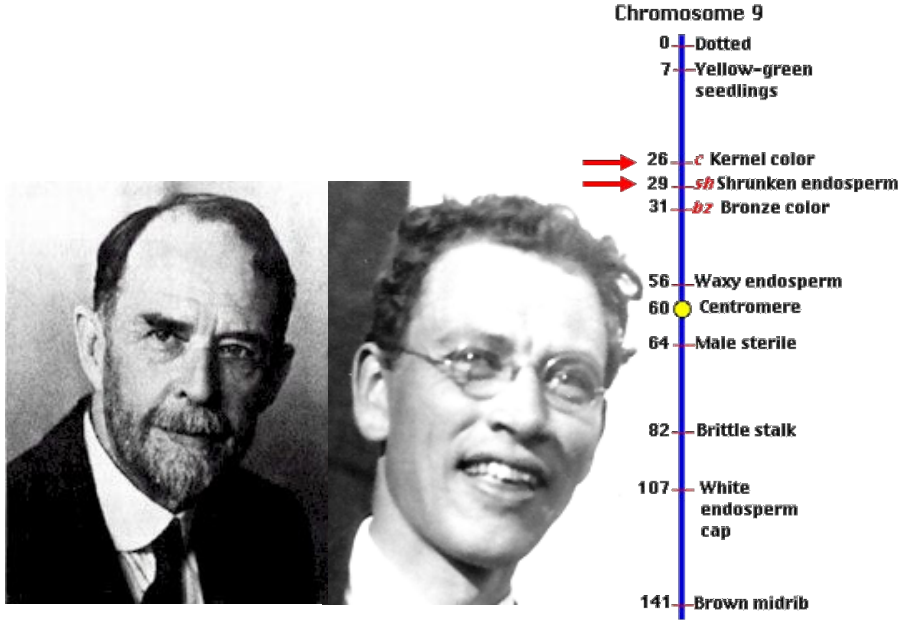
20th Century: Synthesis, DNA, polygenic inheritance



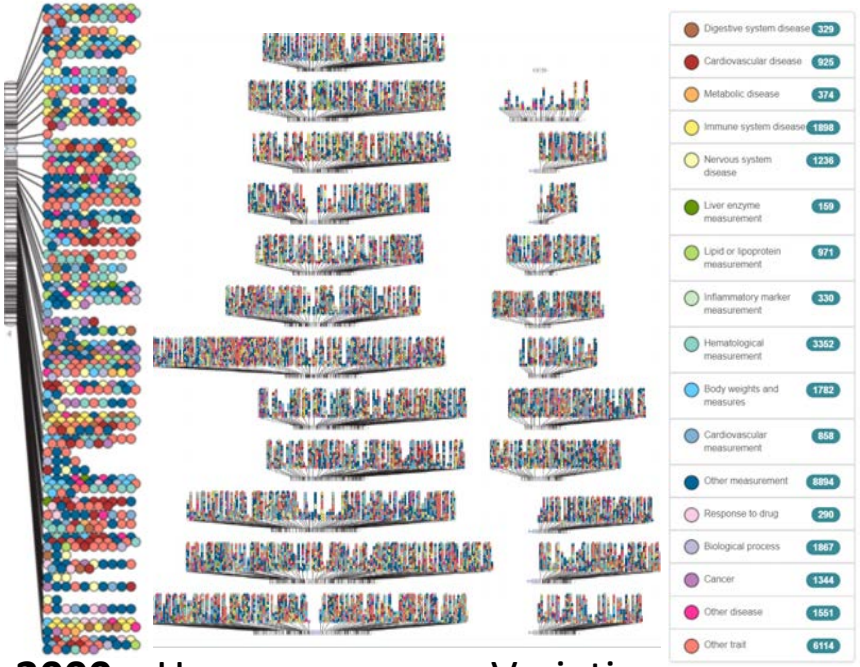
1918: Fisher. Continuous phenotypic variation explained simply by multiple Mendelian loci



1902: Chromosomes, DNA, genetic material
1953: Structure of DNA, basis for inheritance



1913: Linkage/mapping, Morgan, Sturtevant
1980s: Mendelian Trait genes mapped



2000s: Human genome. Variation maps. Haplotypes. GWAS. Common/rare variants.

Types of genetic variation

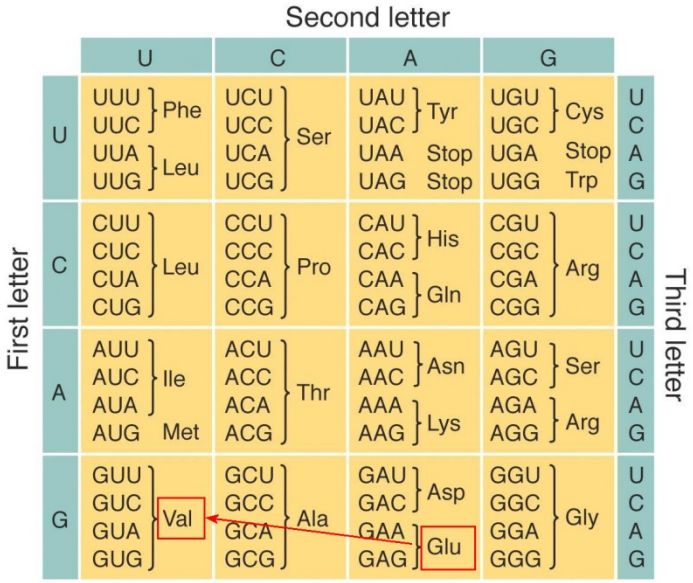
- **99% of DNA is shared** between two individuals
- Variation in the remainder explains all our **predisposition** differences
- **Remaining** phenotypic variation: environmental/stochastic differences

Name	**Example**	**Frequency in one genome**
Single nucleotide polymorphisms (**SNPs**)	GAGGAGAACG[C/G]AACTCCGCCG	1 per 1,000 bp
Insertions/deletions (**indels**)	CACTATTC[C/CTATGG]TGTCTAA	1 per 10,000 bp
Short tandem repeats (**STRs**)	ACGGCAGTCGTCGTCGTCACCGTAT	1 per 10,000 bp
Structural variants (**SVs**) / Copy Number Variants (**CNVs**)	Large (median 5,000 bp) deletions, duplications, inversions	1 per 1,000,000 bp

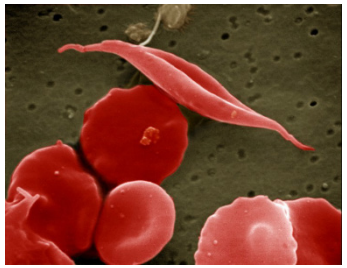
Single-nucleotide polymorphisms (SNPs)

CATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTG

CATGGTGCATCTGACTCCTG<u>T</u>GGAGAAGTCTGCCGTTACTG



glutamic acid > valine



Sickle Cell Anemia

rs189107123

GAGGAGAACG[C/G]AACTCCGCCG

- Many modern analyses (GWAS, eQTL) focus on SNPs/indels
- Often have only two **alleles** (states)
- Identified as reference SNP clusters (**rsid**)
- Submitted sequences containing a variant are clustered to build a database (**dbSNP**)
- To date, >100 M known variants in dbSNP

Short tandem repeats (STRs) + Insertions/deletions (indels)

- Variable number tandem repeats

9 **TCACAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT**

10 **TCACAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT**

12 **TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT**

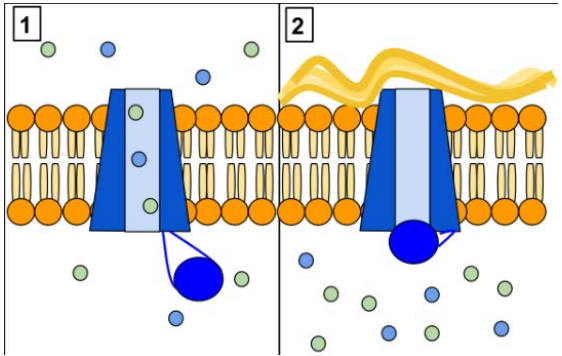
> 30 Huntington's Disease

Abnormal protein, damages neurons, brain cell death, mood, coordination, speaking, dementia, etc

- Insertion/Deletions

Cystic fibrosis transmembrane conductance regulator (CFTR) -> Lung infections, cysts, fibrosis

CATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATA
CATTAAAGAAAATATCATTGGTGTTTCCTATGATGAATA



CFTR Sequence:

Nucleotide	ATC	ATC	TTT	GGT	GTT
Amino Acid	Ile 506	Ile	Phe 508	Gly	Val 510

Deleted in ΔF508

ΔF508 CFTR Sequence:

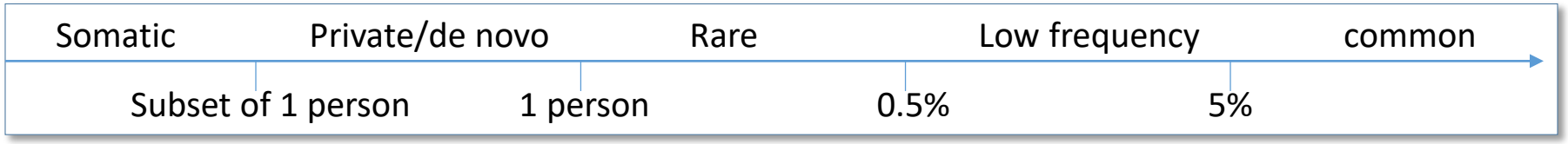
Nucleotide	ATC	ATT	GGT	GTT
Amino Acid	Ile 506	Ile	Gly	Val

SNP alleles: ref/alt; maj/min; risk/prot; anc/der

Referring to the two alleles:

- **Reference/alternate:** Matching the human reference sequence (arbitrary, some random person was sequenced, has rare alleles too)
- **Major/minor:** Being more frequent in the population (population-specific)
- **Ancestral/derived:** Matching the most recent common ancestor between human and chimpanzee (but sometimes chimp doesn't match)
- **Risk/non-risk:** Based on their disease association (but environment specific, e.g. Sickle-cell vs. Malaria)

Classifying variants by minor allele frequency:



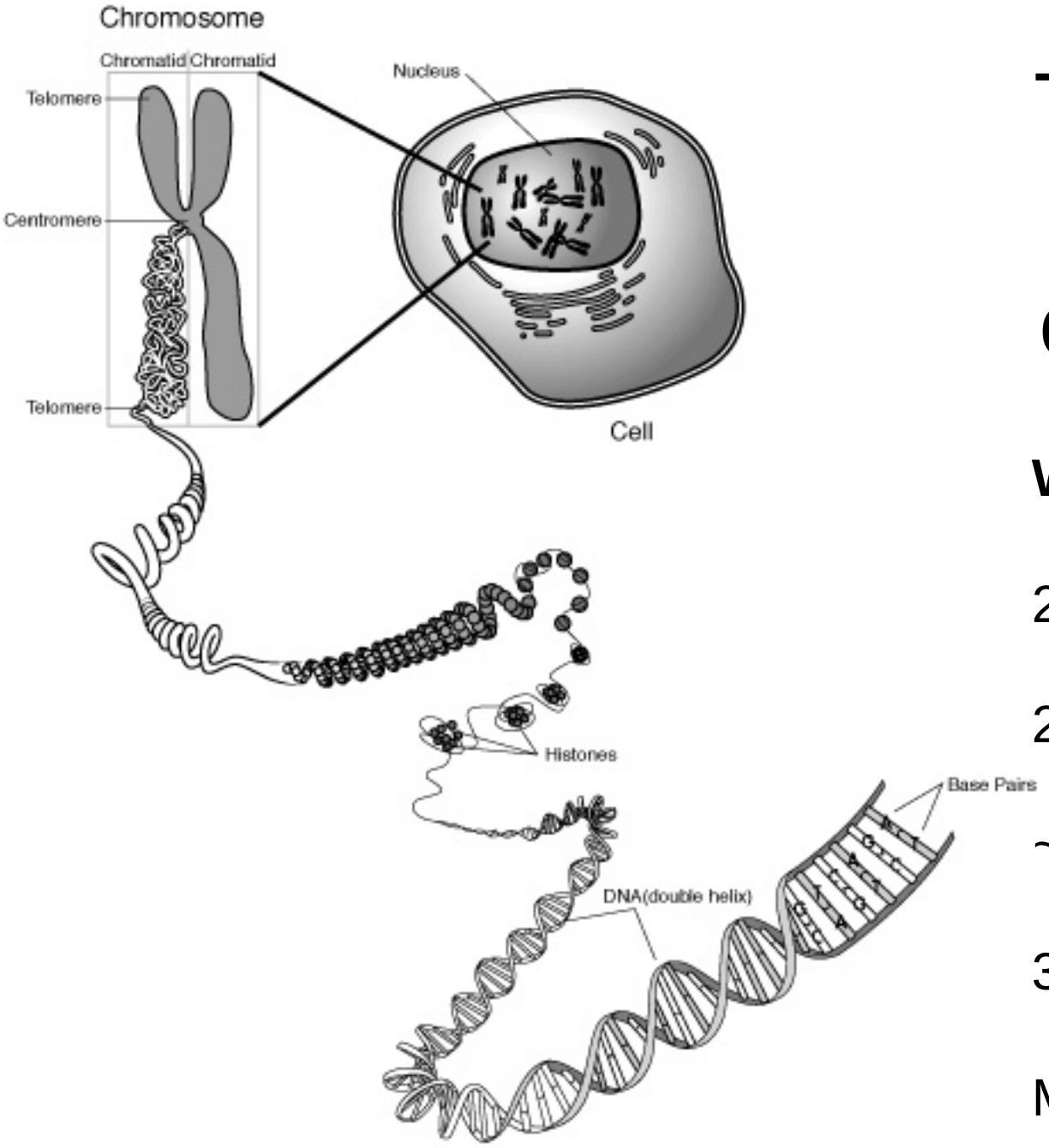
Example: rs189107123

GAGGAGAACG[C/G]AACTCCGCCG

Reference allele: C

Minor allele: G (frequency 0.03 in Europeans)

Ancestral allele: unknown (**why?**)



The scope of the challenge:

Within each cell:

2 copies of the genome

23 chromosomes

~20,000 genes

3.2B letters of DNA

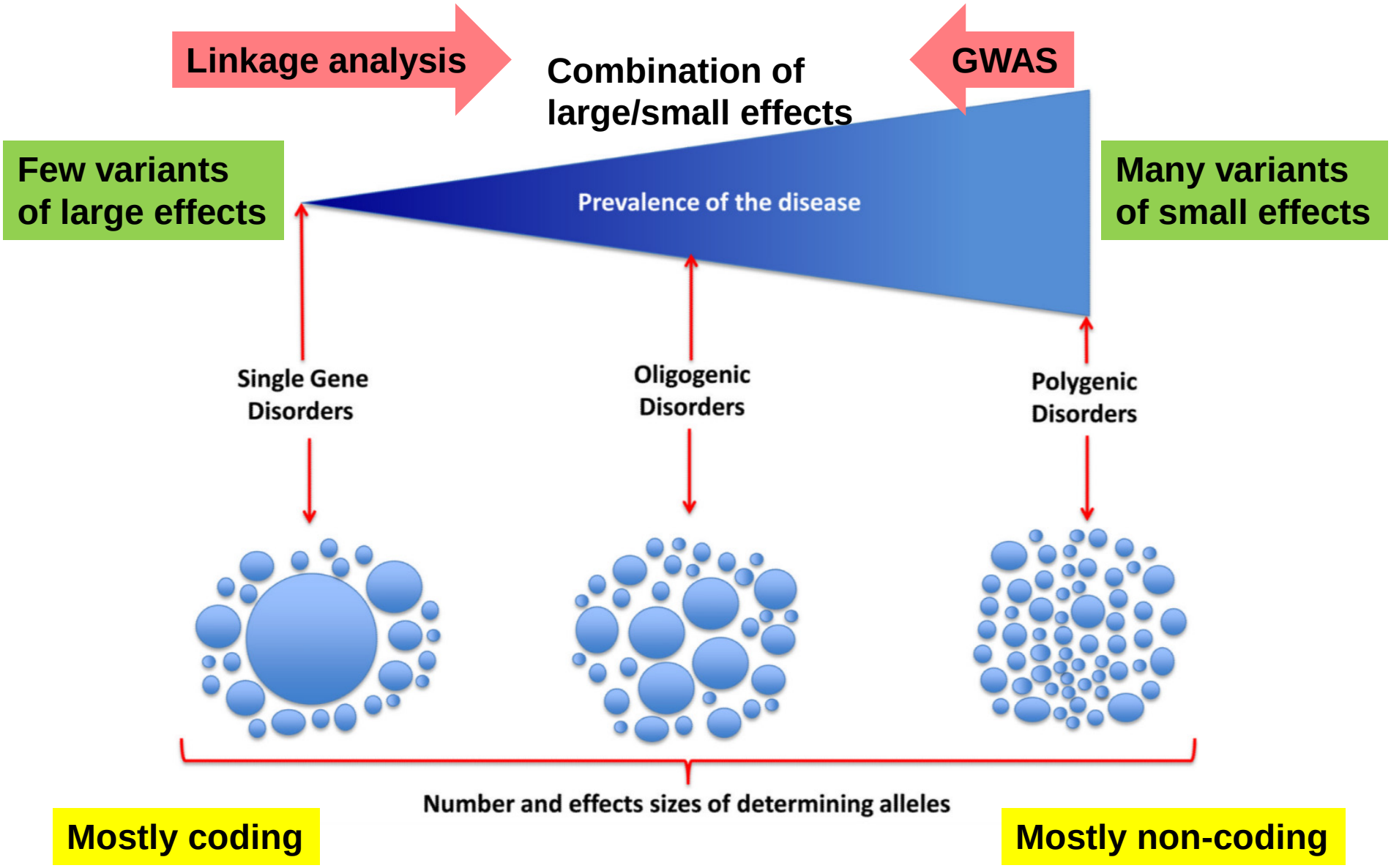
Millions of polymorphic sites

Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting: Common/rare alleles, Linkage vs. GWAS
3. LD, Haplotypes, Co-inheritance, and the challenge of fine-mapping
4. From locus to mechanism - Case study: FTO and Obesity
5. Epigenomics-GWAS integration: ENCODE, Roadmap, EpiMap
6. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA
7. Interpreting non-coding variation: DeepSEA (Jian Zhu guest lecture)

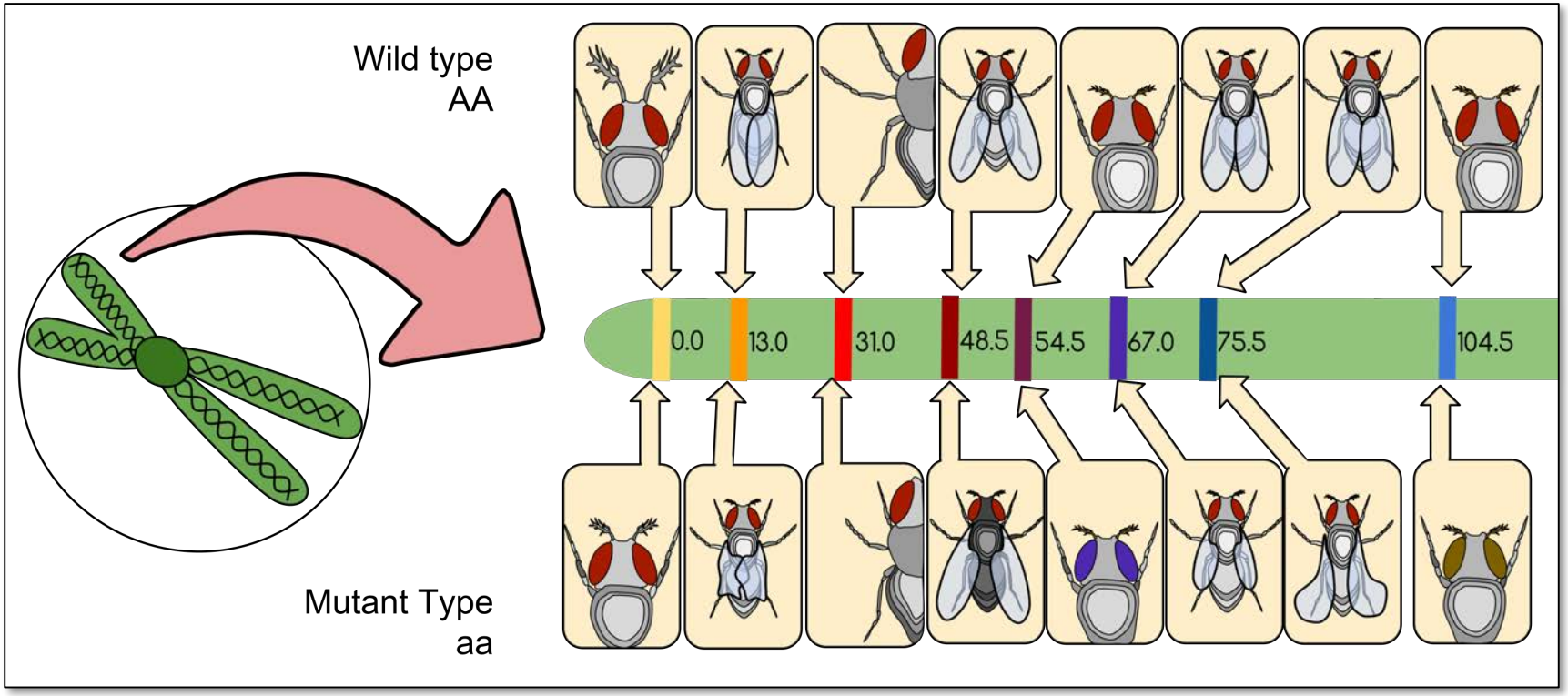
2. ‘Disease gene’ hunting (locus, really):
Common/rare alleles, Linkage vs. GWAS

Monogenic vs. oligogenic vs. polygenic disorders



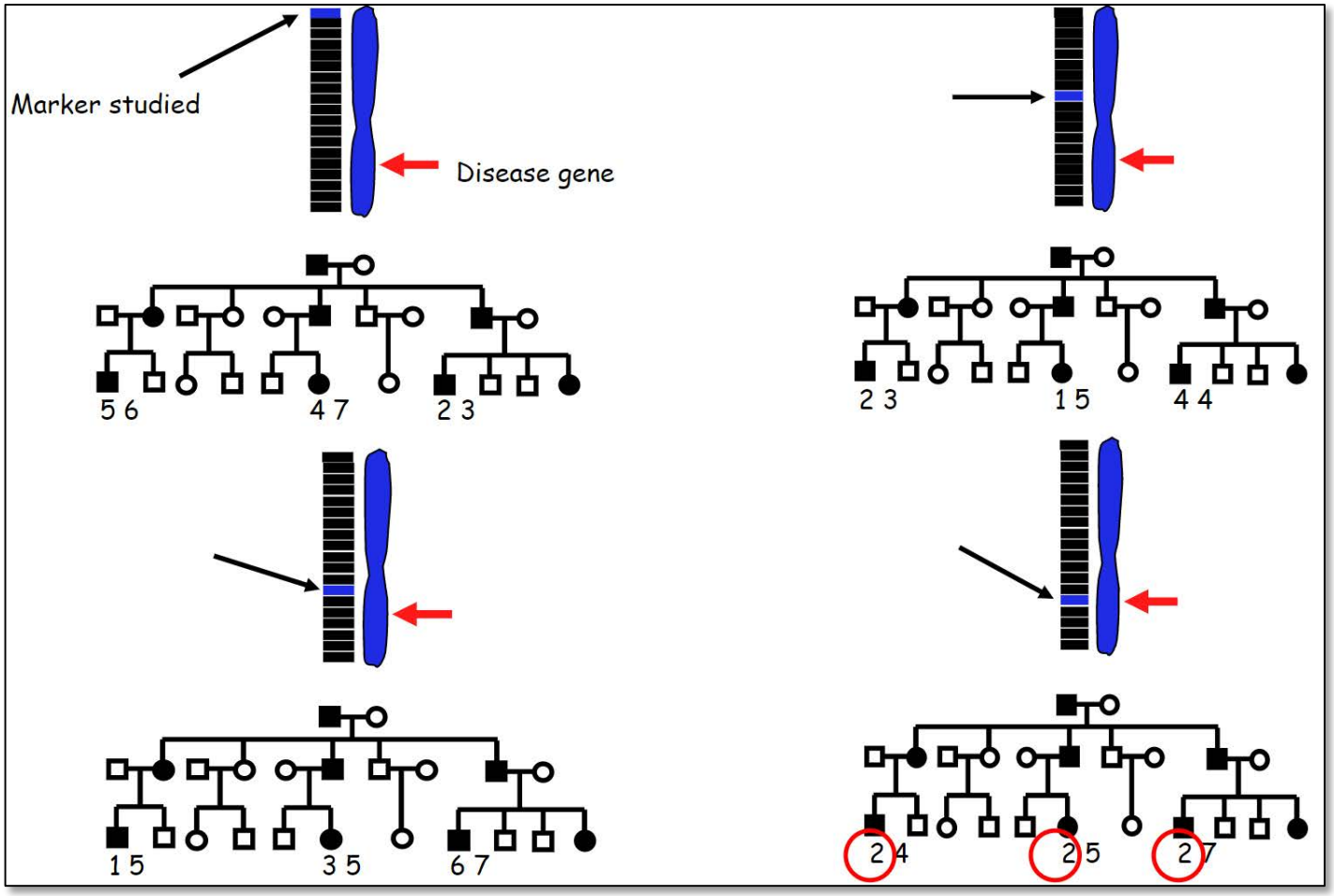
Molecular genetic studies of complex phenotypes. Marian - Translational Research, 2012

Linkage analysis allows mapping of genetic traits



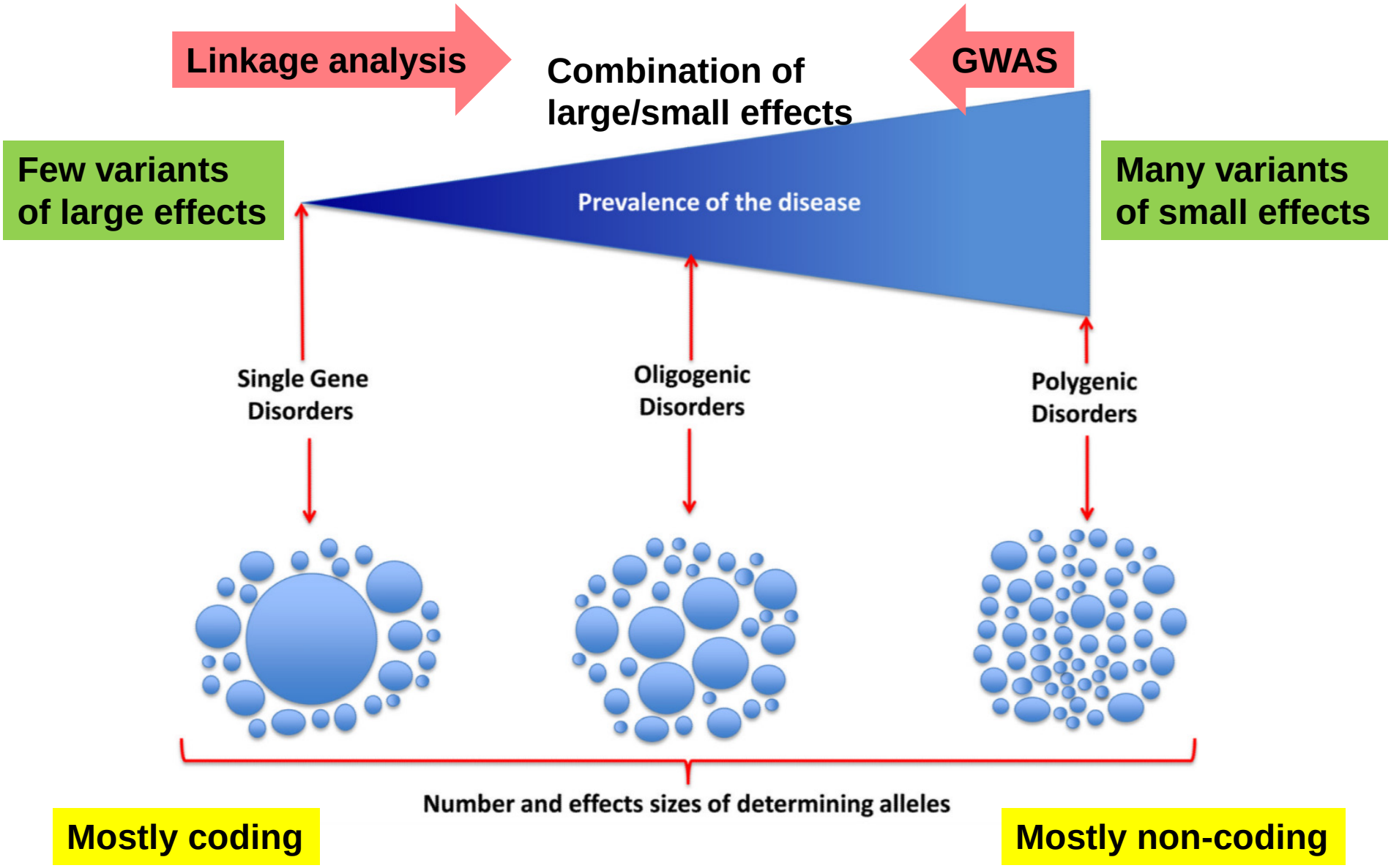
- Frequency of co-inheritance tells you about genetic distance on the chromosome
- Allows making of genetic map of genetic elements (eventually recognized to be genes)
- Used in human to lay out chromosomal maps based on inheritance of STR markers, well before mapping of first genetic trait in human

Linkage analysis allows mapping of disease loci



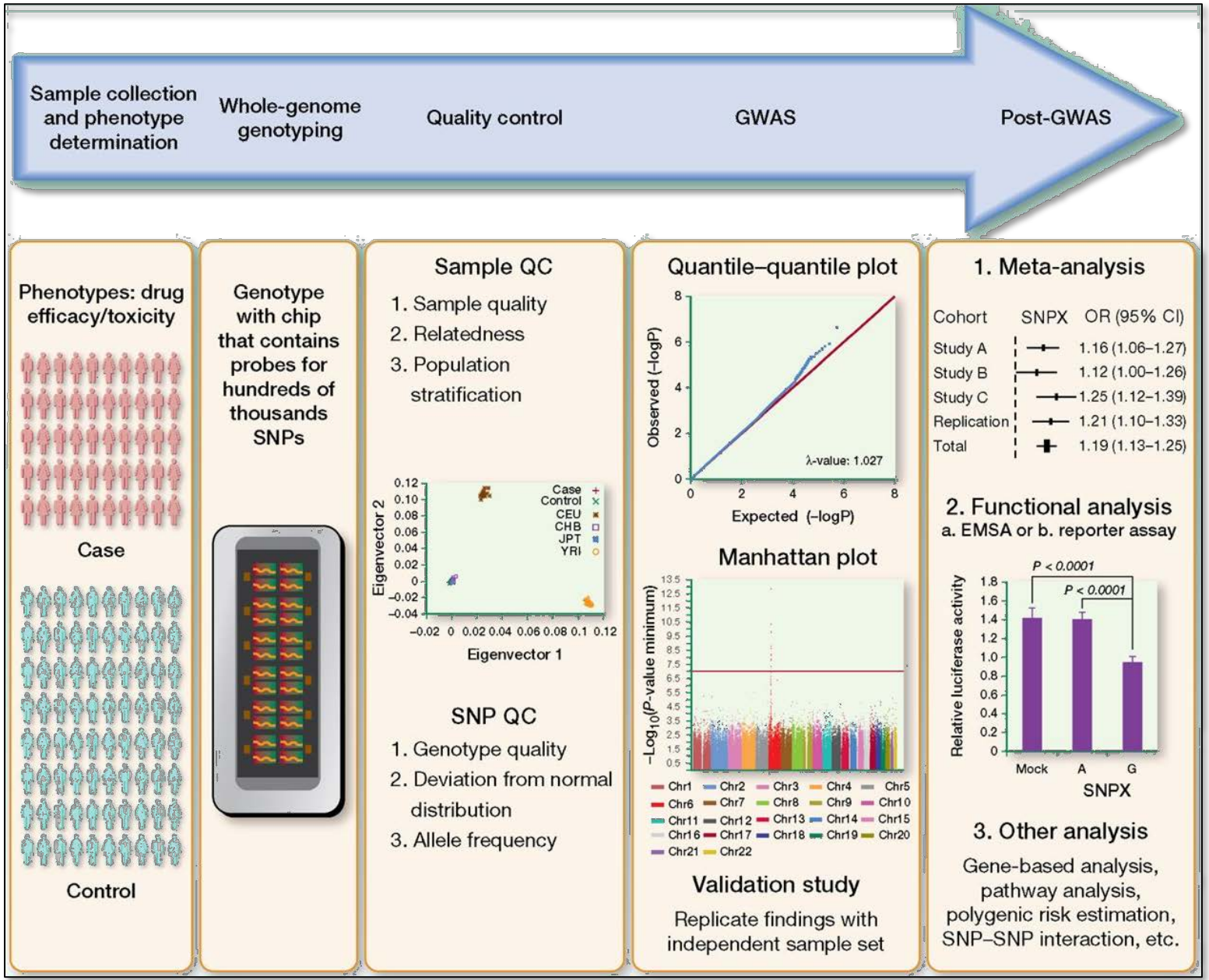
- Exploit human STR marker maps, to search for co-inheritance between STR markers and disease inheritance patterns
- Search for such co-segregation in many independent families (each carries a diff. mutation, but all map to the same region)
- Led to mapping of many Mendelian disease loci in human

Monogenic vs. oligogenic vs. polygenic disorders



Molecular genetic studies of complex phenotypes. Marian - Translational Research, 2012

GWAS: basic study overview



Testing for association

- Most straightforward: compare proportion of each SNP allele in cases and controls

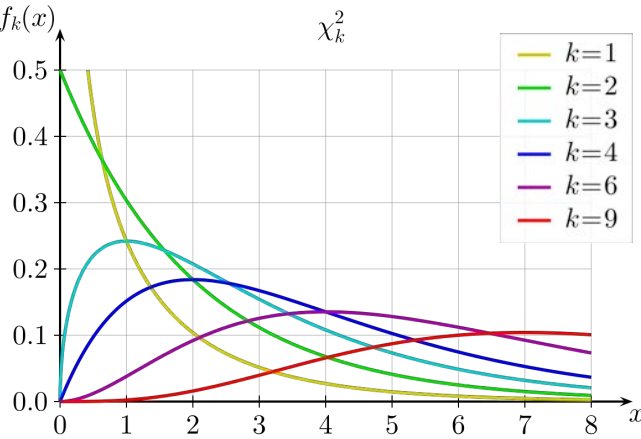
rs11209026	Allele A	Allele G
Cases	22	976
Controls	68	932

Chi-sq = 24.5, p=7.3 x 10^{-7}

Expected	Allele A	Allele G
Cases	47	951
Controls	47	953

(O-E)^2/E	Allele A	Allele G
Cases	13.4	0.7
Controls	9.2	0.5

$$\chi^2 = \sum(O - E)^2/E$$



Simplest tests (single marker regression, χ^2) rule the day.
Association results requiring arcane statistics.
Complex multi-marker models are often less reliable

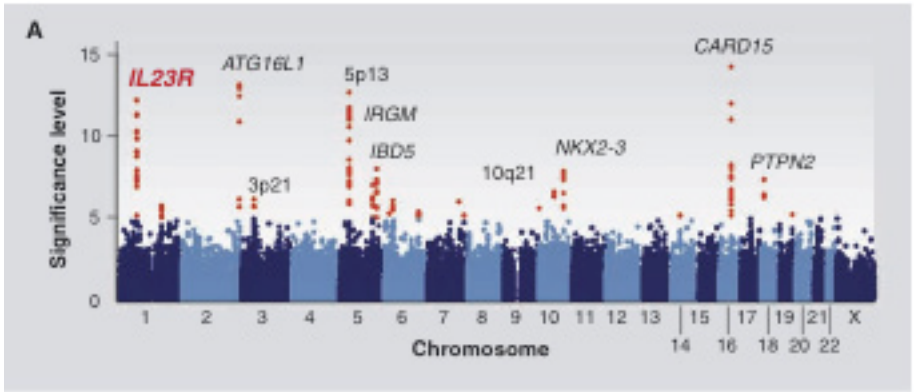
Multiple Testing

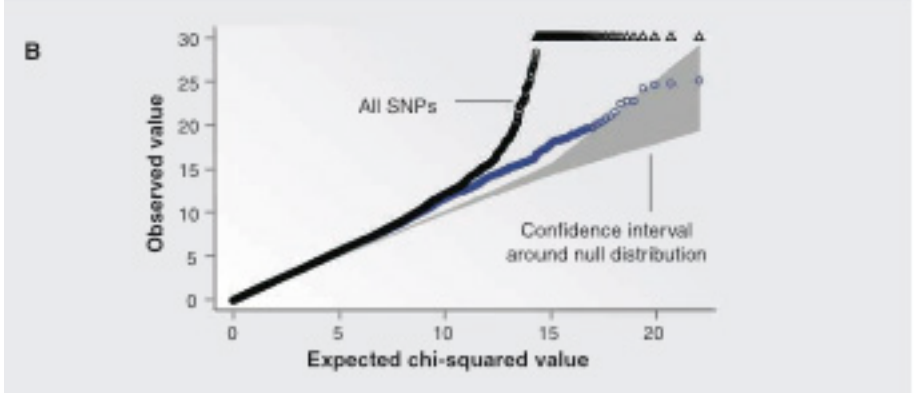
- In linkage, p = .001 (.05 / ~50 chromosomal arms) considered potentially significant

- In GWAS, we're performing $O(10^6)$ tests that are largely independent
  - Each study has hundreds of p<.001 purely by statistical chance (no real relationship to disease)
  - "Genome-wide significance" often set at $p=5x10^{-8}$ (= .05 / 1 million tests)

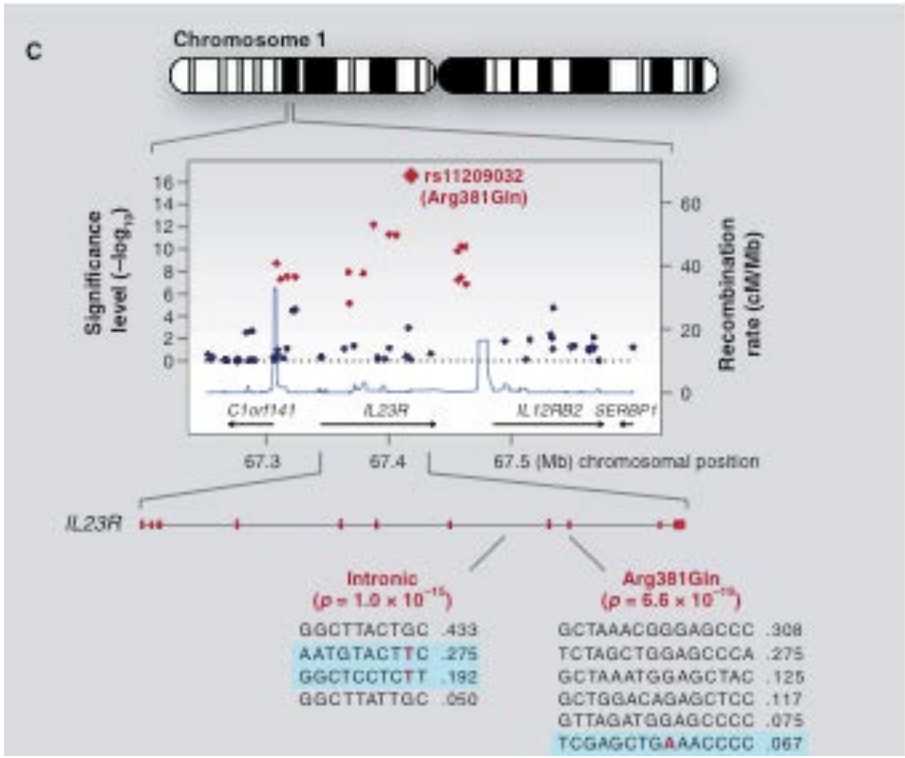
Genome-wide Association

‘Manhattan’ plot

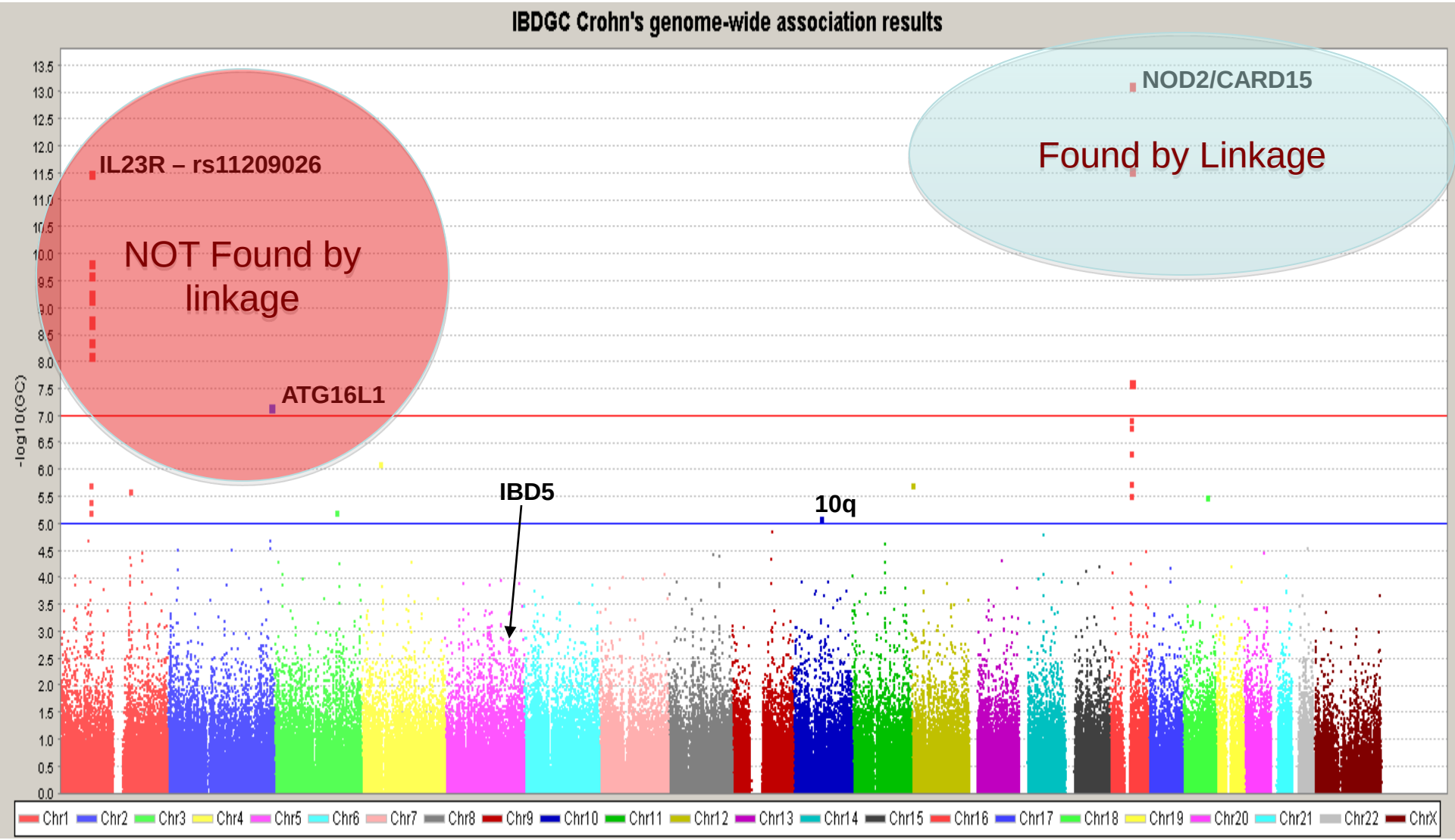




Q-Q plot



Search for gene / mechanism

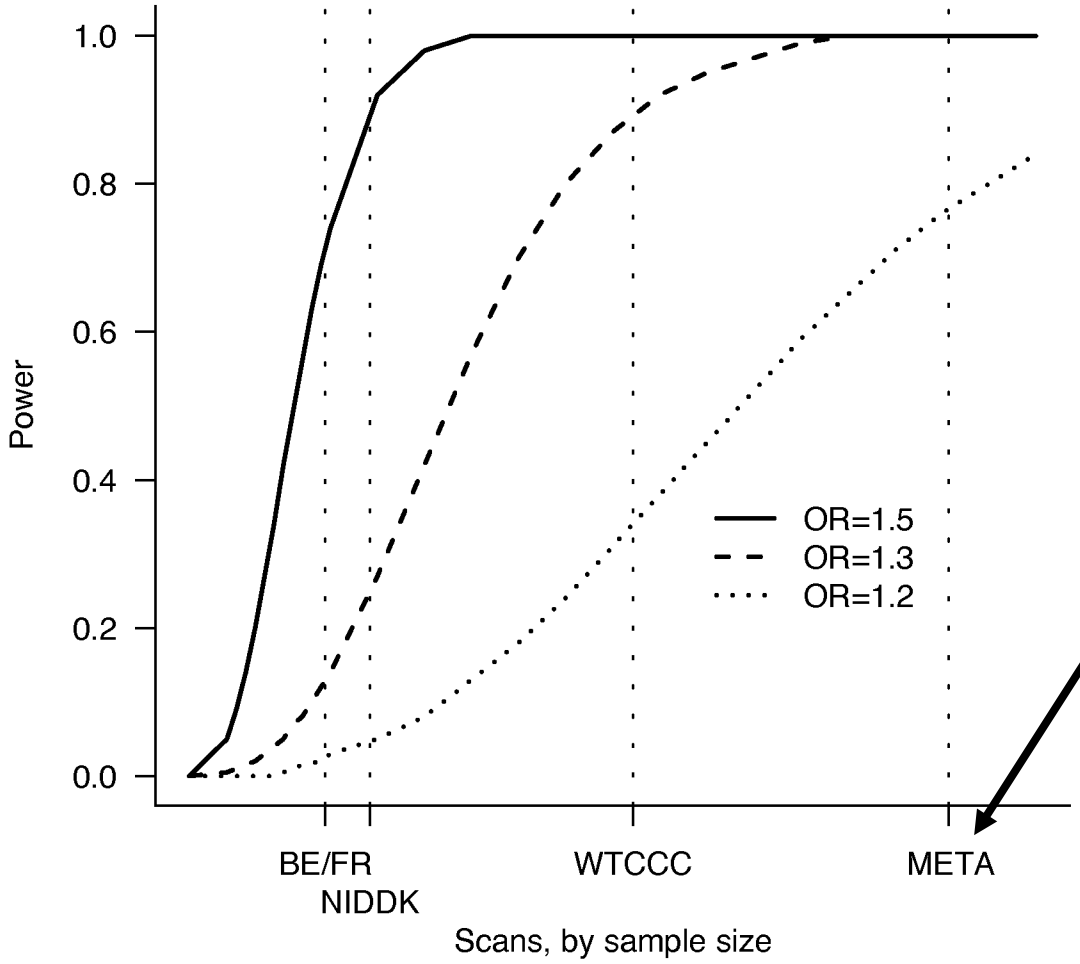

IBDGC Crohn's genome-wide association results
IL23R – rs11209026
NOT Found by linkage
ATG16L1
NOD2/CARD15
Found by Linkage
IBD5
10q
-log10(GC)
Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8 Chr9 Chr10 Chr11 Chr12 Chr13 Chr14 Chr15 Chr16 Chr17 Chr18 Chr19 Chr20 Chr21 Chr22 ChrX

Linkage vs. GWAS capture different variants

NOD2: low-frequency, strong risk variants
IL23R: low-frequency, strong protective variant
ATG16L1: common associated variant

Locus	Frequency	Odds-ratio	ASSOCIATION cases to achieve GWS	LINKAGE Pedigrees to achieve signif.
NOD2 (3 coding SNPs)	5%	3.0	435	1400
IL23R (Arg381Gln)	7%	0.33	817	~30,000
ATG16L1 (Thr300Ala)	50%	1.4	1360	~40,000

Combining studies yields greater power



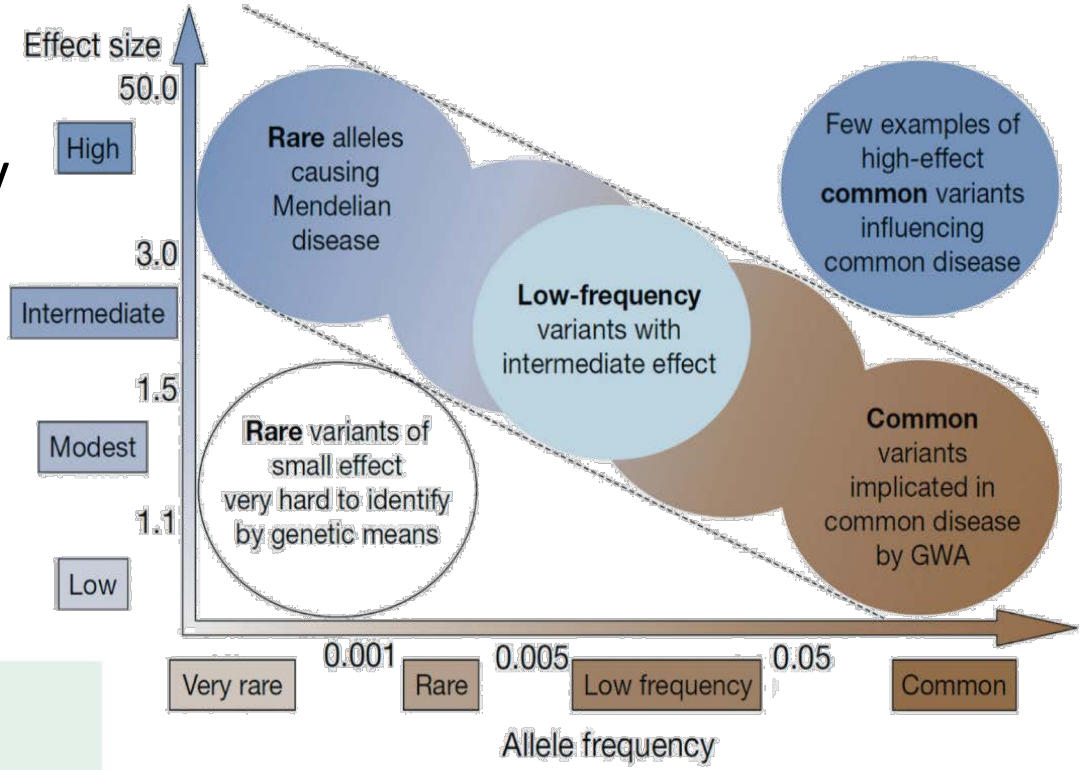
(Example – associated SNP with MAF = 0.20)

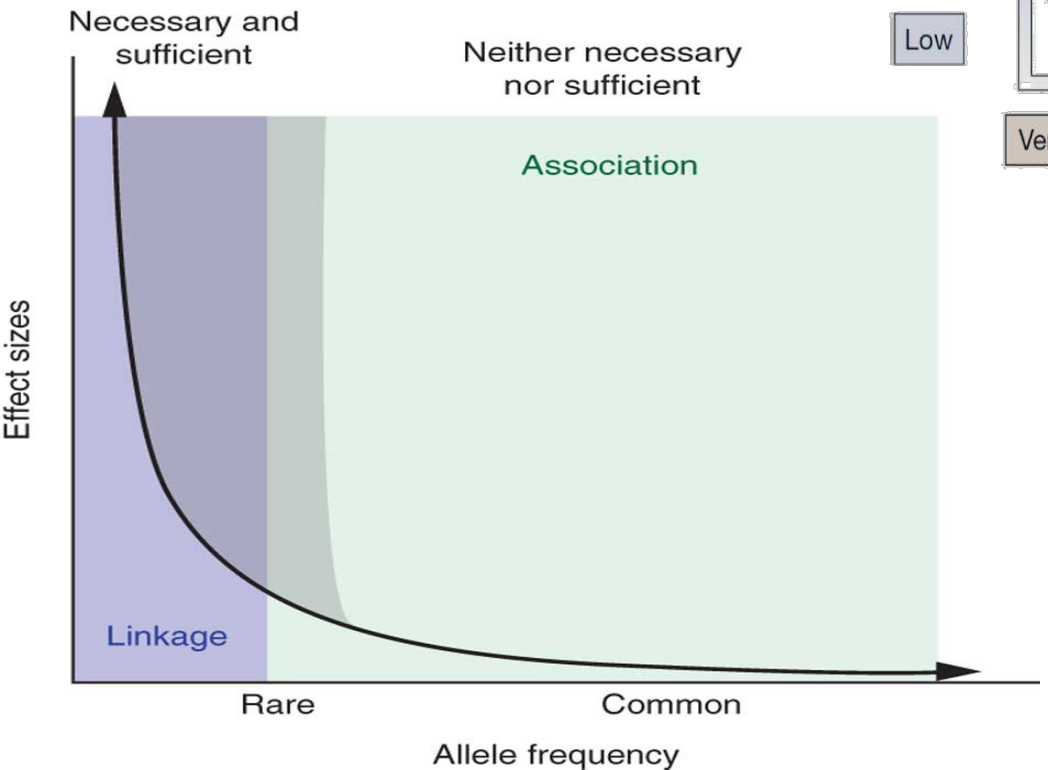
Opportunity: by combining three published studies, we reap the power of an 8000 sample GWAS

Nearly all progress in GWAS has been the result of multiple study meta-analysis

Common alleles typically have small effects

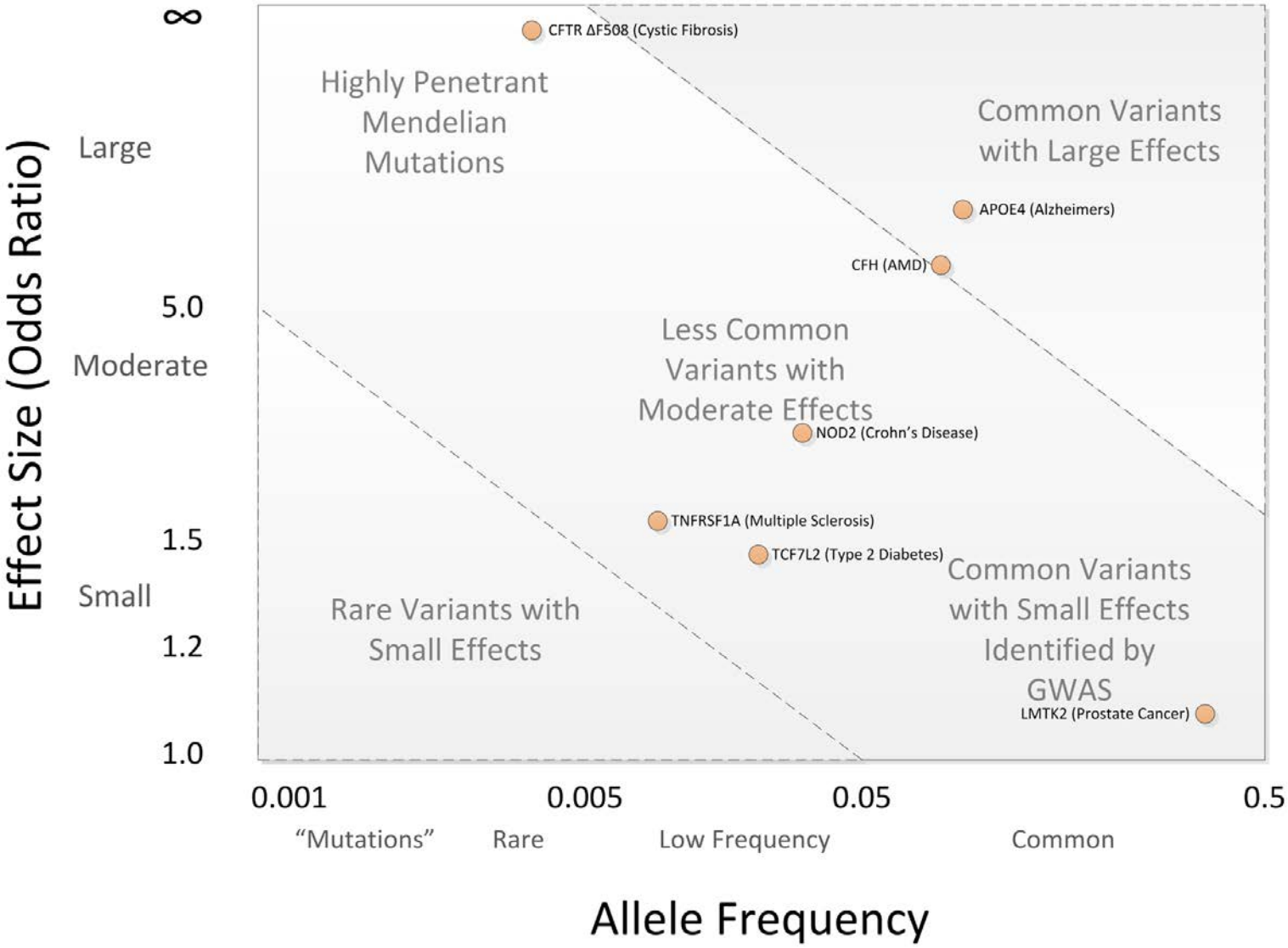
Discovery method tuned to variant effect size/frequency





Discovery method tuned to variant effect size/frequency

GWAS-vs-Linkage best in different freq/effect regimes

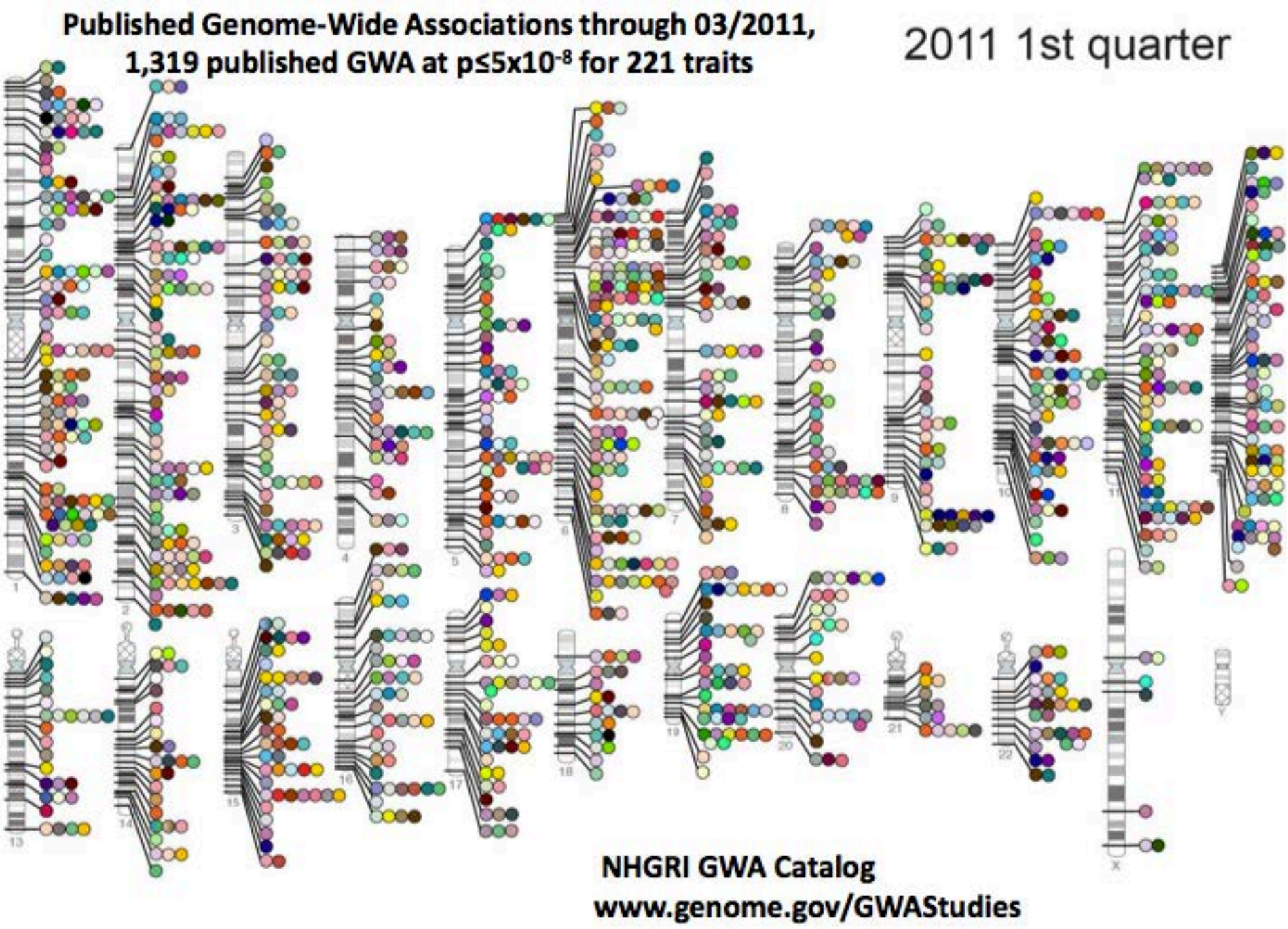


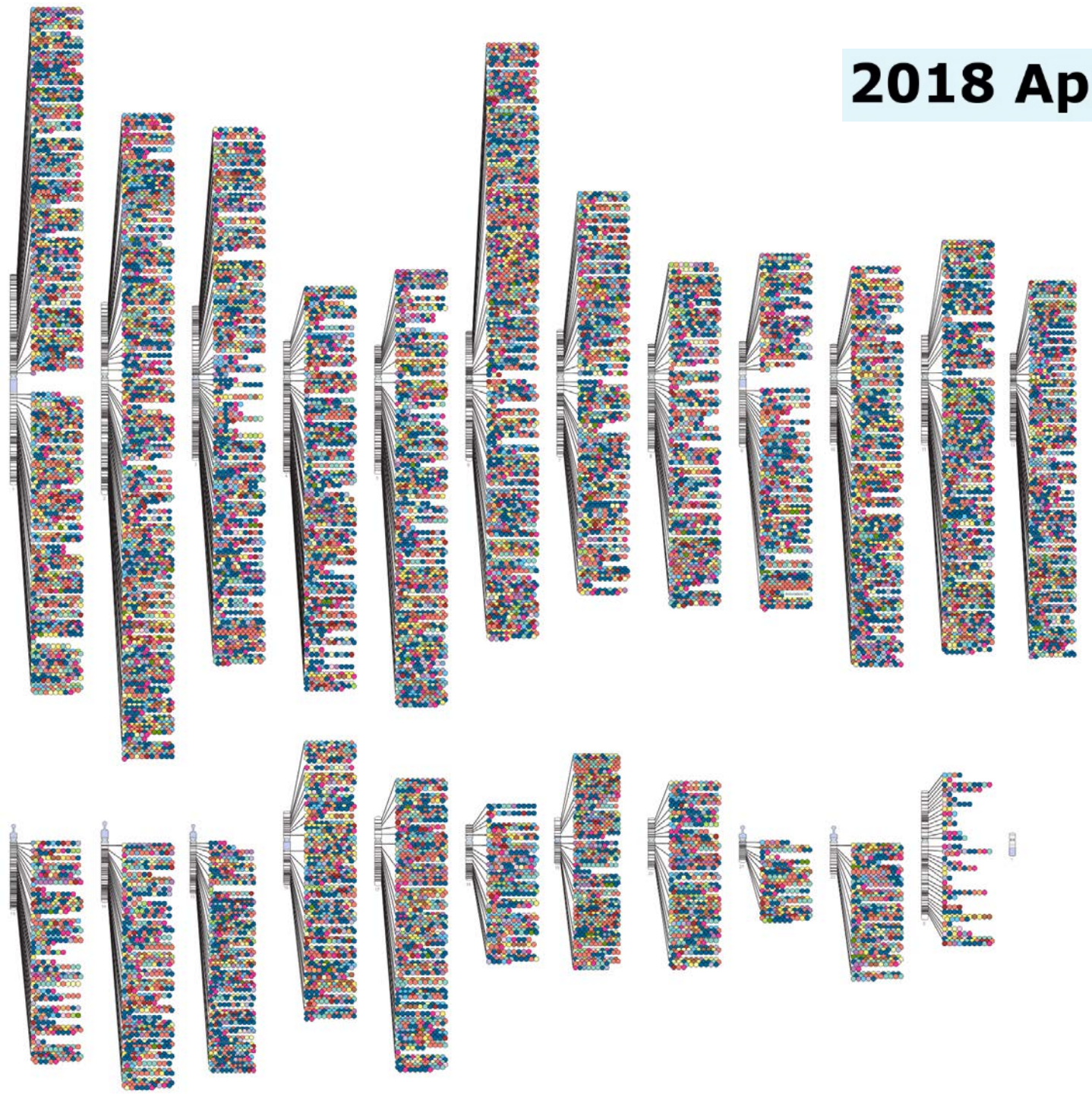
Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting: Common/rare alleles, Linkage vs. GWAS
3. LD, Haplotypes, Co-inheritance, and the challenge of fine-mapping
4. From locus to mechanism - Case study: FTO and Obesity
5. Epigenomics-GWAS integration: ENCODE, Roadmap, EpiMap
6. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA
7. Interpreting non-coding variation: DeepSEA (Jian Zhu guest lecture)

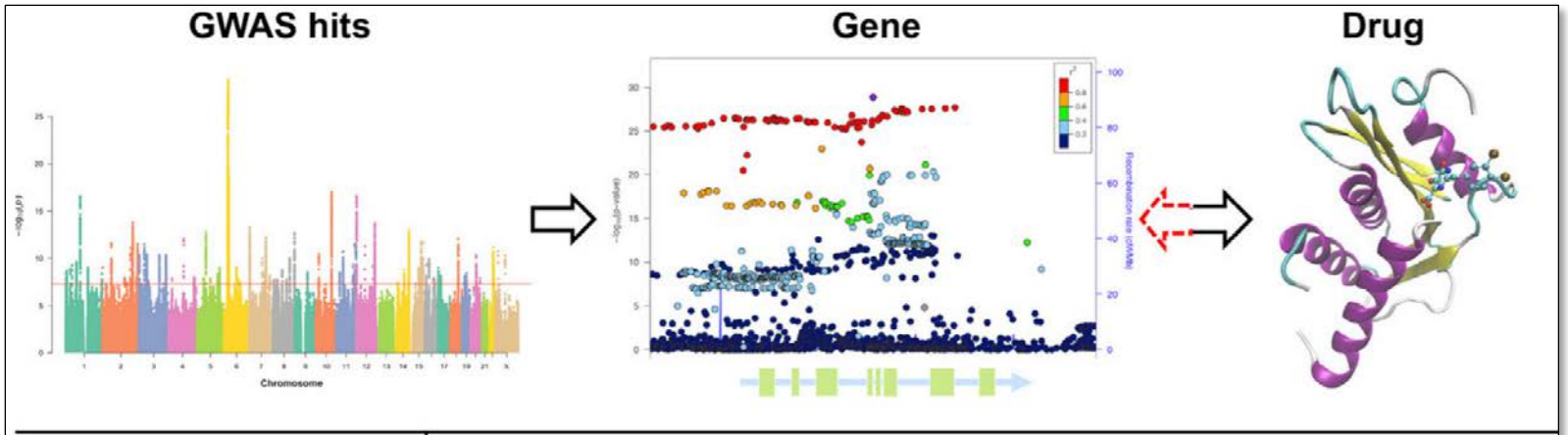
3. LD, haplotypes, and fine-mapping

Linkage disequilibrium, Co-inheritance, Recombination, meiosis, and PRDM9


Published Genome-Wide Associations through 03/2011,
1,319 published GWA at p≤5x10-8 for 221 traits
2011 1st quarter
NHGRI GWA Catalog
www.genome.gov/GWAStudies


2018 Apr
Associations: 69,885
Studies: 5,152
Papers: 3,378
Digestive system disease 968
Cardiovascular disease 674
Metabolic disease 226
Immune system disease 1201
Nervous system disease 1065
Liver enzyme measurement 154
Lipid or lipoprotein measurement 464
Inflammatory marker measurement 326
Hematological measurement 2249
Body weights and measures 1158
Cardiovascular measurement 679
Other measurement 5044
Response to drug 275
Biological process 631
Cancer 979
Other disease 1160
Other trait 3871

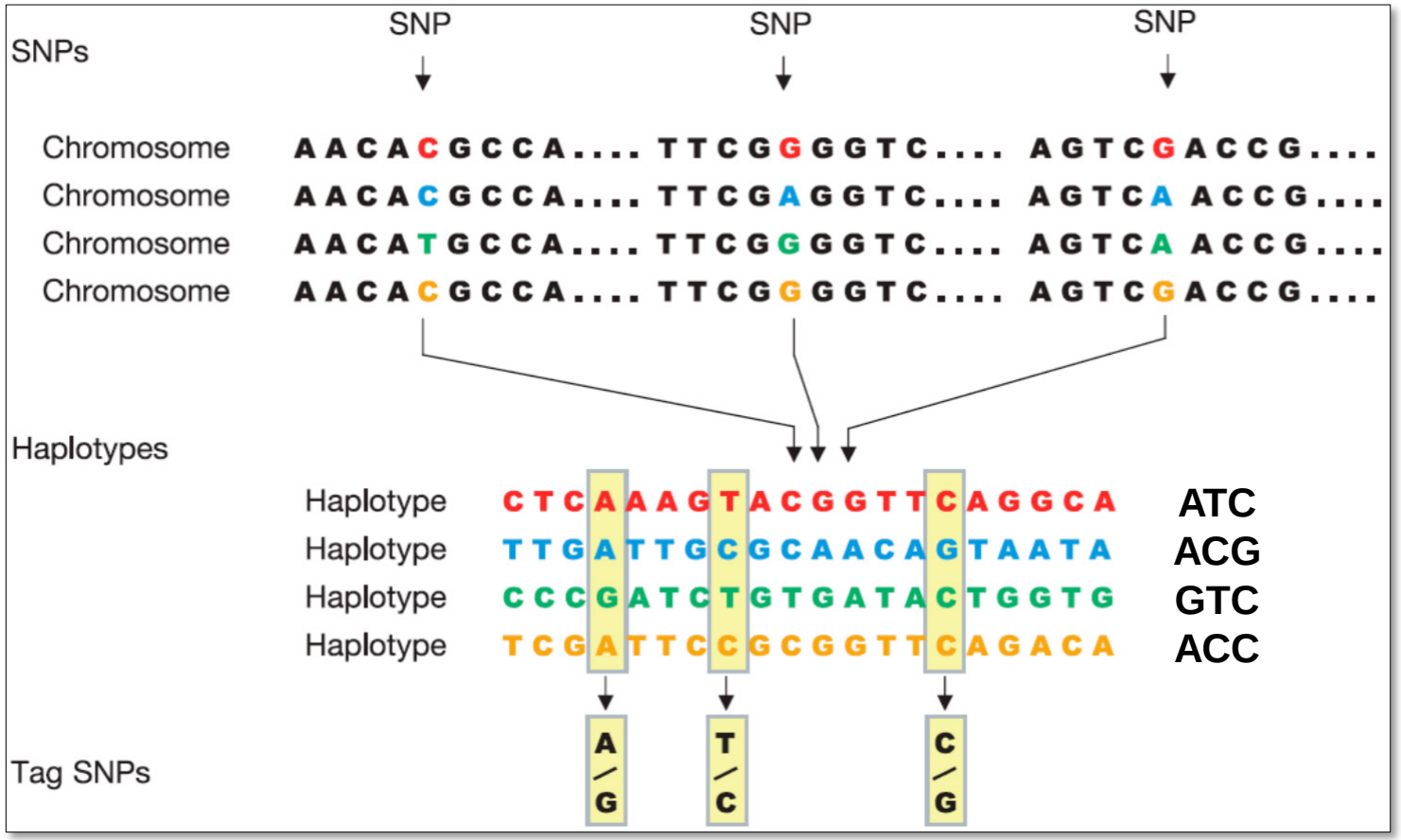
From GWAS hits → to genes → to therapeutics

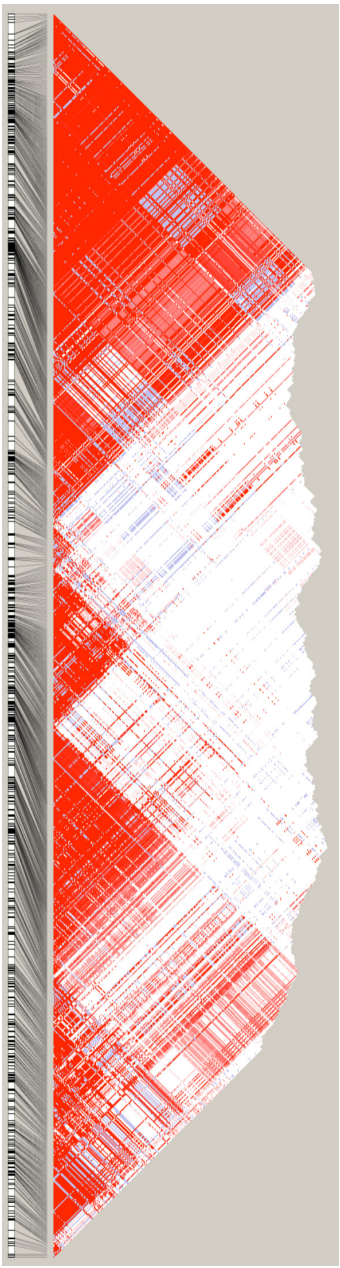


Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

https://www.cell.com/ajhg/fulltext/S0002-9297(17)30240-9

Common variants (SNPs) live in Haplotypes





- Common SNPs only once every 1000 nucleotides or so
- These are co-inherited, so only need to profile a subset
- Markers selected for haplotype profiling are “tag” SNPs

Quantifying Linkage Disequilibrium: D and D'

- Genetic variants do not segregate independently
- D = coeff. of linkage disequilibrium between alleles A and B at loci L1 and L2
  - $D_{AB}=P_{11}P_{00}-P_{10}P_{01}=0.07$
  - Property of the specific **alleles**. Different alleles at these loci will have diff D_{AB}
- If independent, then $D_{AB}=0$ ($P_{11}P_{00}=P_{10}P_{01}$)
- Linkage disequilibrium measures the degree of departure from Mendel's laws of independent assortment

Haplotype AB	Marginal allele frequency
0*	0.54
1*	0.46
*0	0.30
*1	0.60

How to interpret actual values?

- Relative to D_{ABmax}, which depends on frequencies of individual alleles at A, B
- $D_{ABmax}=P_{0*}P_{*1}-P_{1*}P_{*0}=0.138$
- $D'=D/D_{max}=0.51$

➔ 51% of max possible disequilibrium

Haplotype	Expected	Observed
00	0.162	**0.24****
01	0.324	0.31
10	0.138	**0.07****
11	0.276	**0.39****

Quantifying Linkage Disequilibrium: r^2

- Define
- $r^2 = \frac{D^2}{P(A=0)P(B=0)P(A=1)P(B=1)} = 0.37$
- This really is the squared Pearson correlation of the two SNPs
- In practice, Pearson correlation is efficiently computed for all SNPs in windows as $\mathrm{X}'\mathrm{X}/n$
- This is a fundamental quantity for modeling GWAS *z*-scores

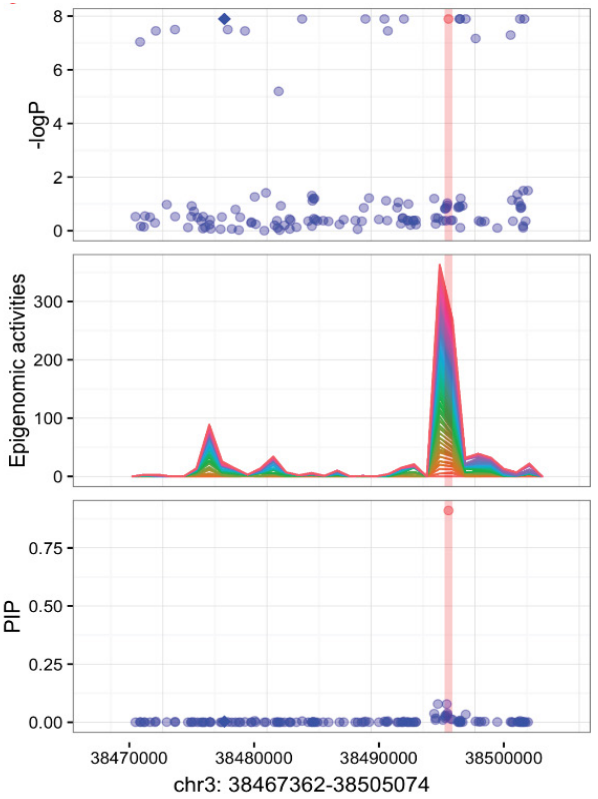
Haplotype AB	Marginal allele frequency
0*	0.54
1*	0.46
*0	0.30
*1	0.60

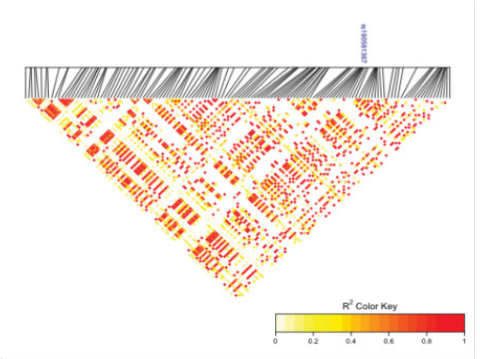
Haplotype	Expected	Observed
00	0.162	0.24
01	0.324	0.31
10	0.138	0.07
11	0.276	0.39

Key property: r^2 correlation for individual SNPs is exactly the r^2 of the GWAS association summary statistics of these SNPs

Fine-mapping disease associations: Epigenomics / functional data

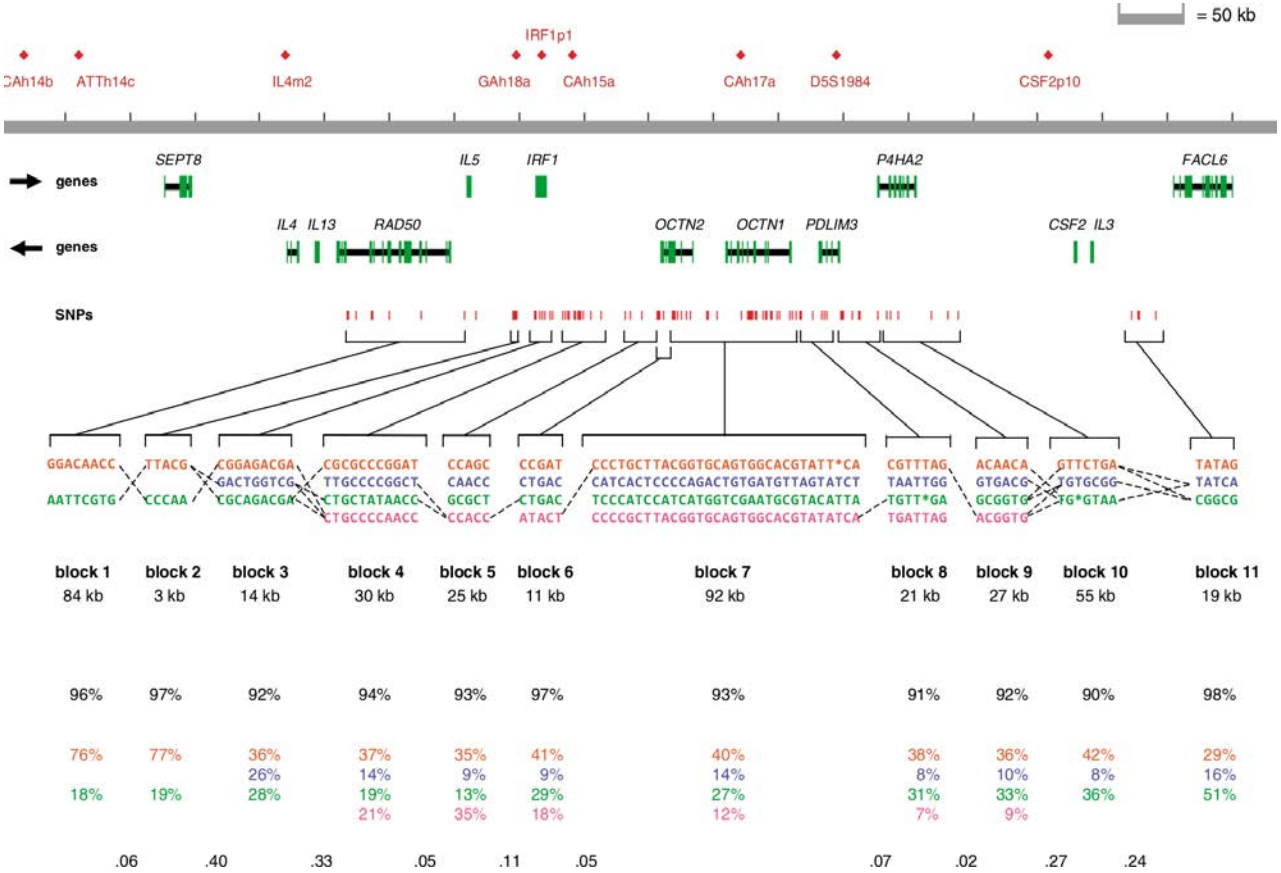
- LD is a **blessing for mapping loci** to disease, as it enables genotyping of just a handful of tag variants
  ➔ enabled the GWAS revolution
- LD is a **curse for fine-mapping** loci into their causal variants
  ➔ many variants are strongly correlated to the true causal variant(s)
  ➔ often indistinguishable scores by genetics alone associations
  ➔ strongest-association SNP might actually be an artifact of LD, and true causal variant may be another one
- Orthogonal data (e.g. epigenomics) often used for fine-mapping





Li and Kellis *BiorXiv* 2016

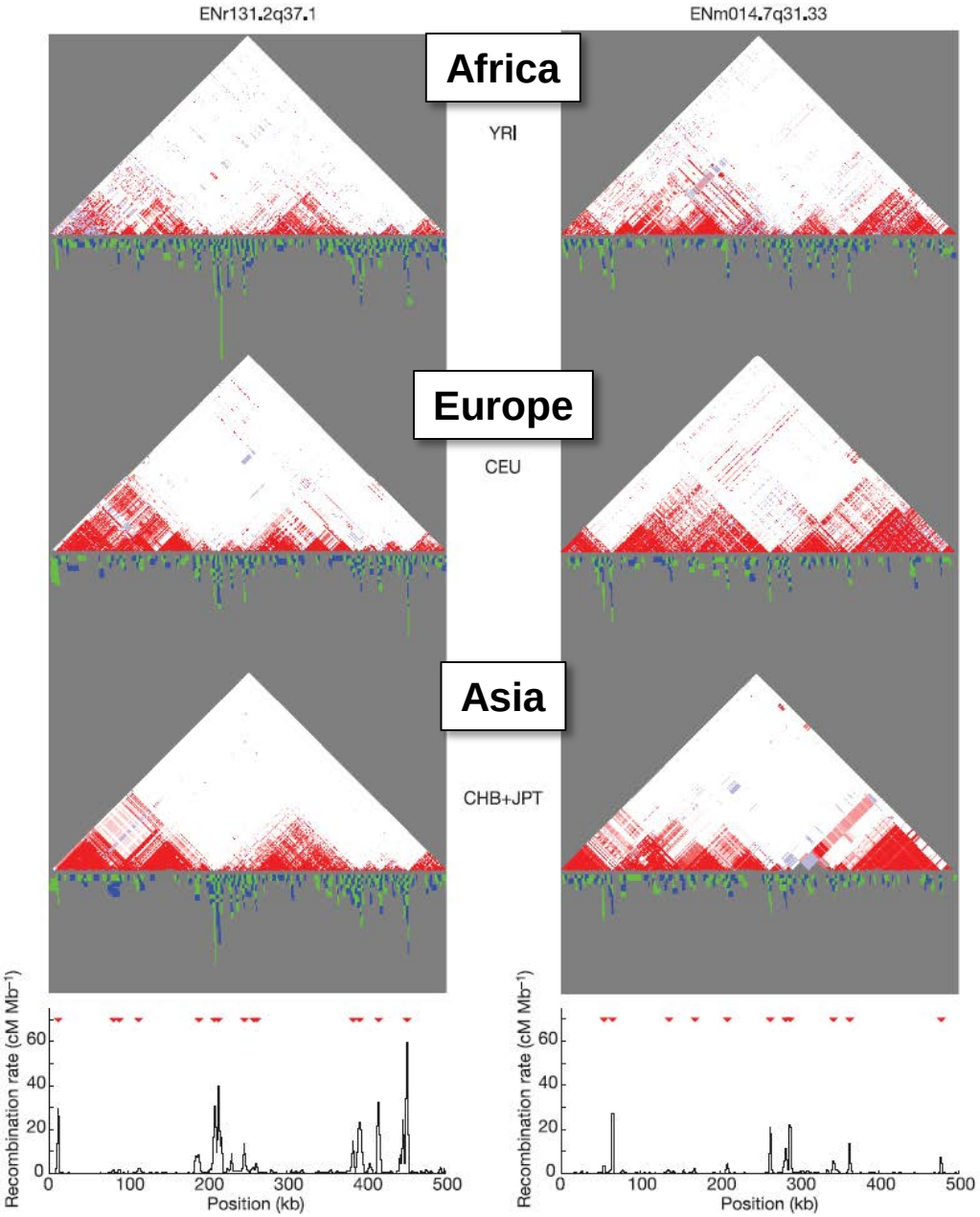
Long-range threading of haplotype blocks



- Relatively few haplotypes exist in the human population (consider 10M SNPs: we don't see 2^{10M} haplotypes!)
- Implies high level of genotype sharing even for unrelated individuals

Daly et al *Nat Genet* 2001

Haplotypes differ across regions/populations

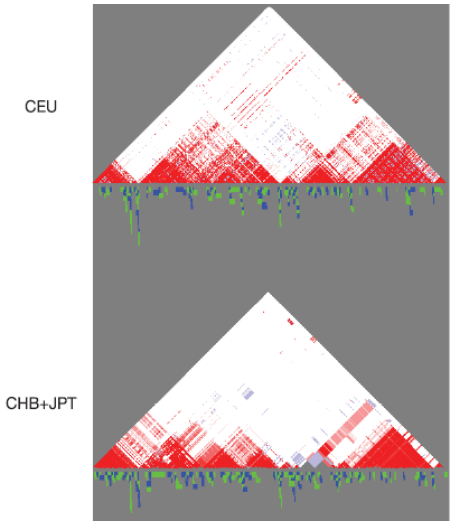


- Recurrent recombination events occur at hotspots
- r^2 correlations between SNPs depend on **historical order** in which they arose

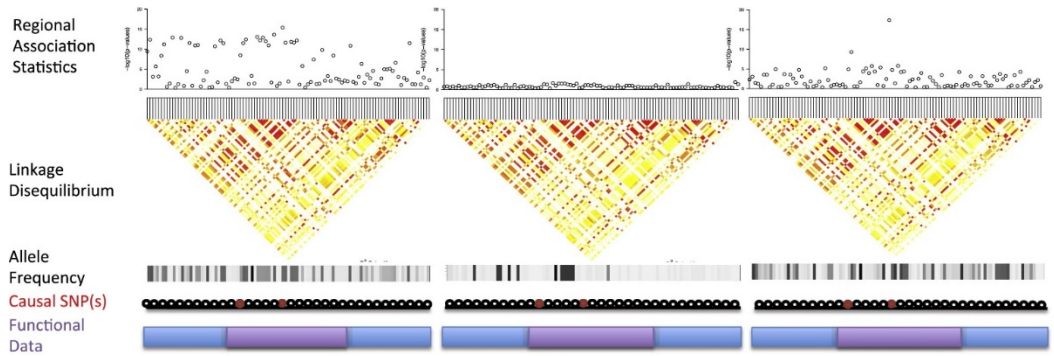
(not in their physical order on the chromosome)

Multi-ethnic analysis can be used for fine-mapping

Case 1: LD boundaries differ



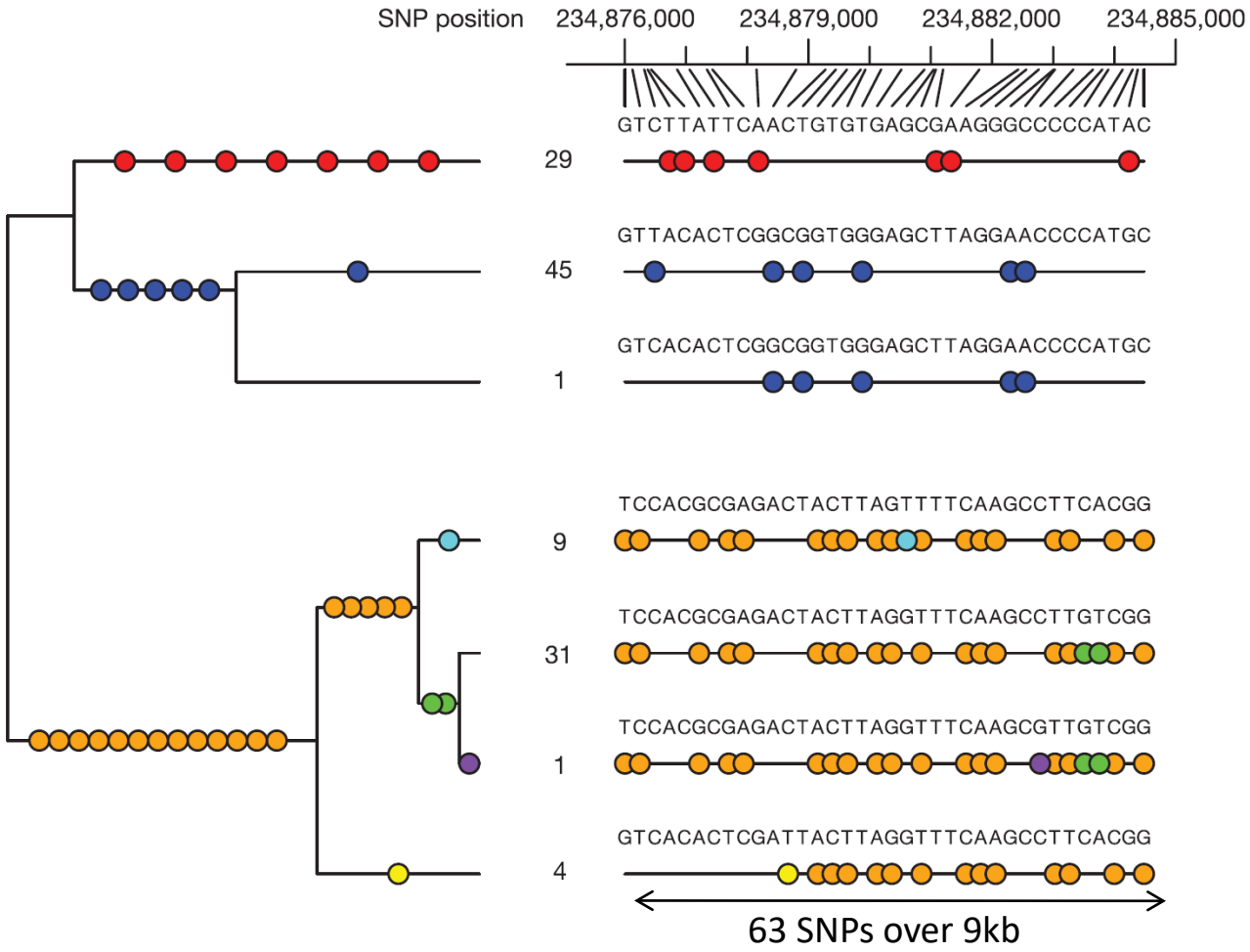
Case 2: allele frequencies differ



- Allele frequencies and LD patterns can differ between populations
- Currently, disease associations are biased for discovery in European cohorts
- As we begin conducting association studies in Asia/Africa, there is a pressing need to develop statistical methods which can account for population genetic differences

Kichaev et al. *Am J Hum Genet* 2015

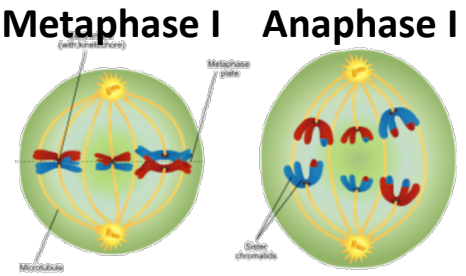
Haplotypes evolve, accumulate mutations

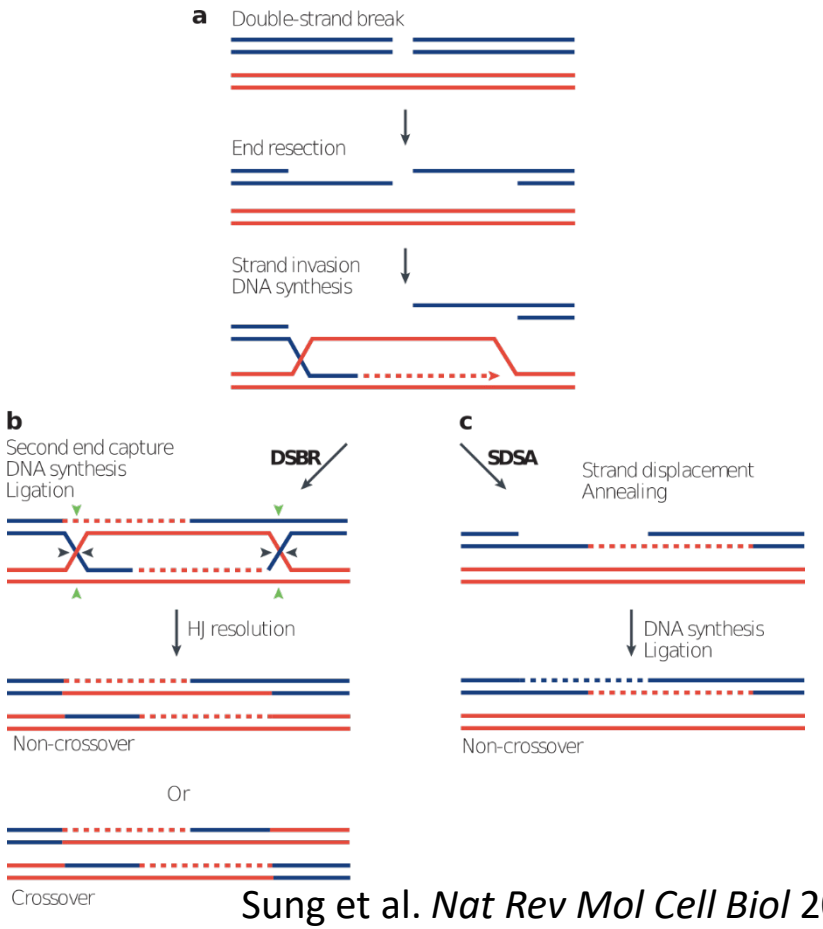


- Example region: 36 SNPs spanning 9kb
- In principle: 2^36 possible allele combinations (haplotypes)
- Sample 120 parental European chromosomes.
- In practice: only 5 recurrent haplotypes seen (and 2 singleton haplotypes)

Haplotypes result from non-uniform recombination

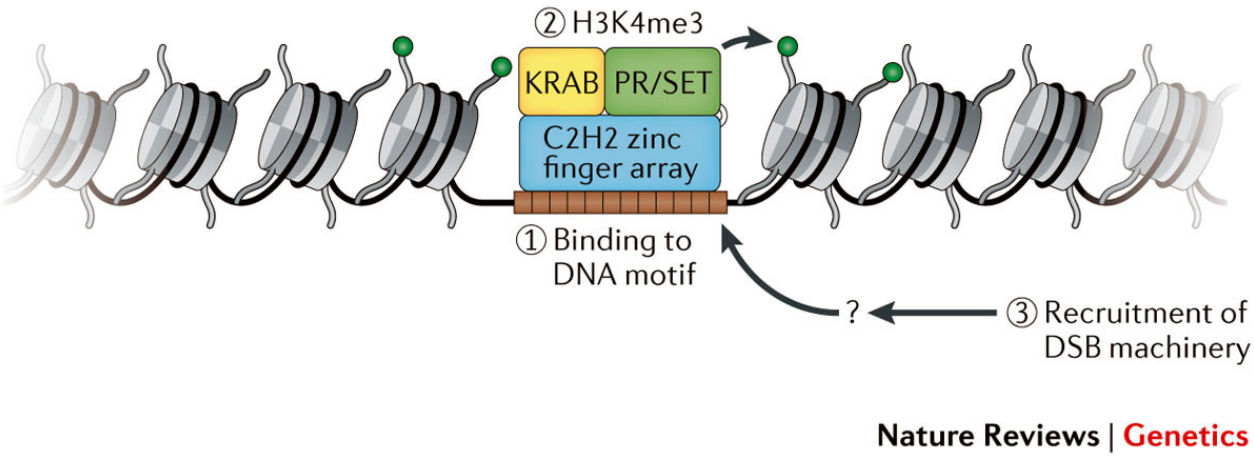
- **Recombination** is crucial for lining up chromosomes during **meiosis** for gamete formation.
- Recombination starts with a **double-stranded break (DSB)**, which is then repaired by strand invasion of the homologous chromosome.
- Repair can lead to either:
  - **Gene conversion**, via strand displacement annealing (SDSA), which transfers a segment of one homologous chromosome into the other, or
  - **Recombination** via cross-over repair of a double-stranded break, leading to new allele combination
- Recombination provides selective **advantage** for **sexual reproduction** (mix and match beneficial alleles)
- Recombination **does not happen uniformly** over each chromosome
- Recombination **hotspots** occur once every 100kb, and recombination occurs hundreds of times more often in hotspots
- Mouse studies revealed the role of **PRDM9** in demarcating hotspots





Sung et al. *Nat Rev Mol Cell Biol* 2006

PRDM9, recombination, and selection (*aka. The tragic love story of PRDM9*)



- PRDM9 is a zinc finger protein which binds to specific DNA motifs, methylates H3K4 surrounding the binding site, and recruits double-strand break enzymes
- PRDM9 is under strong constraint, but the DNA-binding zinc finger array has high mutation rate and is under **positive selection**
- More than 40 known PRDM9 alleles, each with different DNA-binding specificity
- The repaired double strand break no longer contains the PRDM9 motif, leading to evolutionary competition between the protein and its motif

Baudat et al. *Nat Rev Genet* 2013

Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting: Common/rare alleles, Linkage vs. GWAS
3. LD, Haplotypes, Co-inheritance, and the challenge of fine-mapping
4. From locus to mechanism - Case study: FTO and Obesity
5. Epigenomics-GWAS integration: ENCODE, Roadmap, EpiMap
6. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA
7. Interpreting non-coding variation: DeepSEA (Jian Zhu guest lecture)

4. From locus to mechanism

Case study: FTO and Obesity

The NEW ENGLAND JOURNAL of MEDICINE

FTO Obesity Variant Circuitry and Adipocyte Browning in Humans

Melina Claussnitzer, Ph.D., Simon N. Dankel, Ph.D., Kyoung-Han Kim, Ph.D., Gerald Quon, Ph.D., Wouter Meuleman, Ph.D., Christine Haugen, M.Sc., Viktoria Glunk, M.Sc., Isabel S. Sousa, M.Sc., Jacqueline L. Beaudry, Ph.D., Vijitha Puviindran, B.Sc., Nezar A. Abdennur, M.Sc., Jannel Liu, B.Sc., Per-Arne Svensson, Ph.D., Yi-Hsiang Hsu, Ph.D., Daniel J. Drucker, M.D., Gunnar Mellgren, M.D., Ph.D., Chi-Chung Hui, Ph.D., Hans Hauner, M.D., and Manolis Kellis, Ph.D.

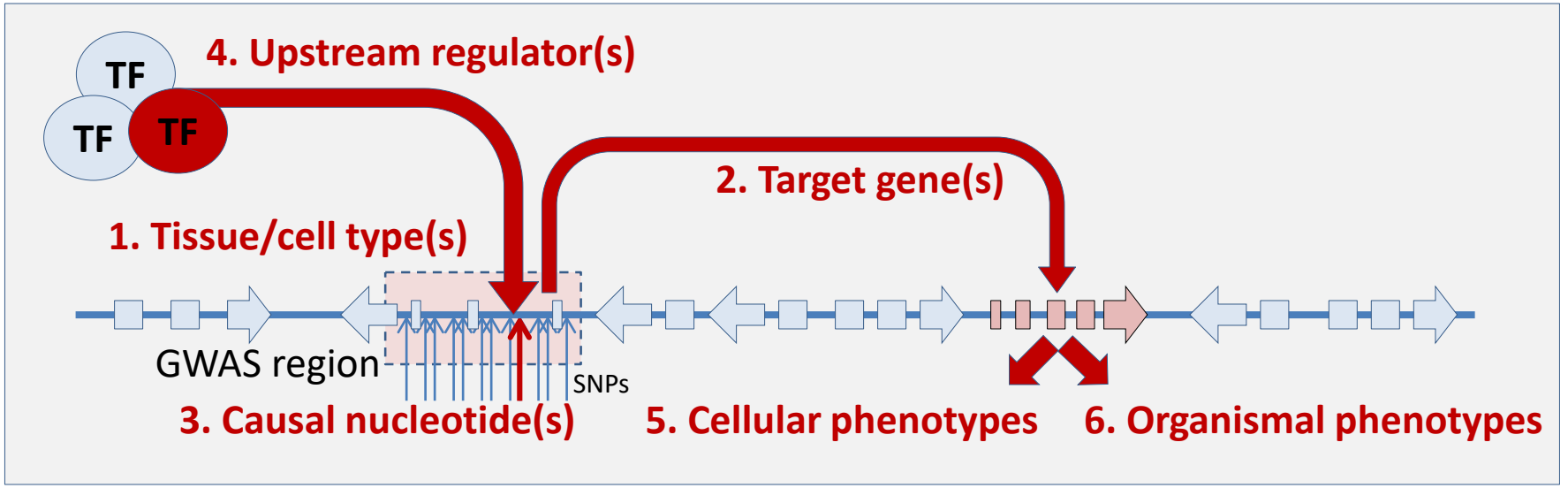
SEPTEMBER 3, 2015 VOL. 373 NO. 10 N Engl J Med 2015;373:895-907.

Mechanistic dissection of a non-coding disease locus

- Identify cell type, causal SNP, regulator, targets, process
- Genome editing demonstrates variant causality
- Adipocyte browning drivers of obesity

Melina Claussnitzer

Dissecting non-coding genetic associations

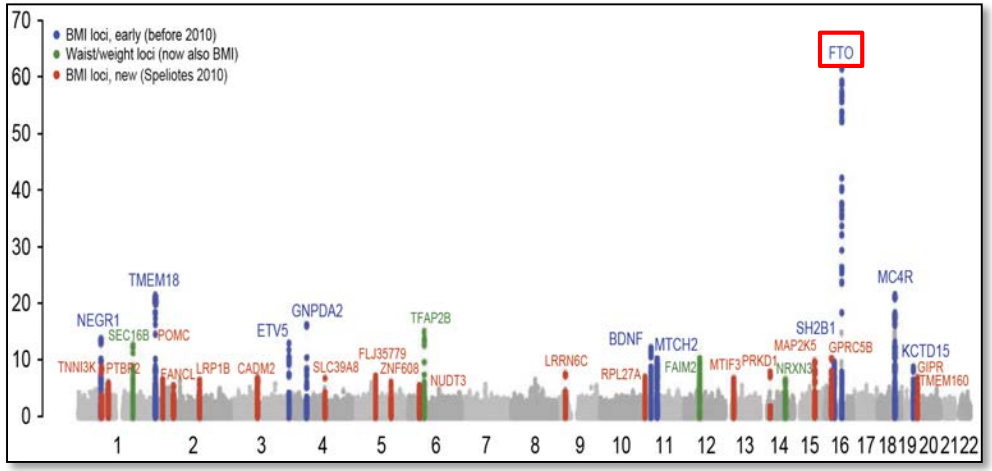


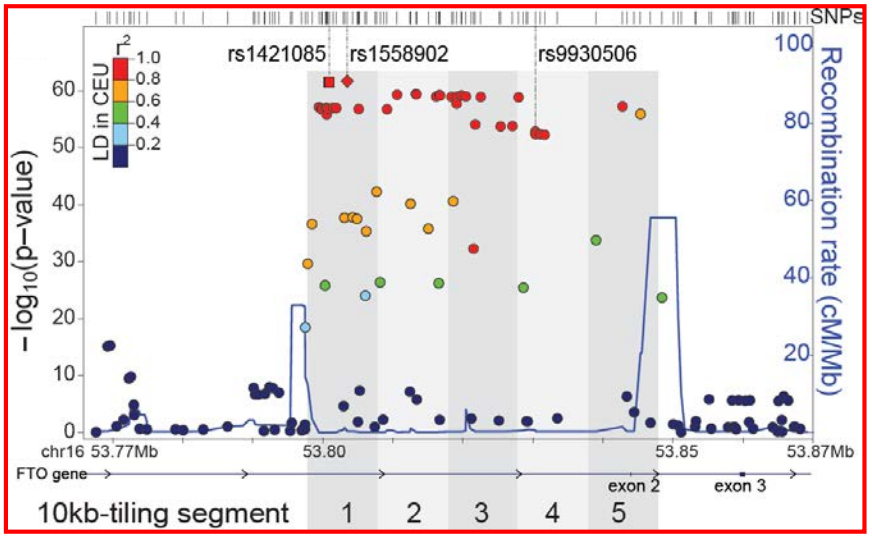
1. Establish relevant **tissue/cell type**
2. Establish downstream **target** gene(s)
3. Establishing **causal** nucleotide variant
4. Establish upstream **regulator** causality
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

Goal:

Apply these to the FTO locus in obesity

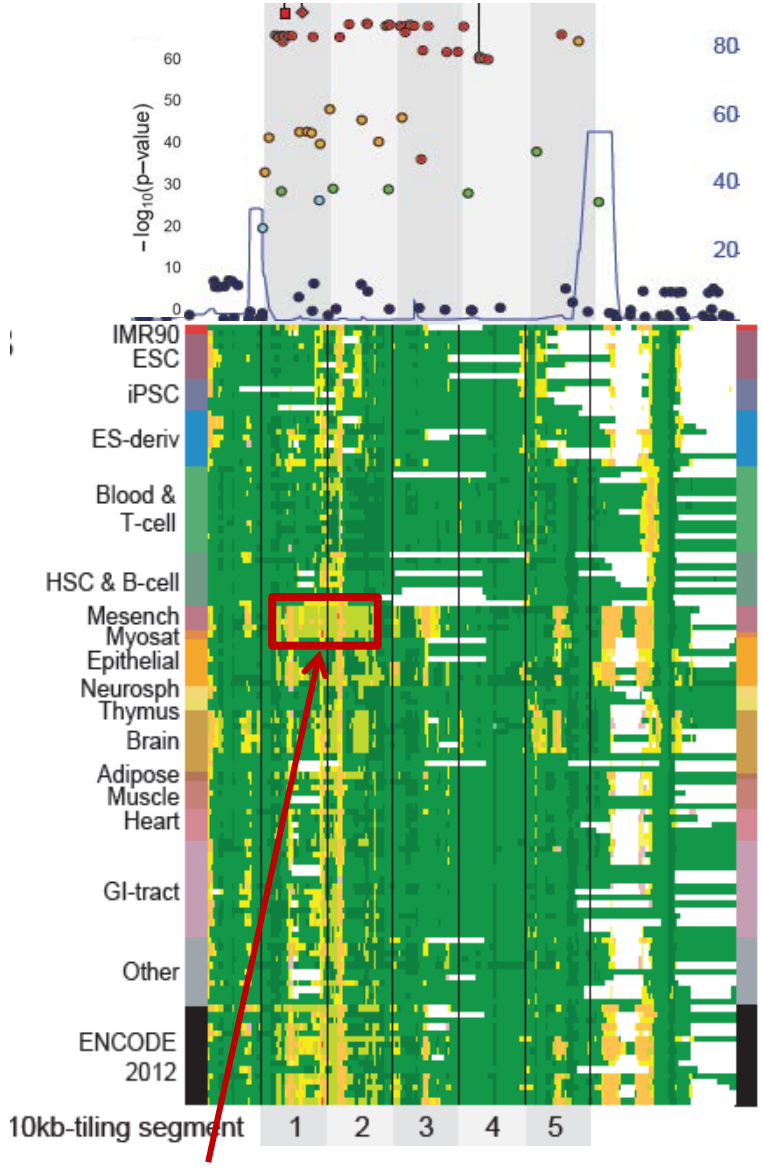
FTO region: strongest association with obesity



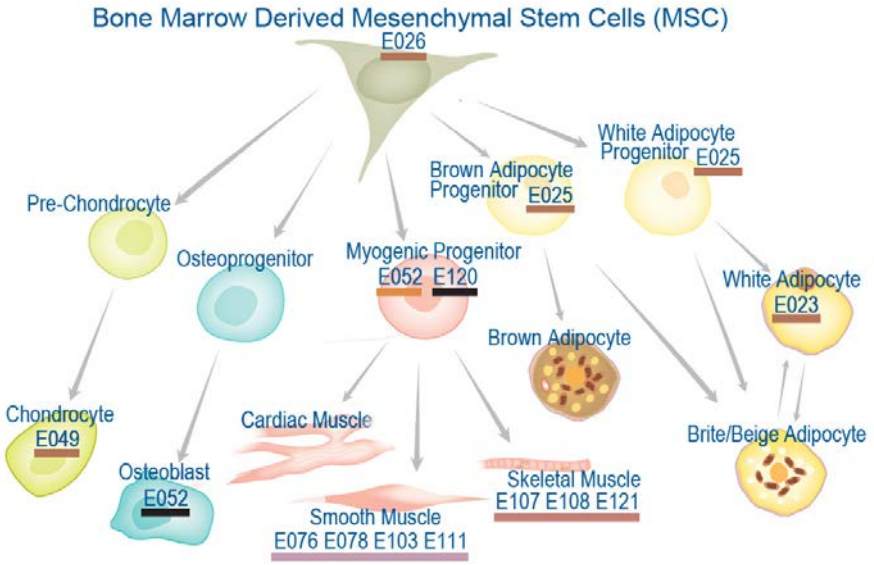


- First and strongest association with obesity (not just ‘your fault’)
- Associated with **obesity**, Type 2 Diabetes, Cardiovascular traits
- 89 variants in LD, spanning 47kb, intron 1 of FTO gene
- No protein-altering variants: regulatory role? Target gene, tissue?

1. Tissue: Chromatin states predict adipocyte function

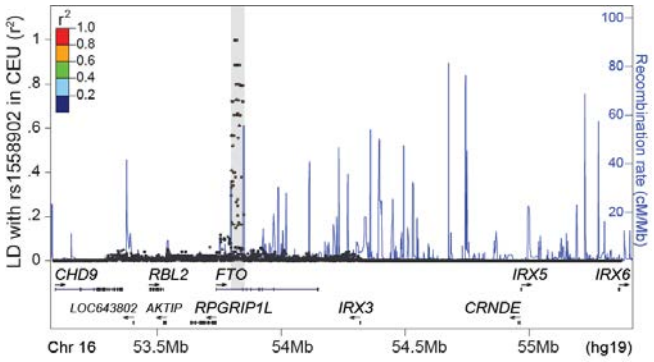


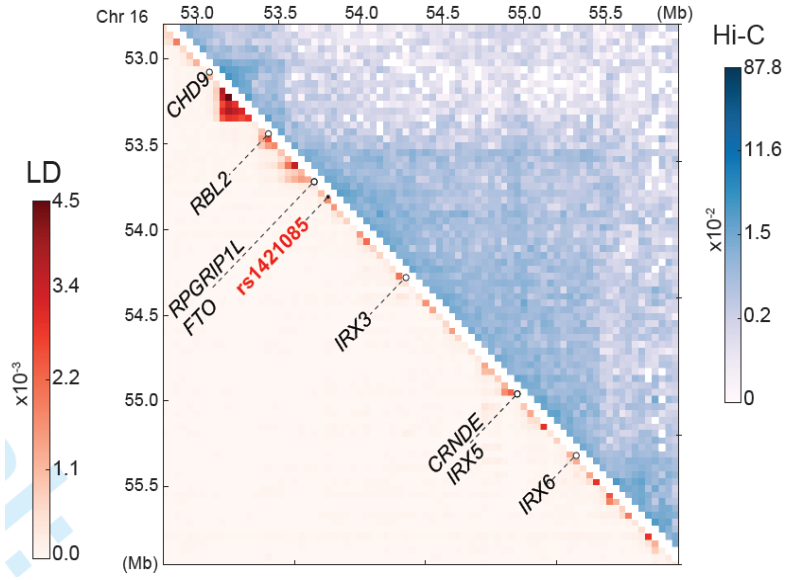
12 kb super-enhancer



Progenitors of white/beige adipocytes

2. Targets: 3D folding and expr. genetics indicate IRX3+IRX5

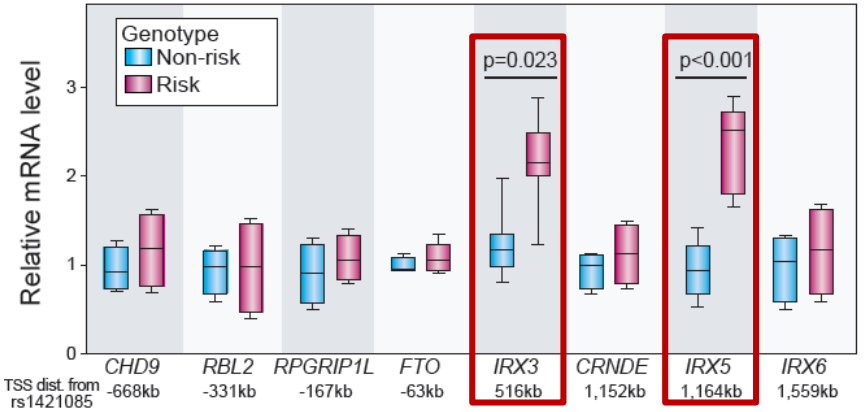




Dixon, Nature 2012

Topological domains span 2.5Mb
Implicate **8 candidate genes**

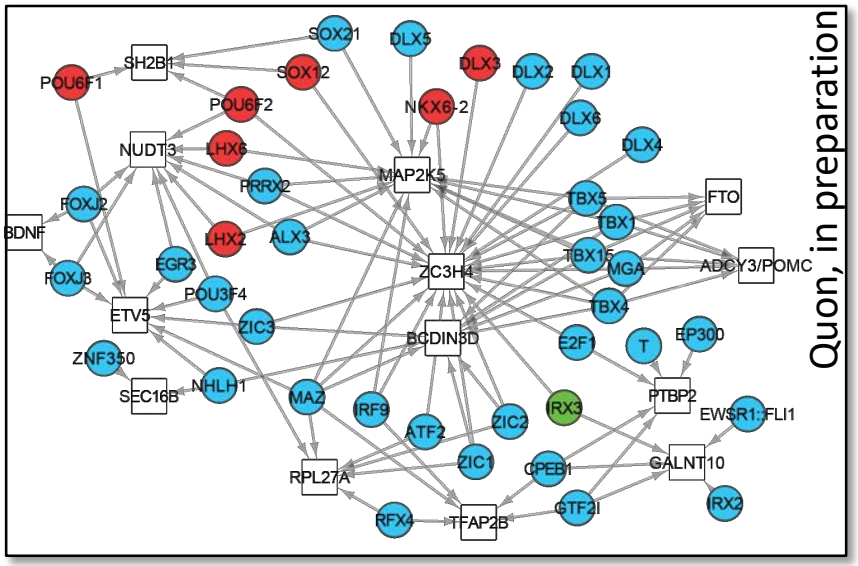
Cohort of **20 homozygous risk** and **18 homozygous non-risk** individuals:
Genotype-dependent expression?



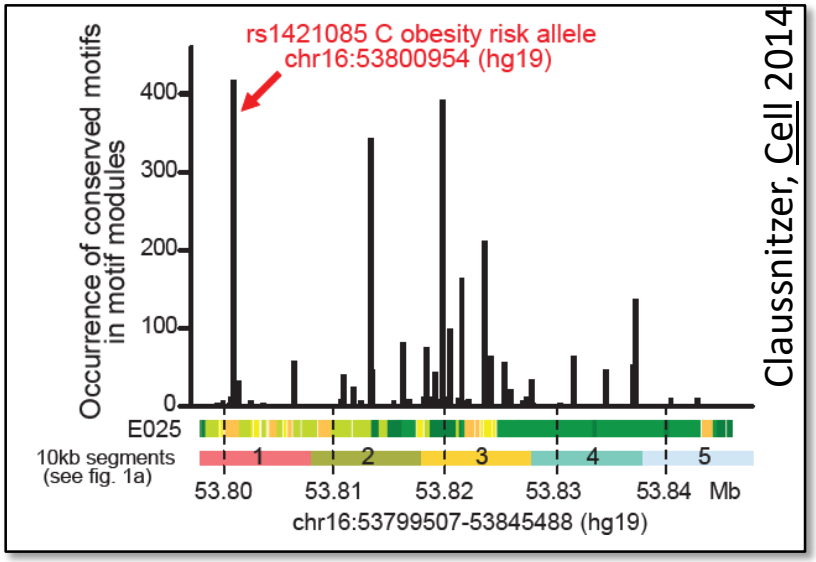
eQTL targets: IRX3 and IRX5

Risk allele: increased expression (gain-of-function)

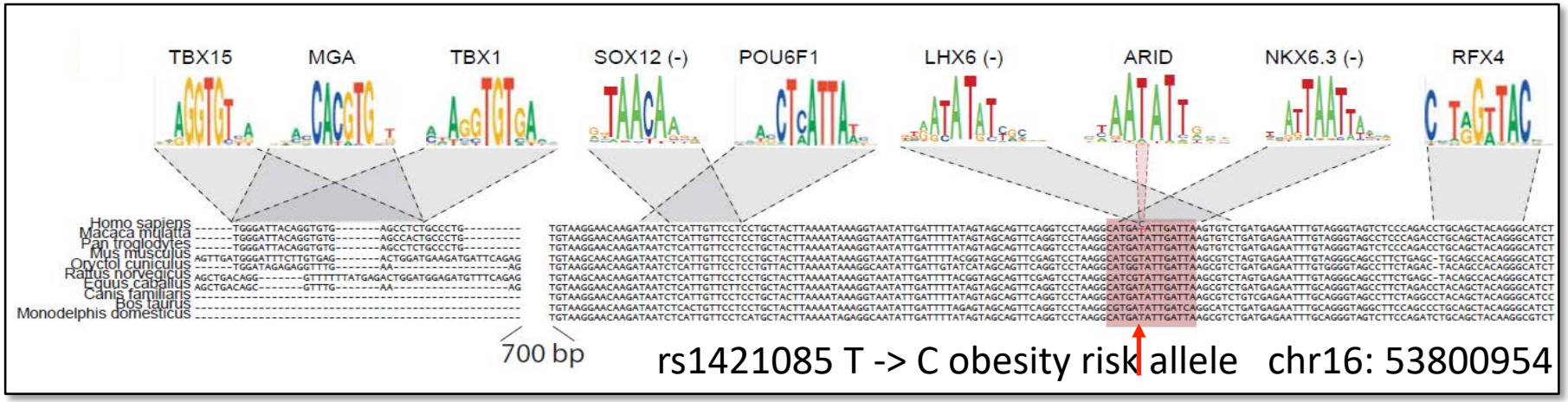
3. Causal SNP: motif enrichment + conservation: rs1421085



Regulatory motifs enriched in BMI GWAS hits

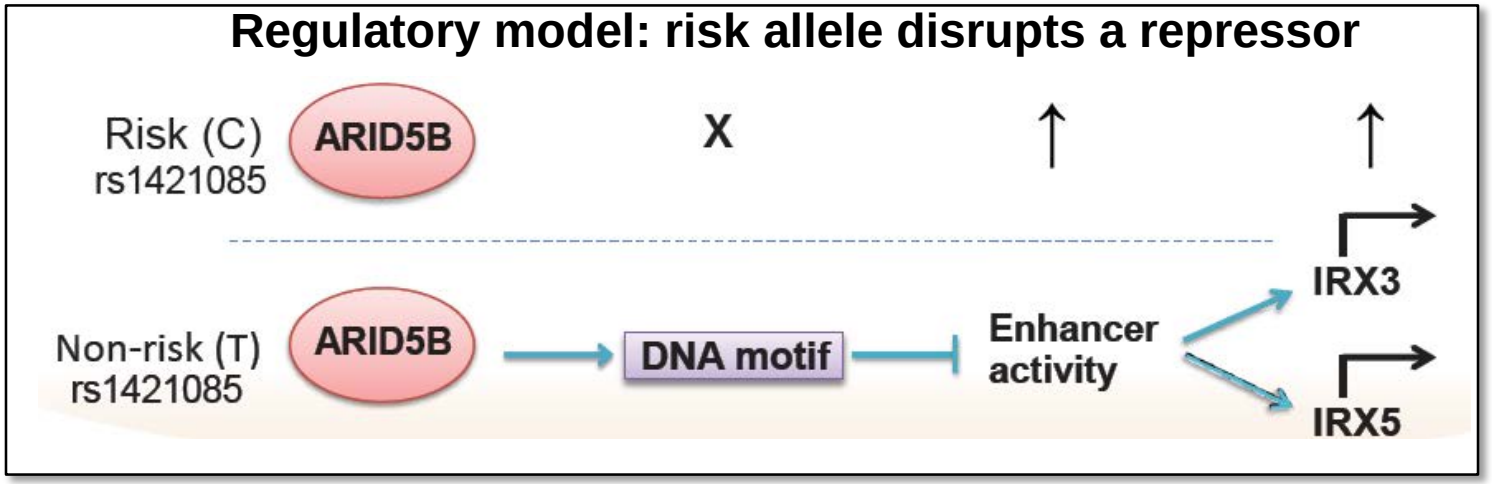


Regulatory motif combinations conserved across mammals



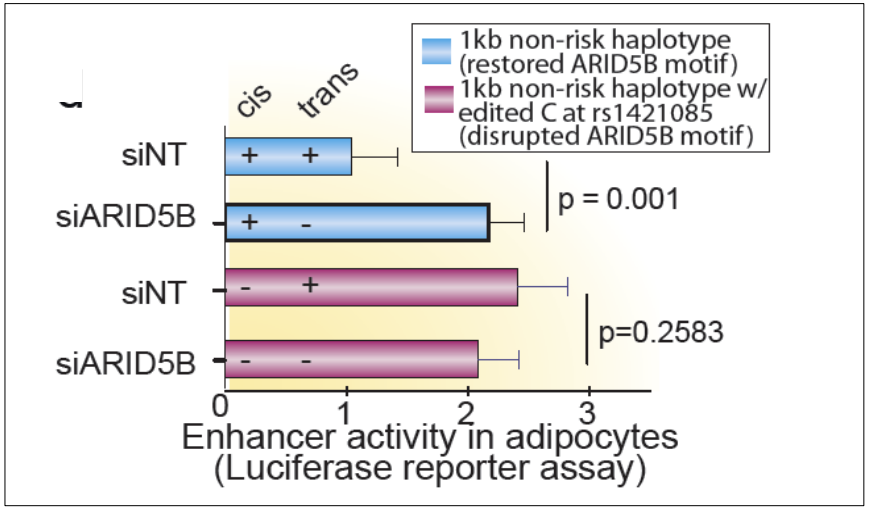
Causal nucleotide rs1421085: risk alters T to C, abolishes AT-rich motif

4. Regulator: Causality and epistasis of ARID5B repressor

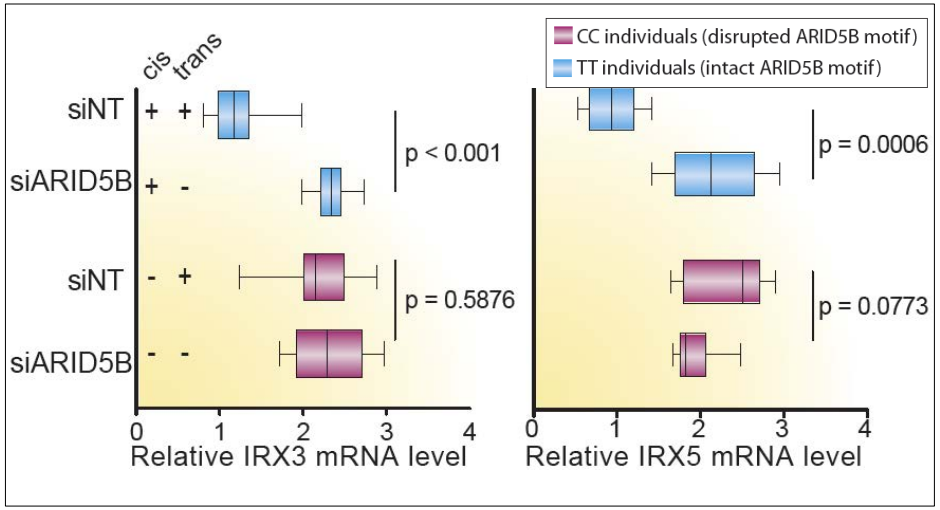


Cis/trans conditional analysis

Enhancer activity

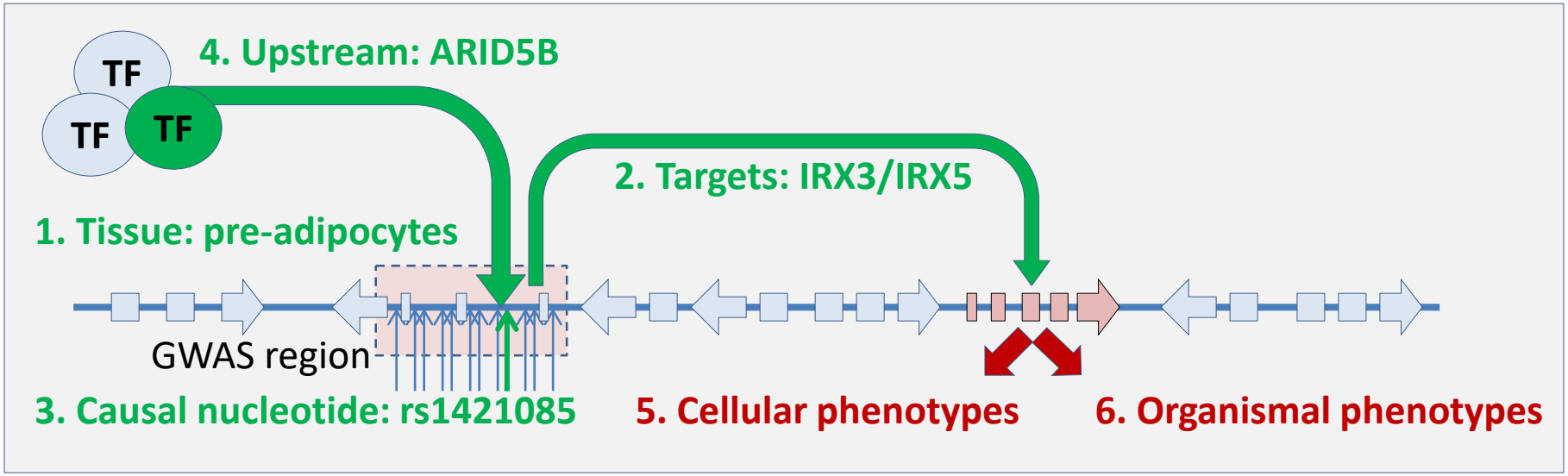


IRX3/5 expression



- ***Repression of enhancer, IRX3 and IRX5 all require both TF and motif***
- ***Disrupting motif (CC), or repressing ARID5B (siRNA) ➔ de-repression***

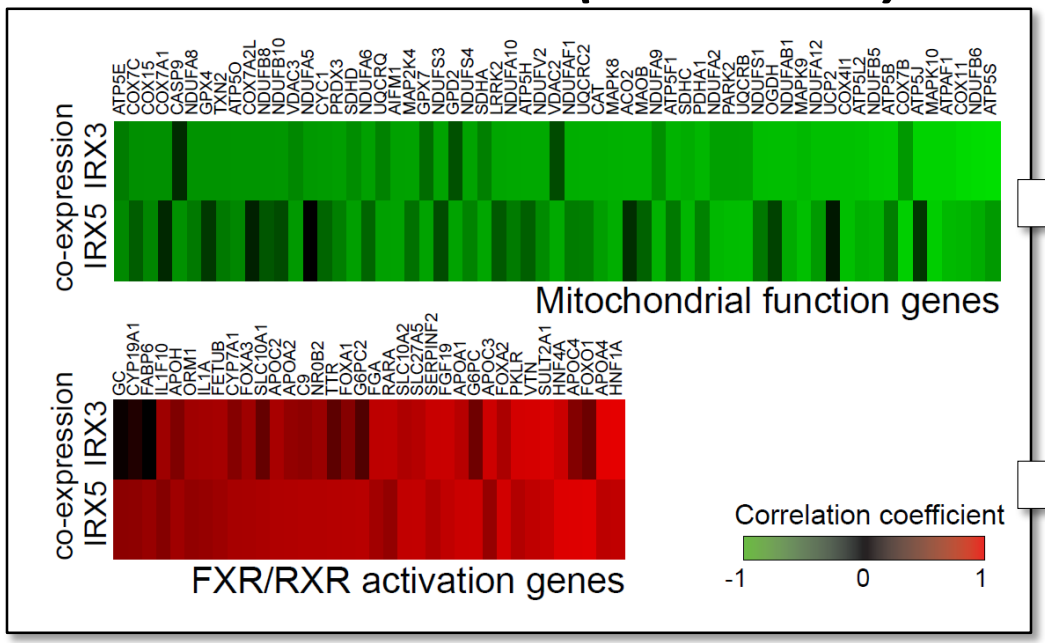
Steps 5-6. Does this circuitry actually lead to obesity?



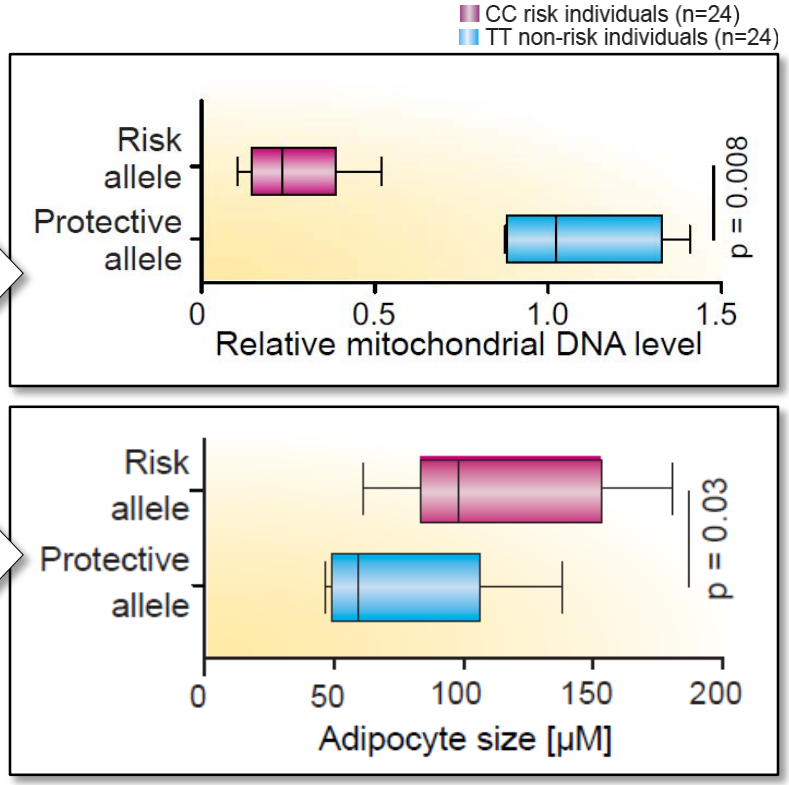
1. Establish relevant **tissue/cell type:** **pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant: **rs1421085**
4. Establish upstream **regulator** causality: **ARID5B**
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

Expression analysis to recognize target processes

Search for genes co-expressed with IRX3 and IRX5 (n=20 indiv.)



Reflected in cellular phenotypes

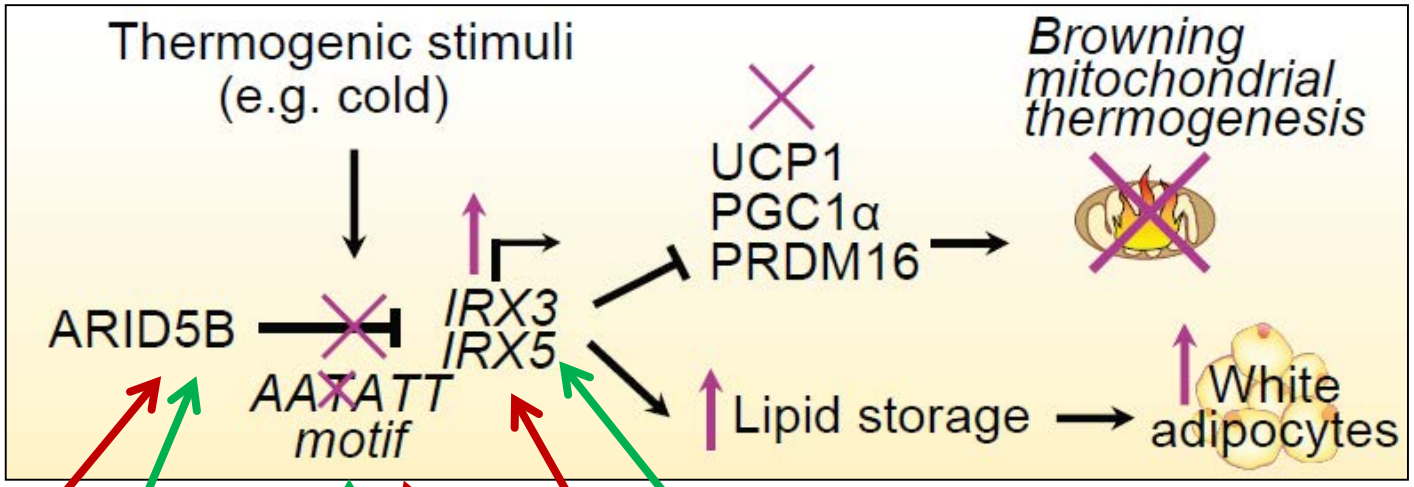


Negative correlation: mitochondria
Positive correlation: lipid storage

Risk carriers: increased mito
Non-risk: increased adipocytes

Risk allele: shift from dissipation to storage

Test model by systematic perturbations



ARID5B KD (obesity)

ARID5B OE (anti-obesity)

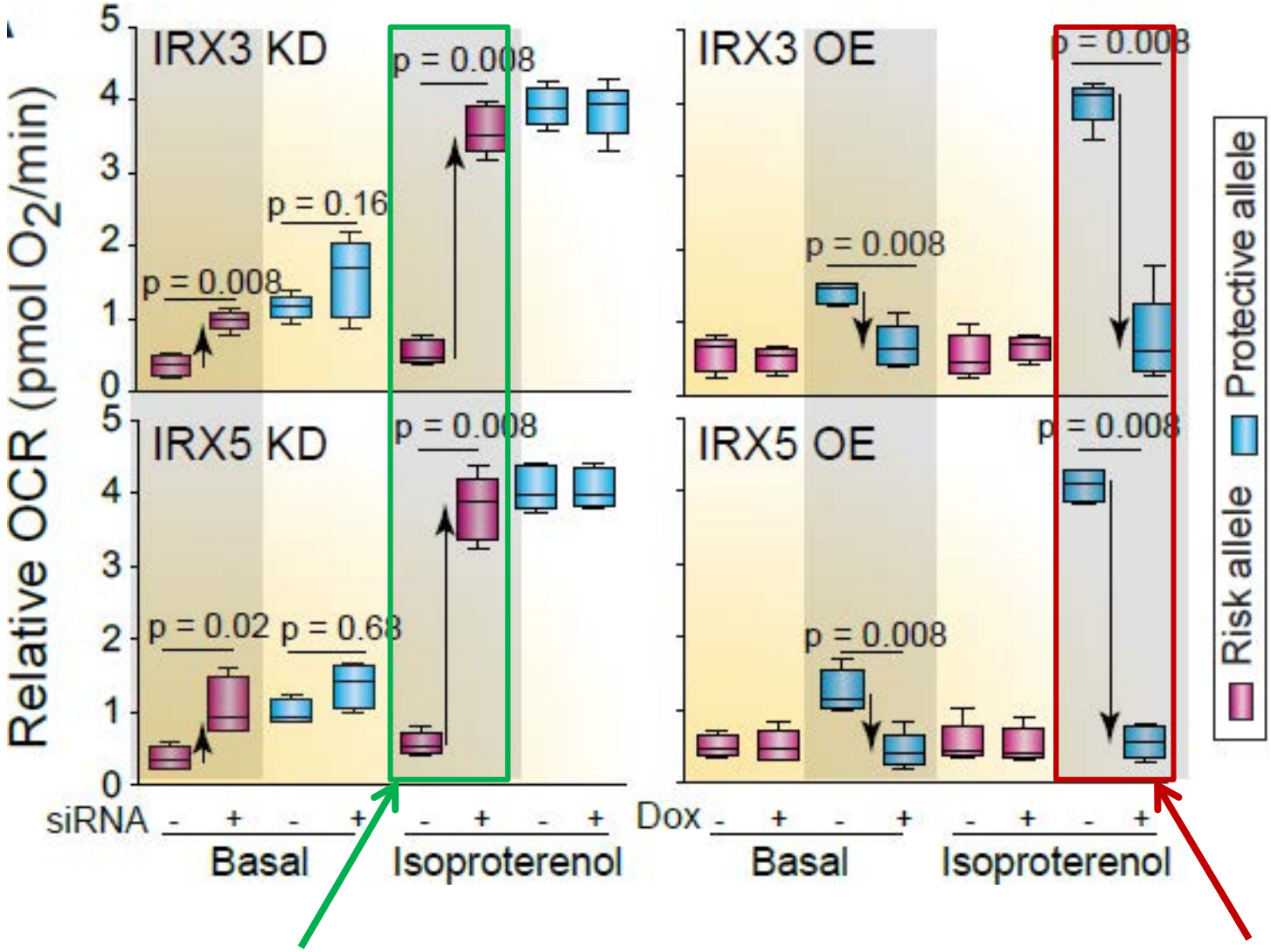
IRX3, IRX5 knock-down ★ (anti-obesity phenotypes)

IRX3, IRX5 overexpression (pro-obesity phenotypes)

★ C-to-T motif rescue (anti-obesity phenotypes)

T-to-C motif disruption (pro-obesity phenotypes)

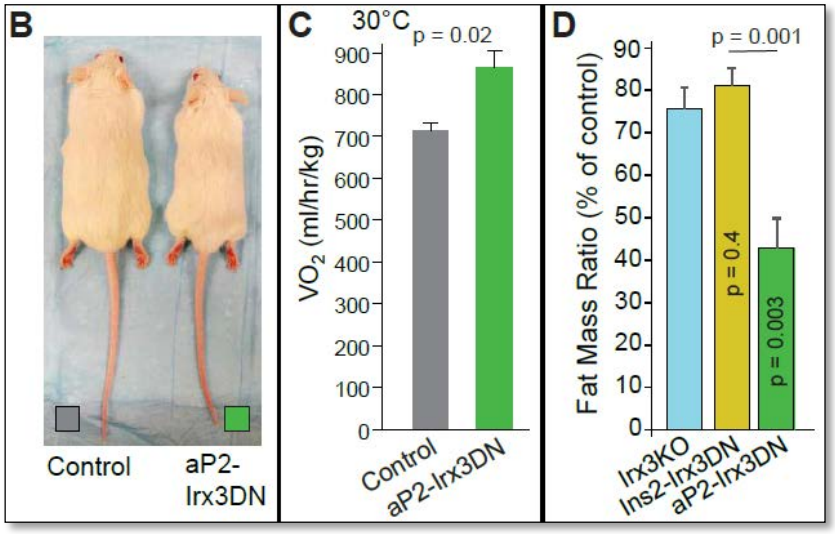
IRX3+IRX5 expression impacts energy utilization



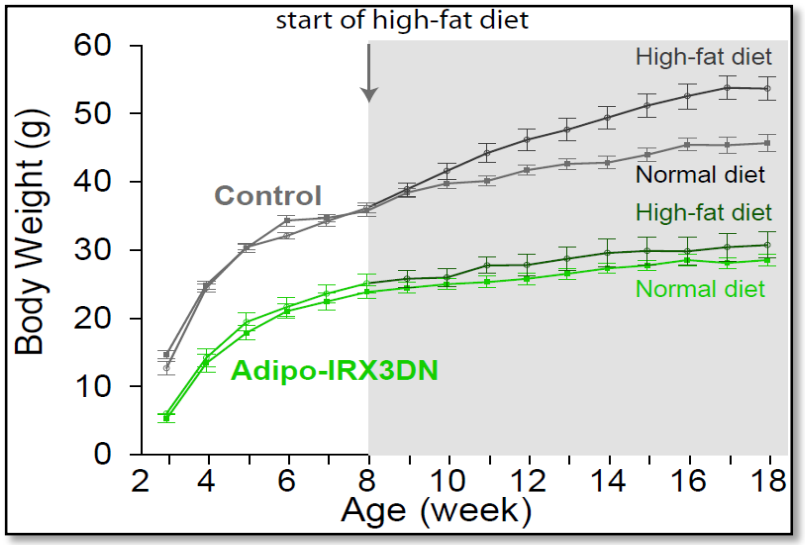
***Risk individuals*: IRX3/5 repression restores respiration,thermogenesis**

***Non-risk*: IRX3/5 overexpression disrupts respiration,thermogenesis**

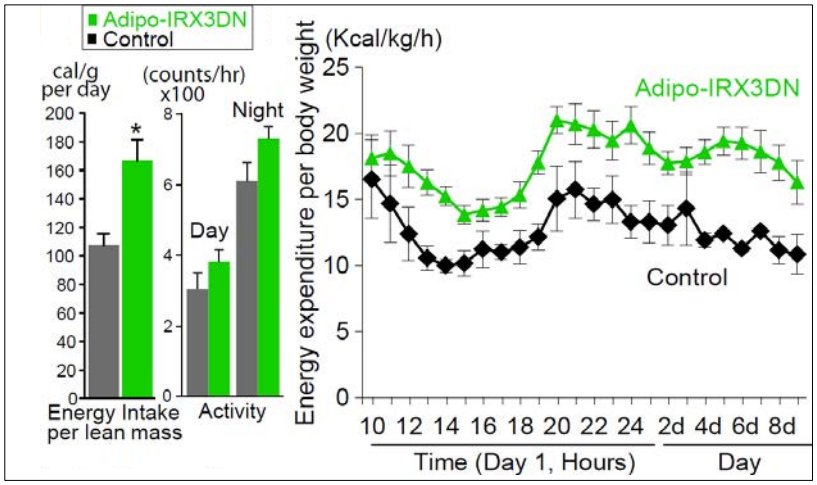
Irx3 adipose repression: anti-obesity phenotypes in mice



54% reduced body weight



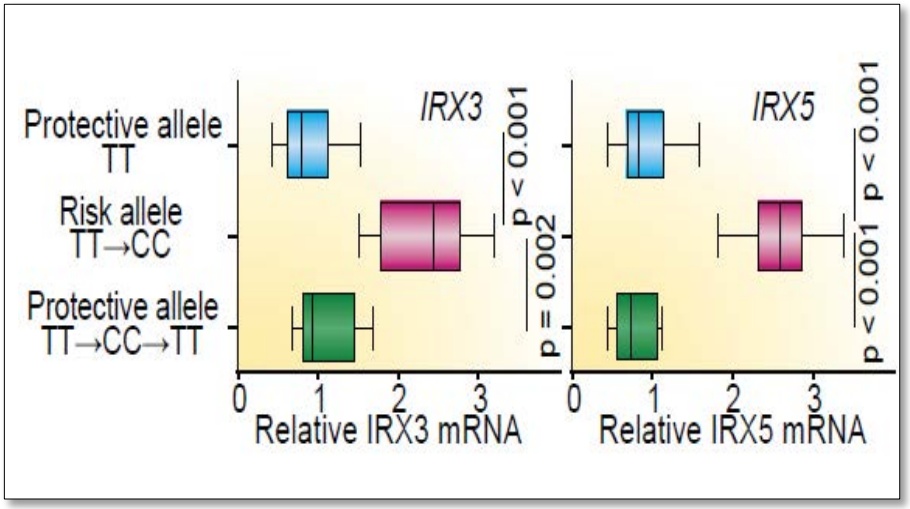
Resistance to high-fat diet



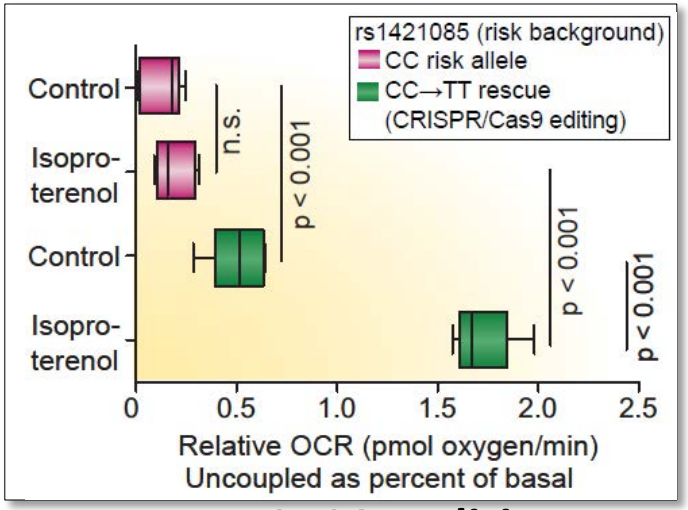
Increased energy dissipation

- *No reduction in appetite*
- *No increase in exercise*
- *In thermoneutral conditions*
- *Day and night (not exercise)*

Single-nucleotide editing reverses thermogenesis in humans



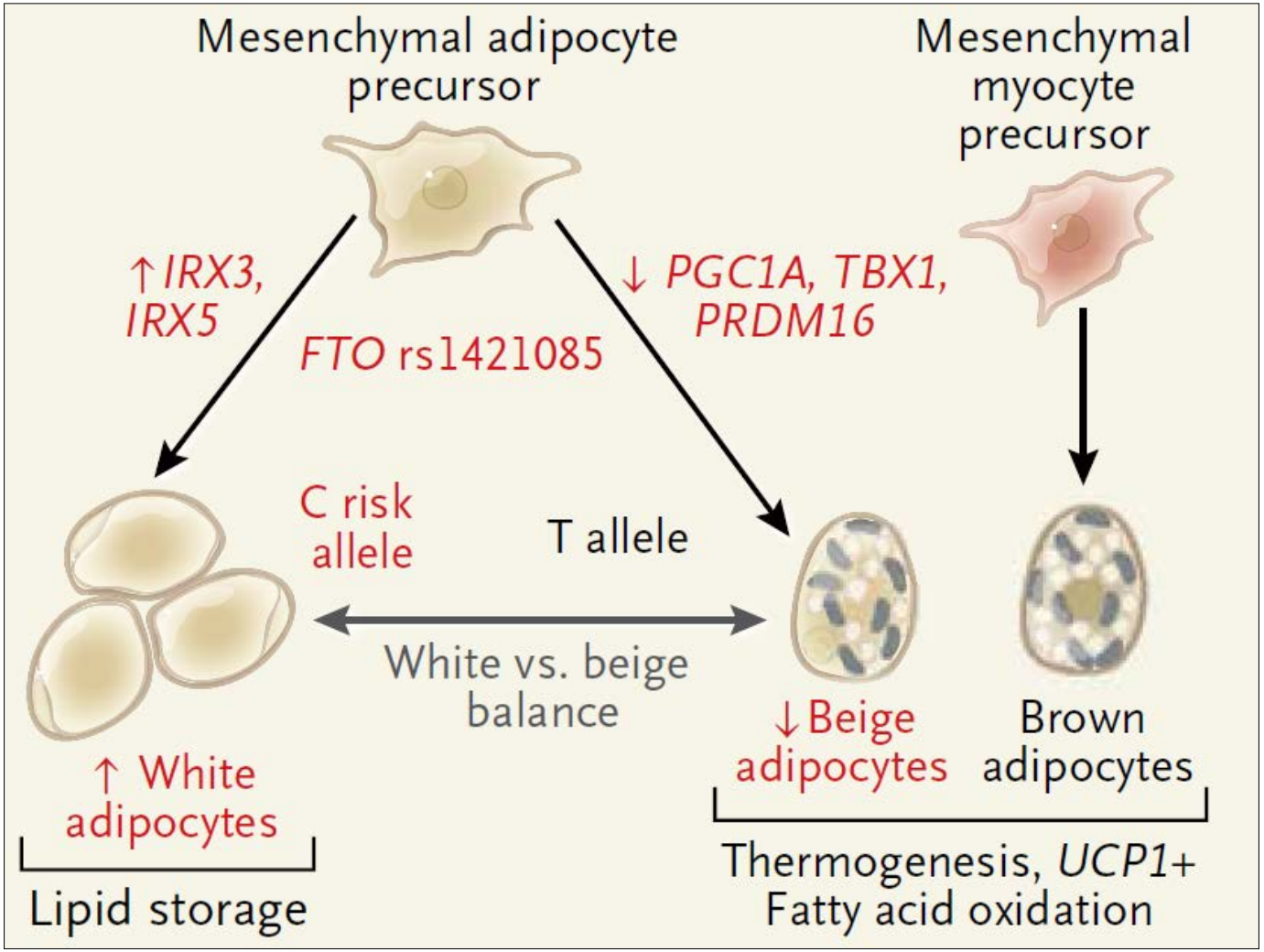
rs1421085 editing alters IRX3+IRX5 expression (500,000 and 1 million nucleotides away!)



rs1421085 editing restores thermogenesis

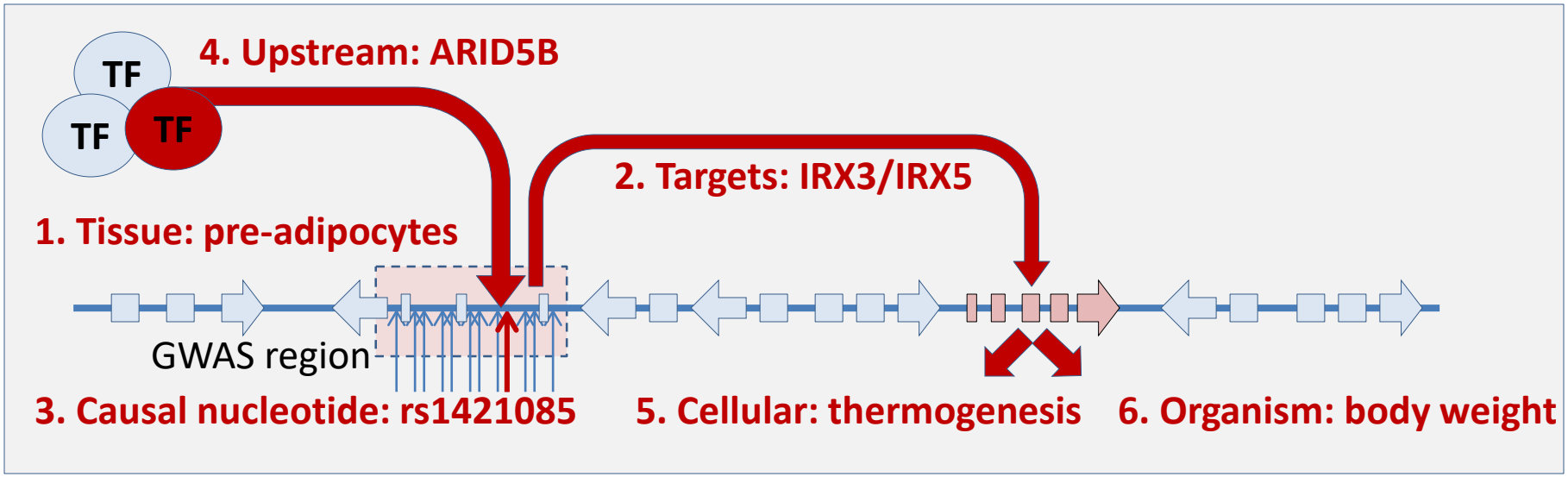
rs1421085 causality: C-to-T editing rescues IRX3/IRX5 expression, ARID5B repression, thermogenesis, developmental expression

Model: beige ⇔ white adipocyte development



Shift therapeutic focus from brain to adipocytes

FTO obesity locus mechanistic dissection



1. Establish relevant **tissue/cell type:** **pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant: **rs1421085**
4. Establish upstream **regulator** causality: **ARID5B**
5. Establish **cellular** phenotypic consequences: **thermogenesis**
6. Establish **organismal** phenotypic consequences: **body weight**

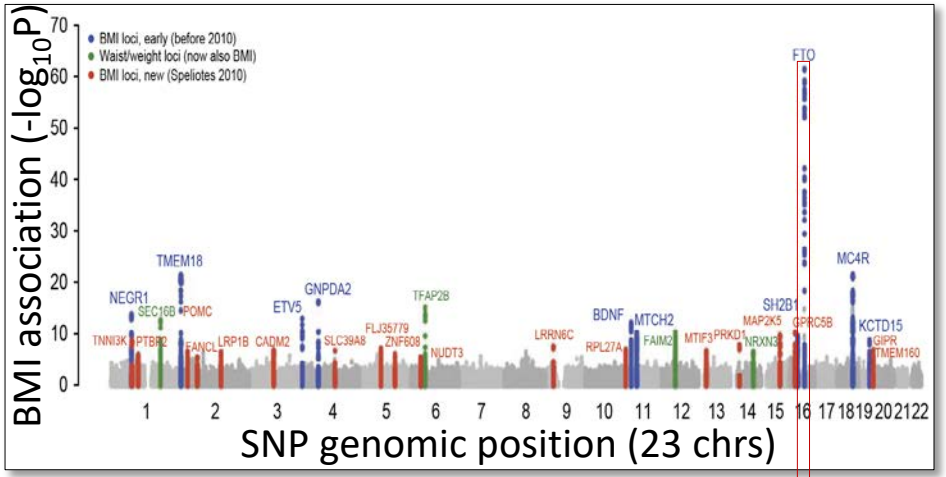
Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting: Common/rare alleles, Linkage vs. GWAS
3. LD, Haplotypes, Co-inheritance, and the challenge of fine-mapping
4. From locus to mechanism - Case study: FTO and Obesity
5. Epigenomics-GWAS integration: ENCODE, Roadmap, EpiMap
6. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA
7. Interpreting non-coding variation: DeepSEA (Jian Zhu guest lecture)

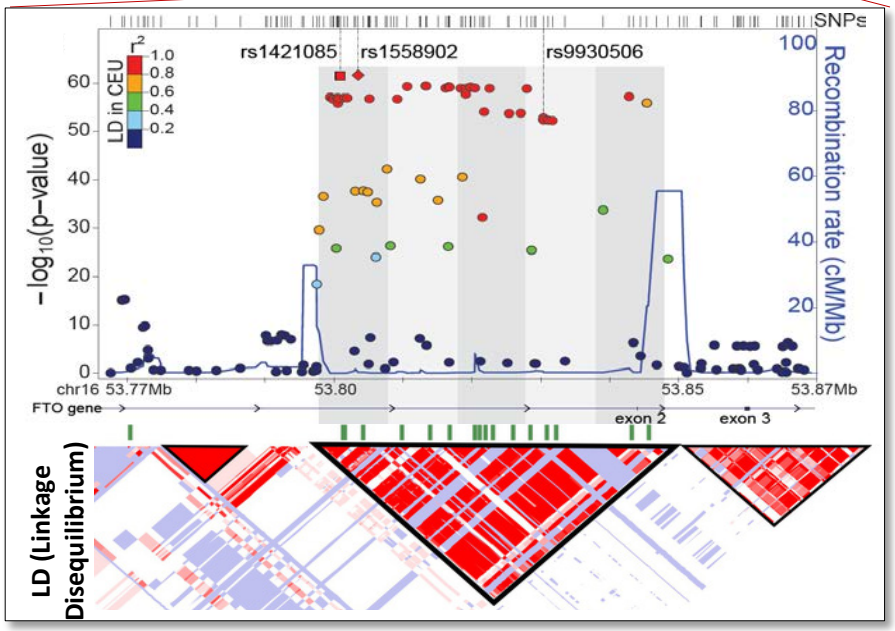
5. Predicting disease-relevant Tissues, Regulators, Cell Types, Target Genes

Genomic medicine today: challenge and promises

GWAS Manhattan Plot: simple χ^2 statistical test



Speliotes NG 2010



Dina NG 2007, Frayling Science 2007, Claussnitzer NEJM 2015

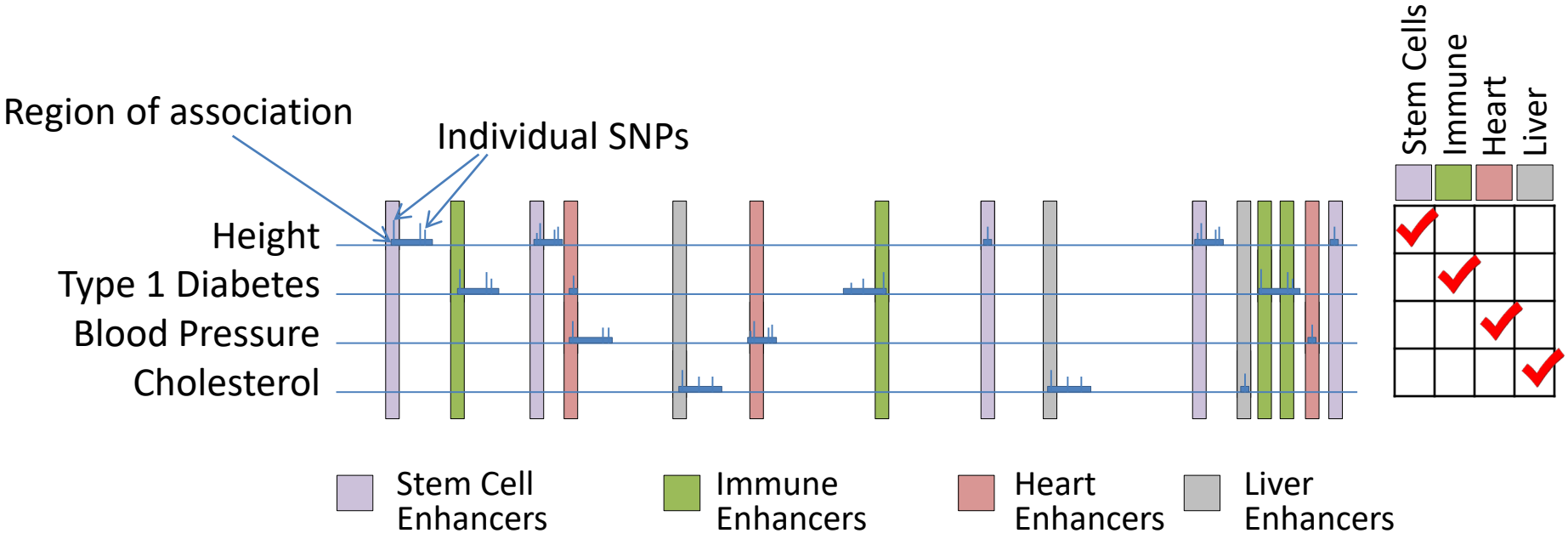
The promise of genetics

- Unbiased, Causal, Uncorrected
- New disease mechanisms
- New target genes
- New therapeutics
- Personalized medicine

The challenge of mechanism

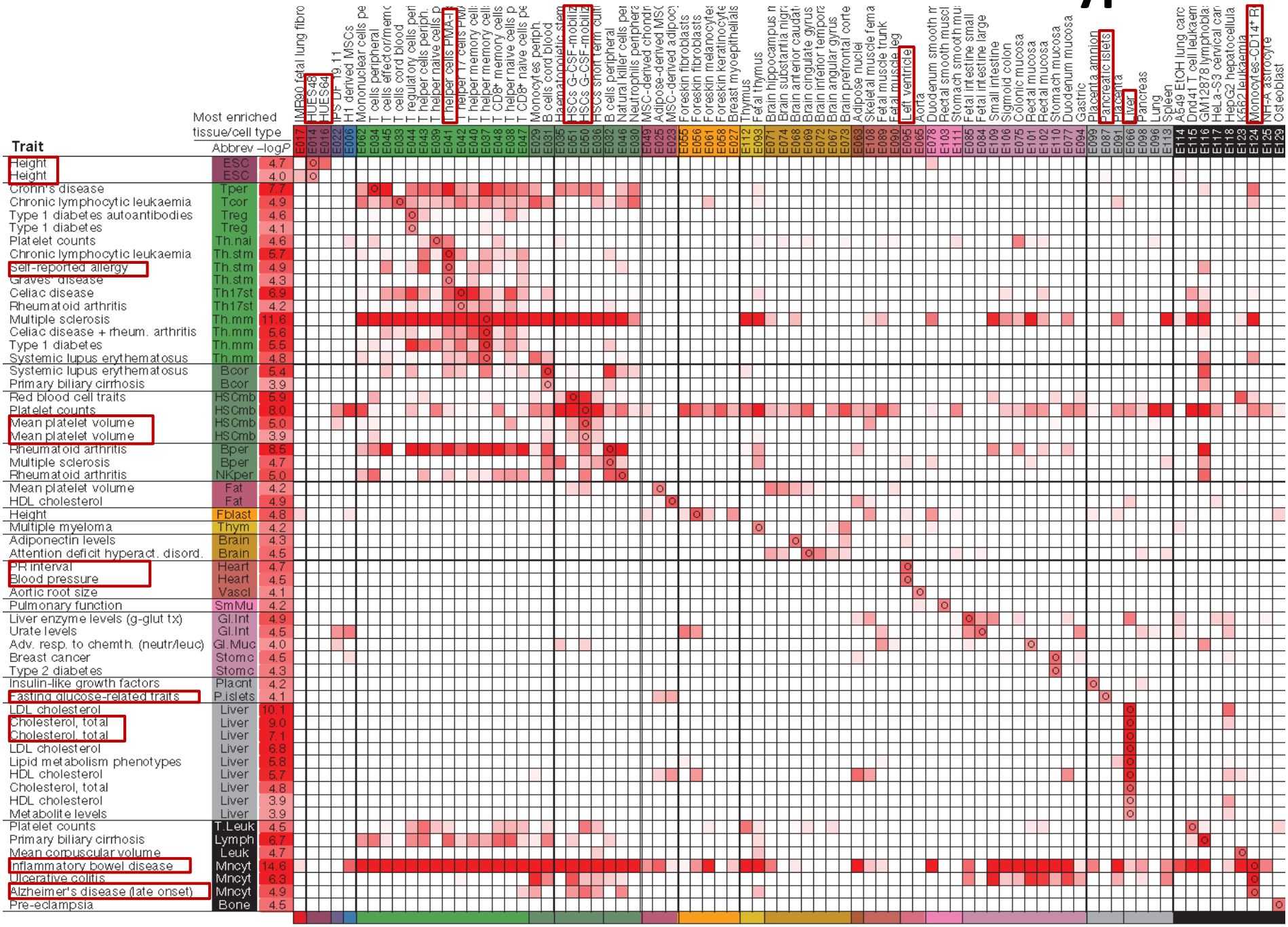
- **90+% disease hits non-coding**
- Target gene not known
- Causal variant not known
- Cell type of action not known
- Relevant pathways not known
- Mechanism not known

Identifying disease-relevant cell types

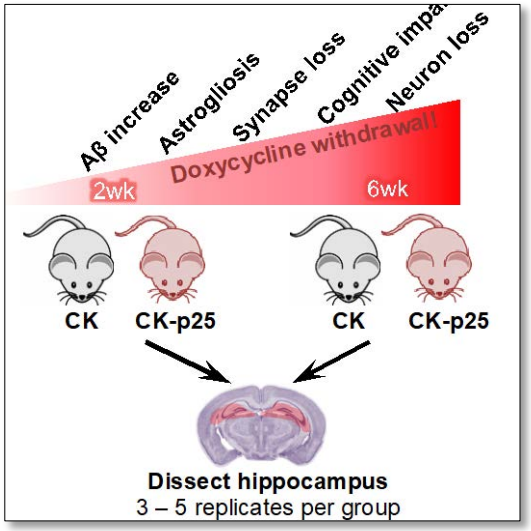


- For every trait in the GWAS catalog:
  - Identify all associated regions at P-value threshold
  - Consider all SNPs in credible interval ($R^2 \geq .8$)
  - Evaluate overlap with tissue-specific enhancers
  - Keep tissues showing significant enrichment ($P<0.001$)
- Repeat for all traits (rows) and all cell types (columns)

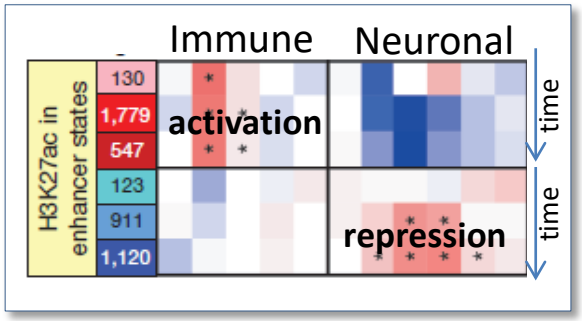
GWAS hits in enhancers of relevant cell types



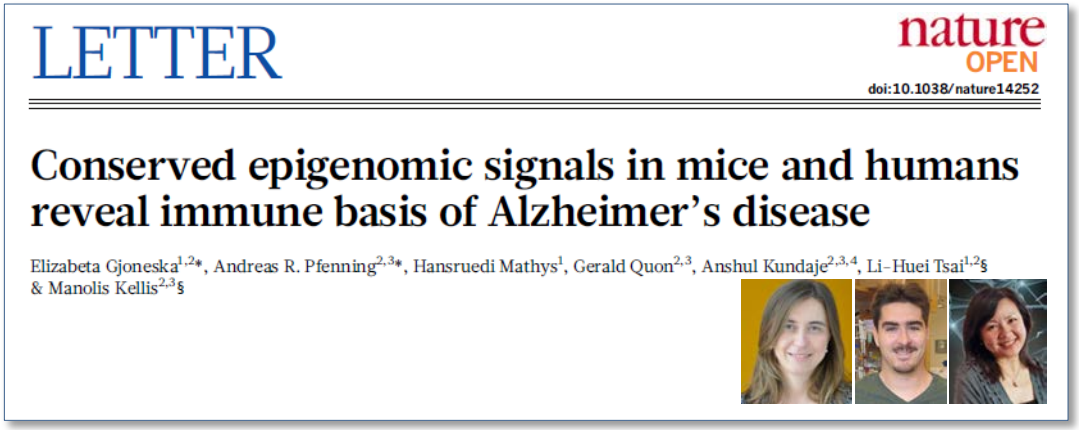
Immune activation + neural repression in human + mouse

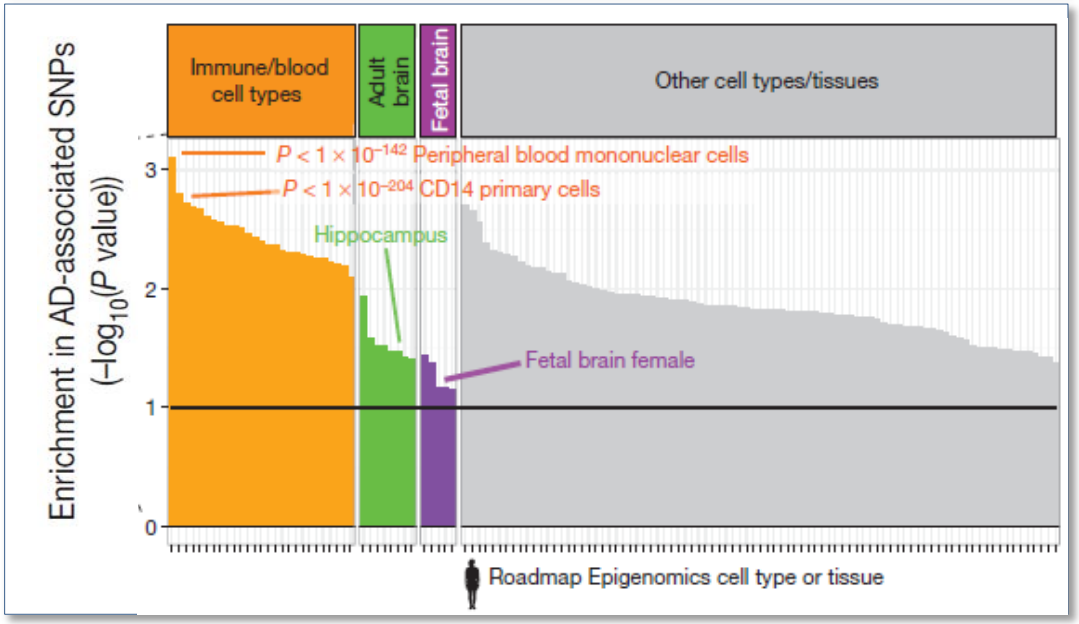


Epigenomics of AD progression



Immune activation precedes neuronal repression

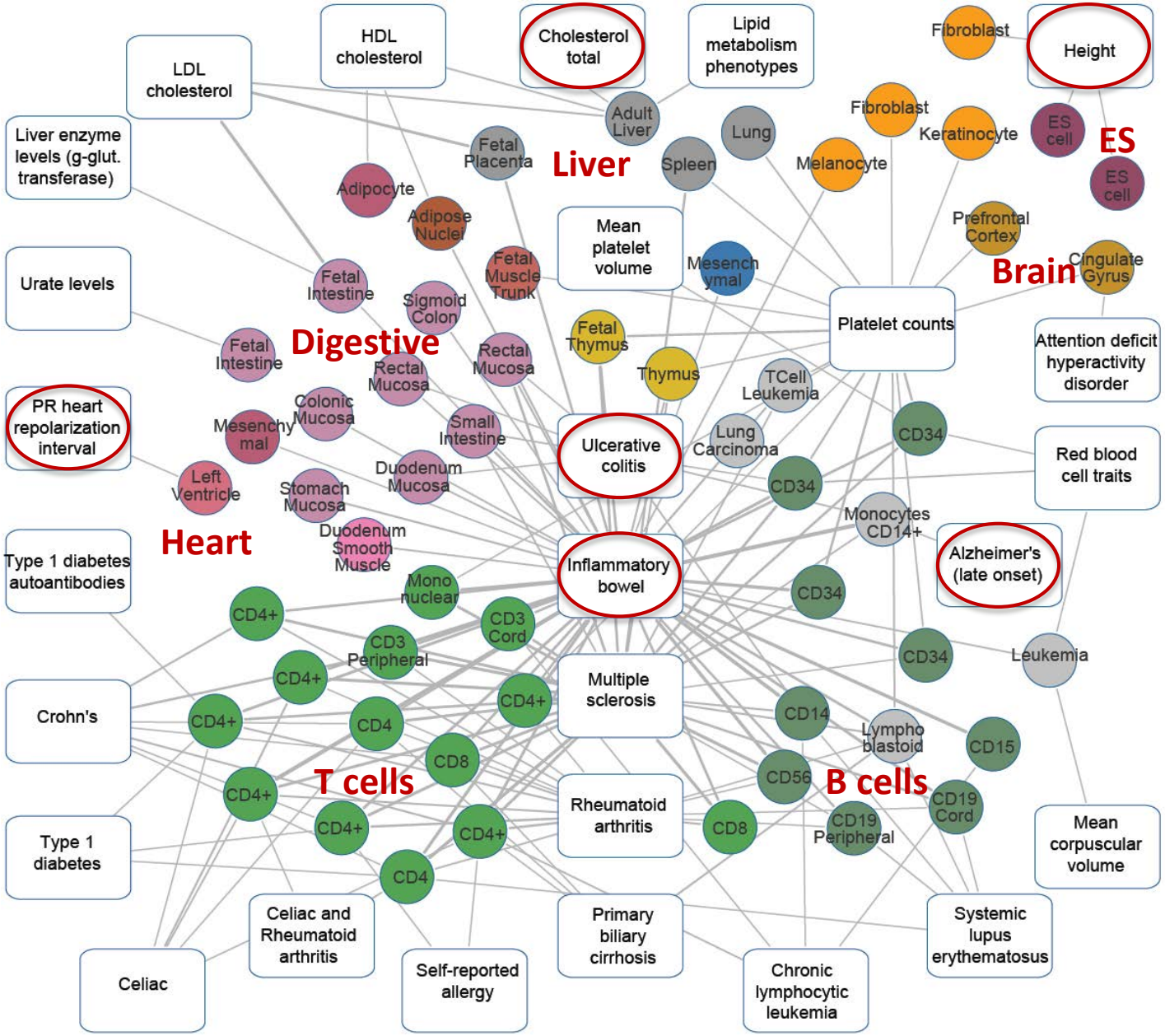




AD variants localize in immune cells, not neuronal

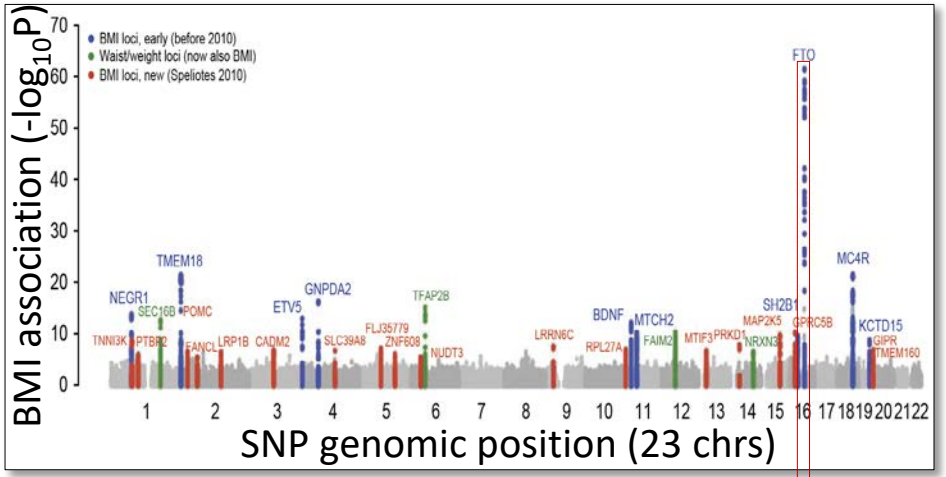
Inflammation as the causal component of Alzheimer's disease

Linking traits to their relevant cell/tissue types

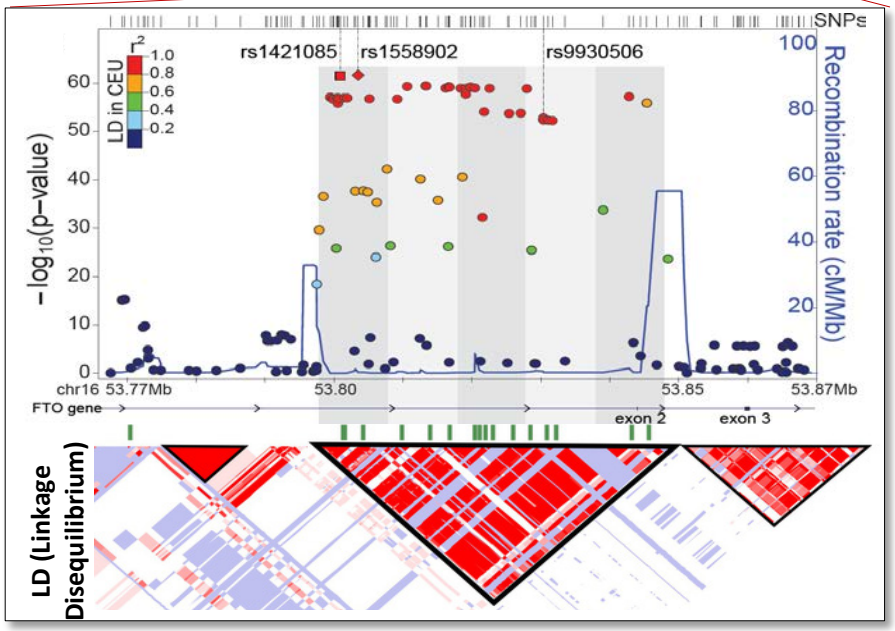


Genomic medicine: challenge and promises

GWAS Manhattan Plot: simple χ^2 statistical test



Speliotes NG 2010



Dina NG 2007, Frayling Science 2007, Claussnitzer NEJM 2015

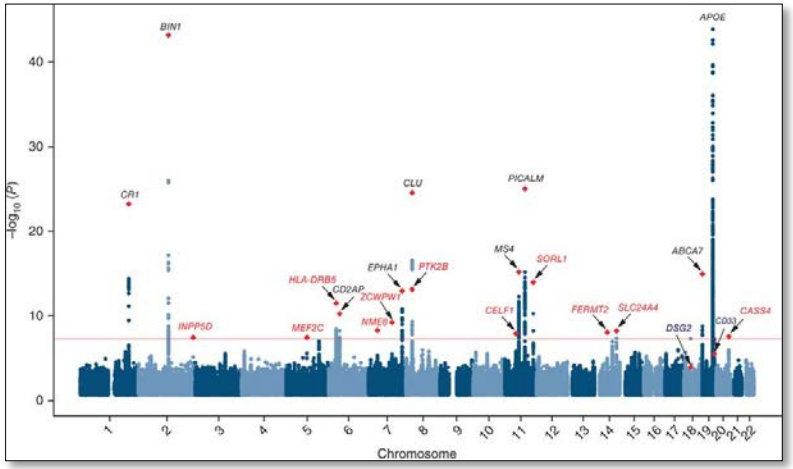
The promise of genetics

- Disease mechanism
- New target genes
- New therapeutics
- Personalized medicine

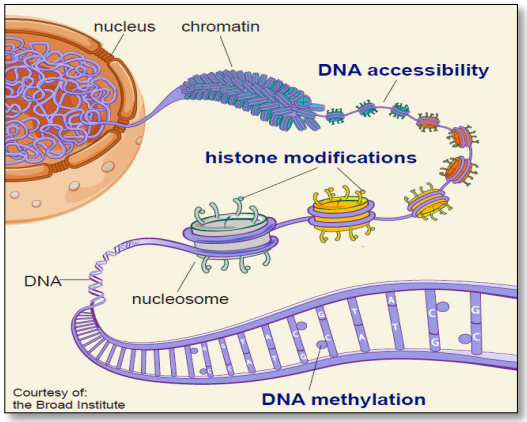
The challenge of mechanism

- **90+% disease hits non-coding**
- Target gene not known
- Causal variant not known
- Cell type of action not known
- Relevant pathways not known
- Mechanism not known

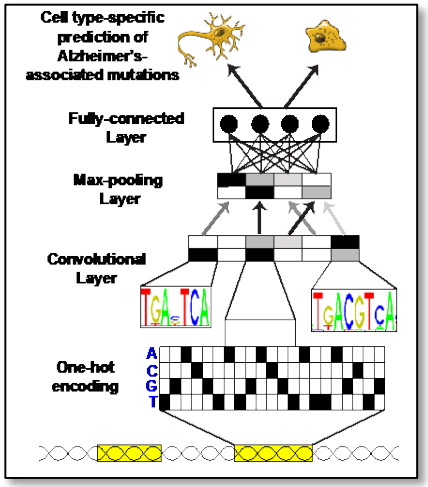
Summary: Dissect circuitry of disease-associated regions



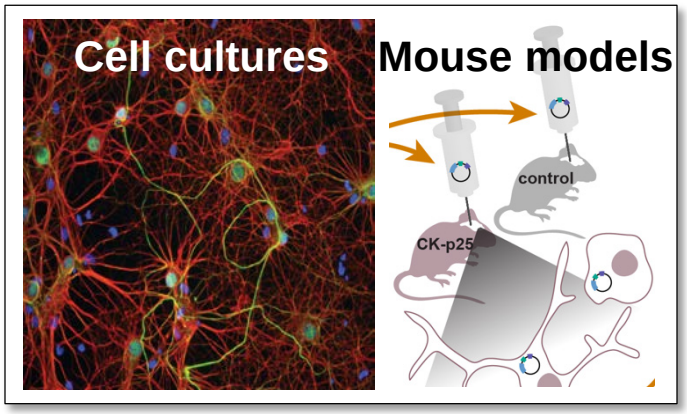
1. Disease genetics reveals common + rare variants/regions



2. Profile RNA + Epigenome in healthy + disease samples



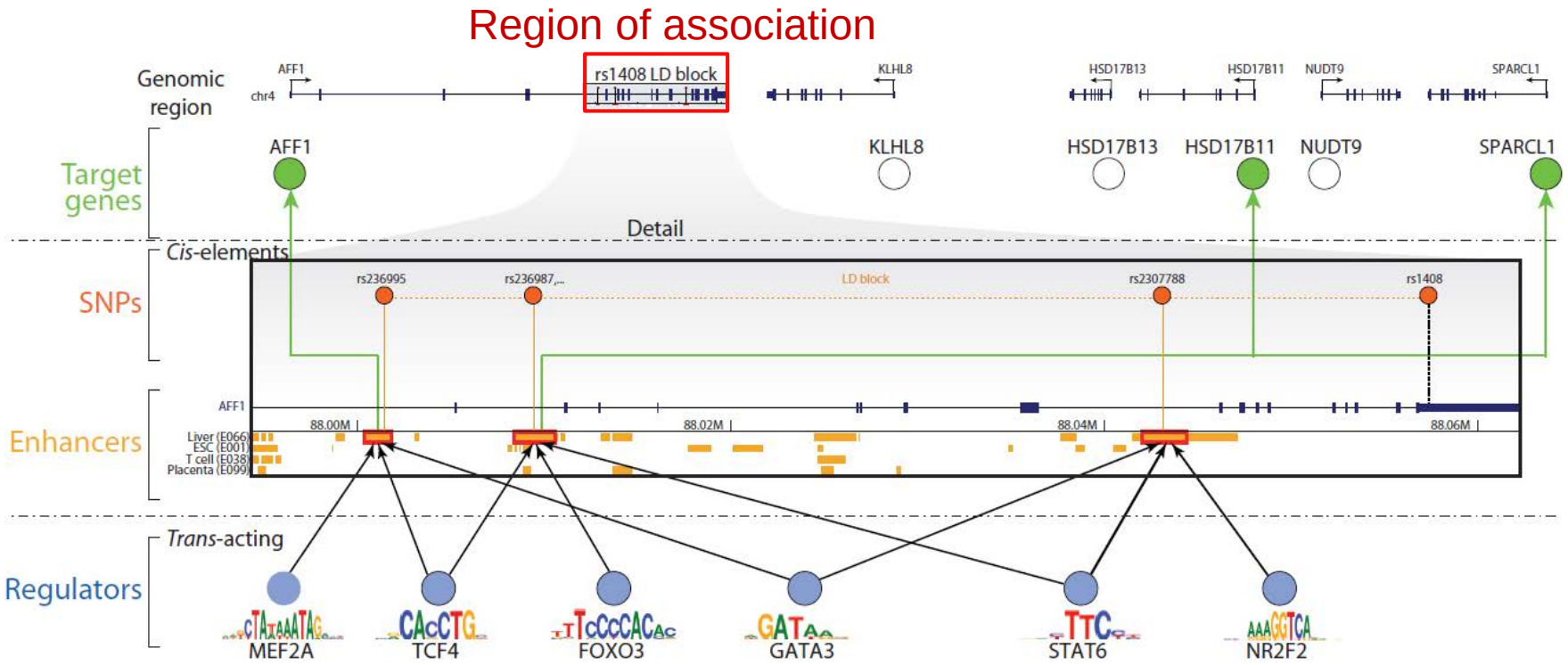
3. Integrate data to predict driver genes, regions, cell types



4. Validate predictions in human cells + mouse models

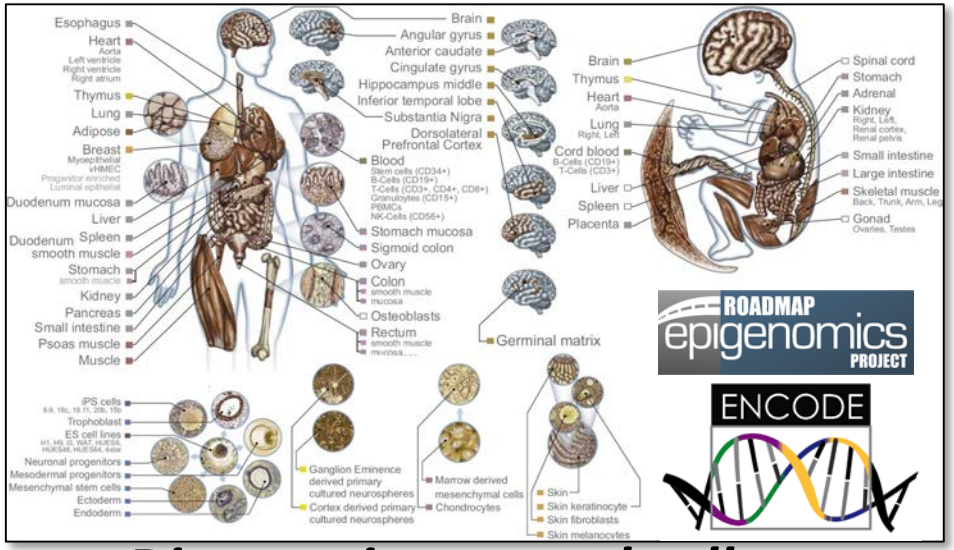
5. Disseminate results

Regulatory circuitry of GWAS loci



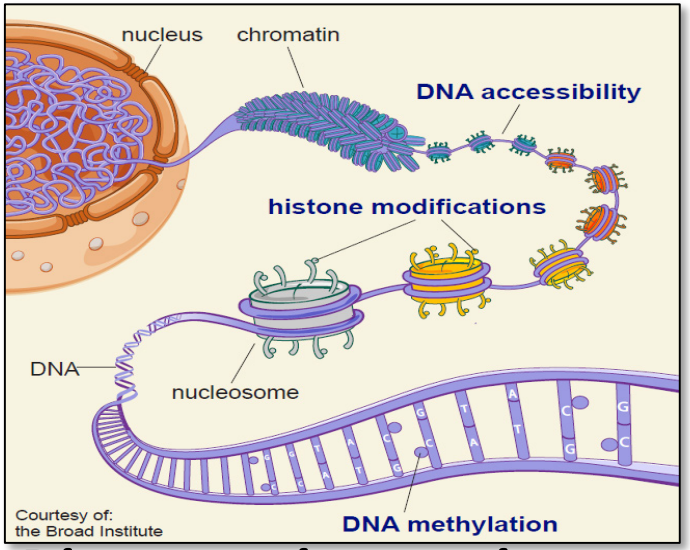
- Expand each GWAS locus using SNP linkage disequilibrium (LD)
  - Recognize **relevant cell types**: tissue-specific enhancer enrichment
  - Recognize **driver TFs**: enriched motifs in multiple GWAS loci
  - Recognize **target genes**: linked to causal enhancers

Epigenomic mapping across 800+ tissues/cell types

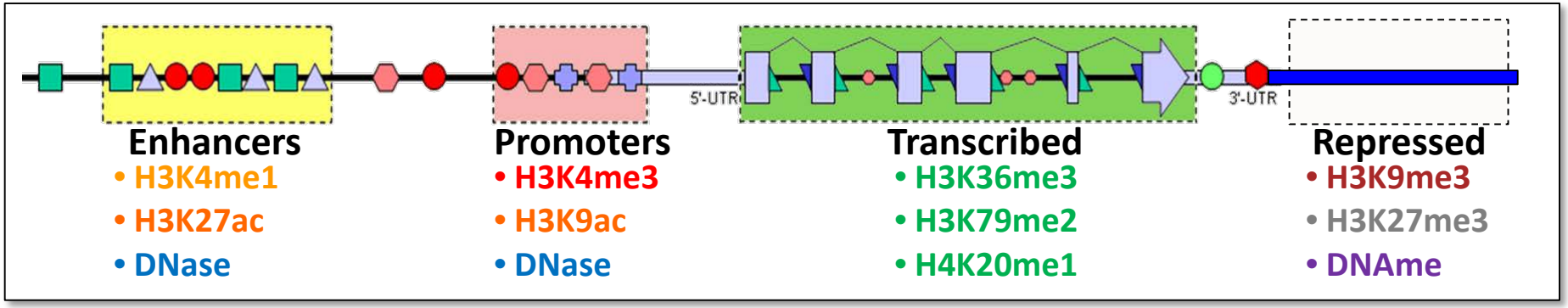


Diverse tissues and cells

X

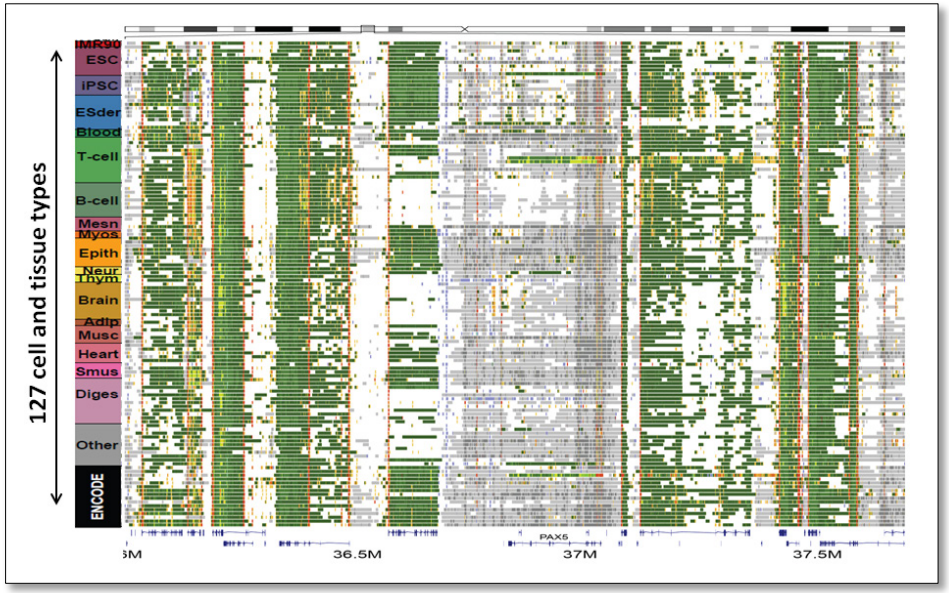


Diverse epigenomic assays

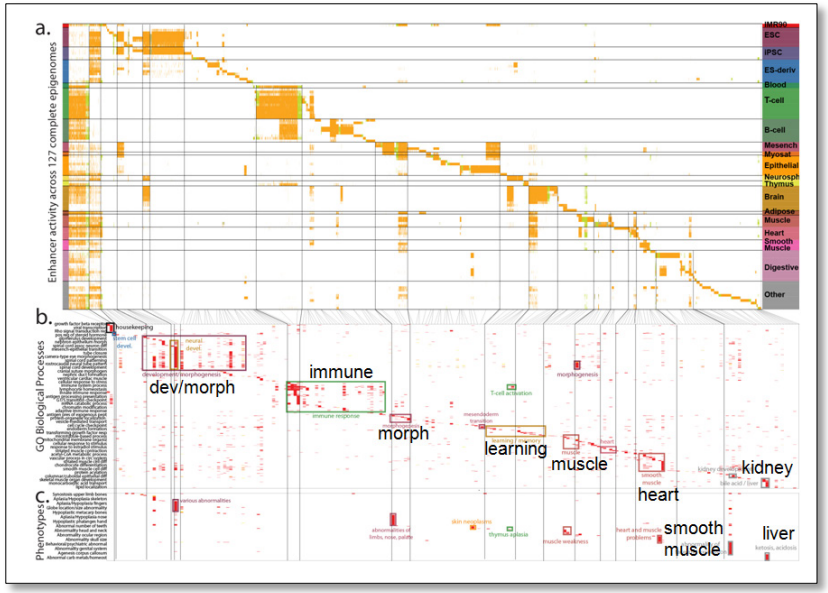


*Their **combinations** define diverse classes of elements*

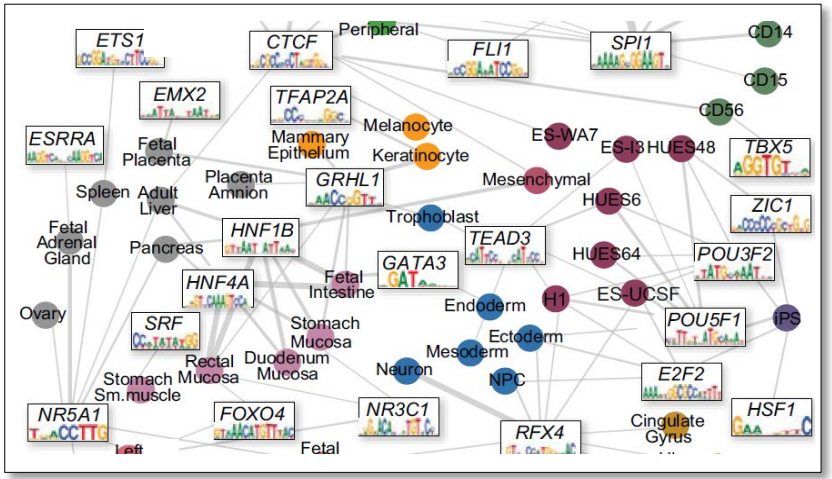
Enhancer modules, regulators, and target genes



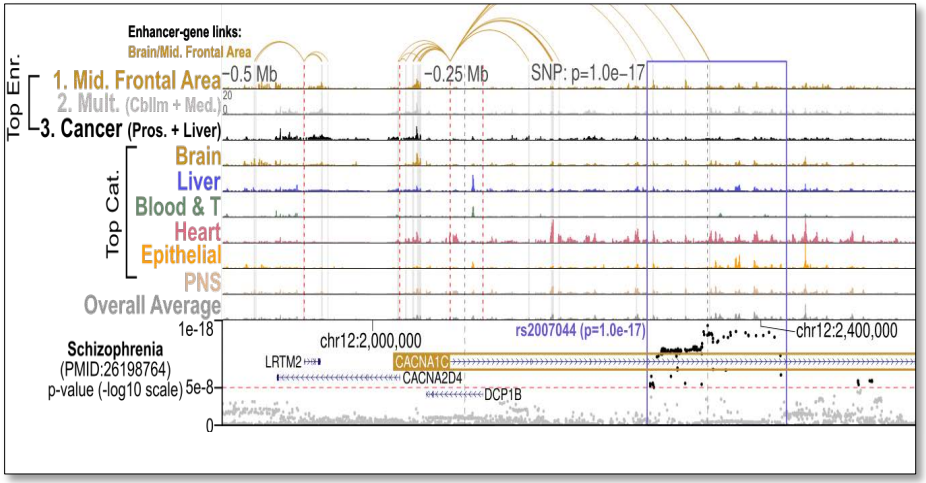
1. *Map chromatin states across 127 tissue/cells*



2. *Group enhancers into modules of common function*



3. *Predict module regulators using motif enrichment*

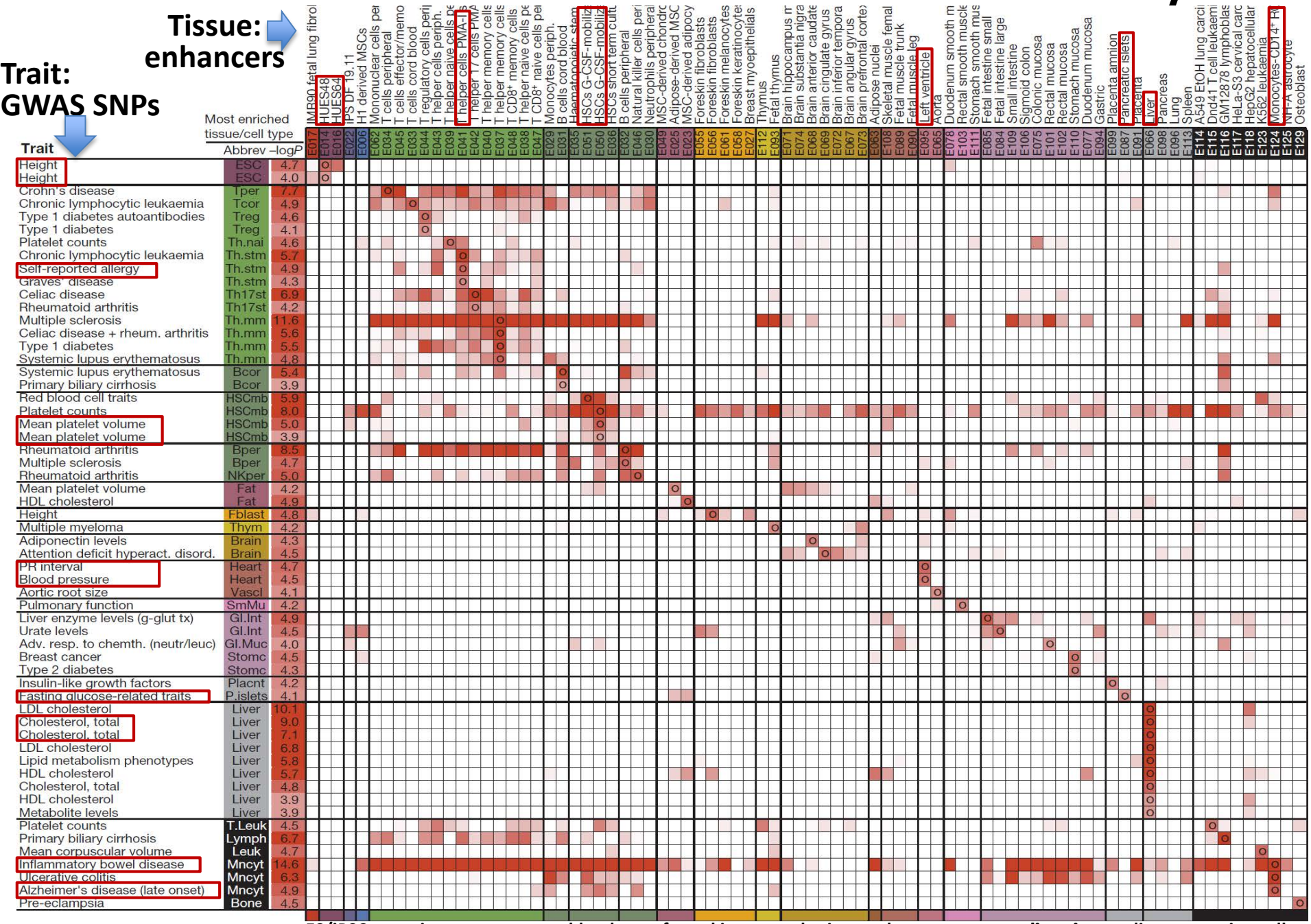


4. *Predict target genes using activity correlation, Hi-C, eQTLs*

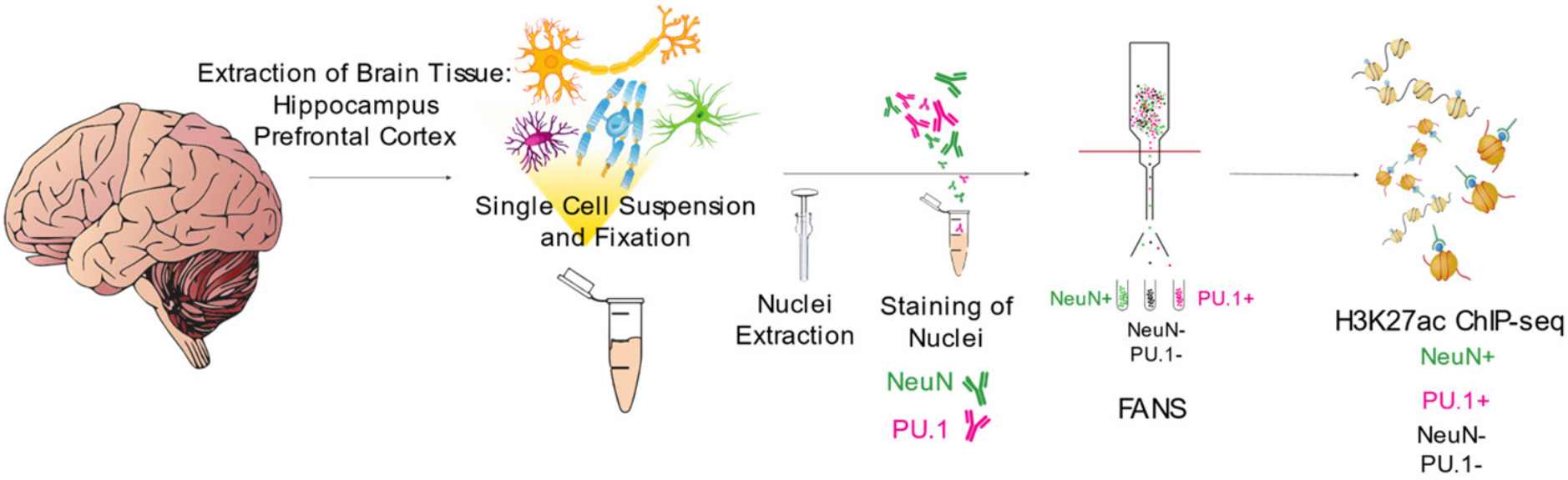
Enhancer enrichment reveals trait-relevant tissues/cells

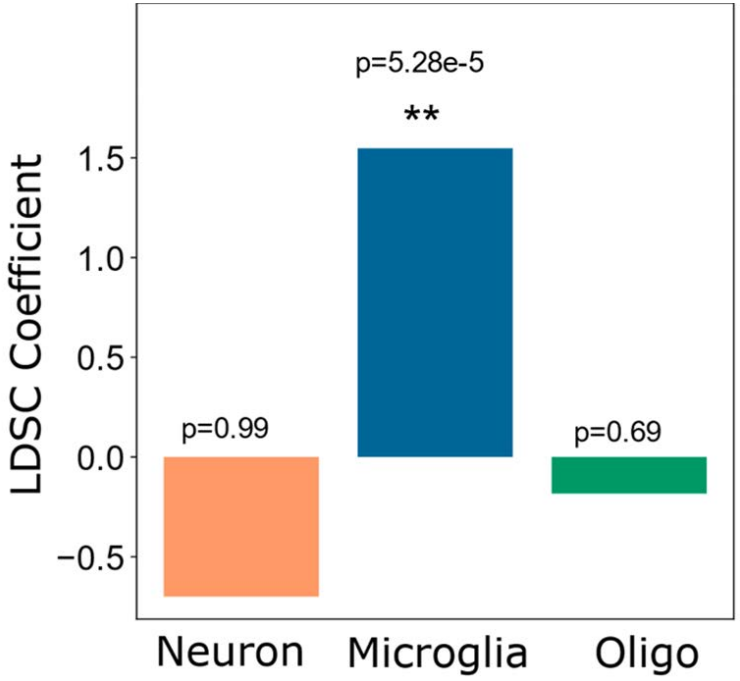
Tissue: enhancers

Trait: GWAS SNPs



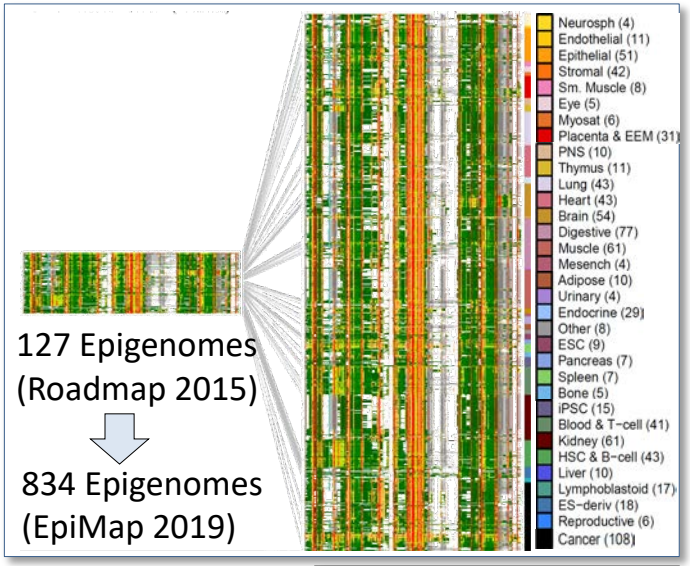
Cell-sorted H3K27ac ➔ AD variants in microglia, not neurons





- **No enrichment found in whole-brain samples**
- **Cell-sorted H3K27ac shows strong enrichment for AD variants in microglia**
- **No enrichment found in neurons or oligodendrocyte H3K27ac for AD variants**

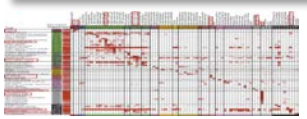
EpiMap: 834 tissue/cell types ➔ 30k GWAS SNPs in 534 traits



127 Epigenomes (Roadmap 2015)

➔

834 Epigenomes (EpiMap 2019)

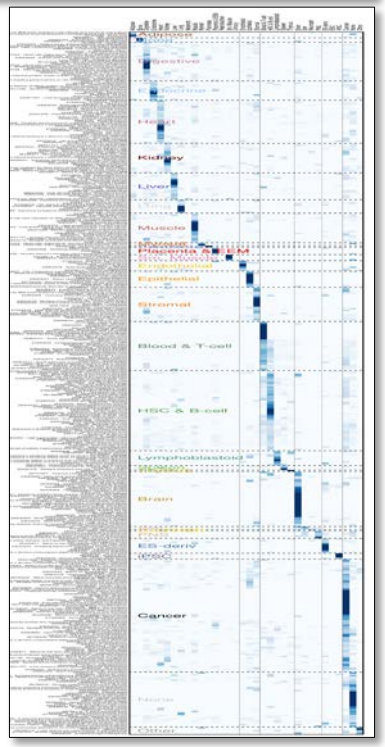


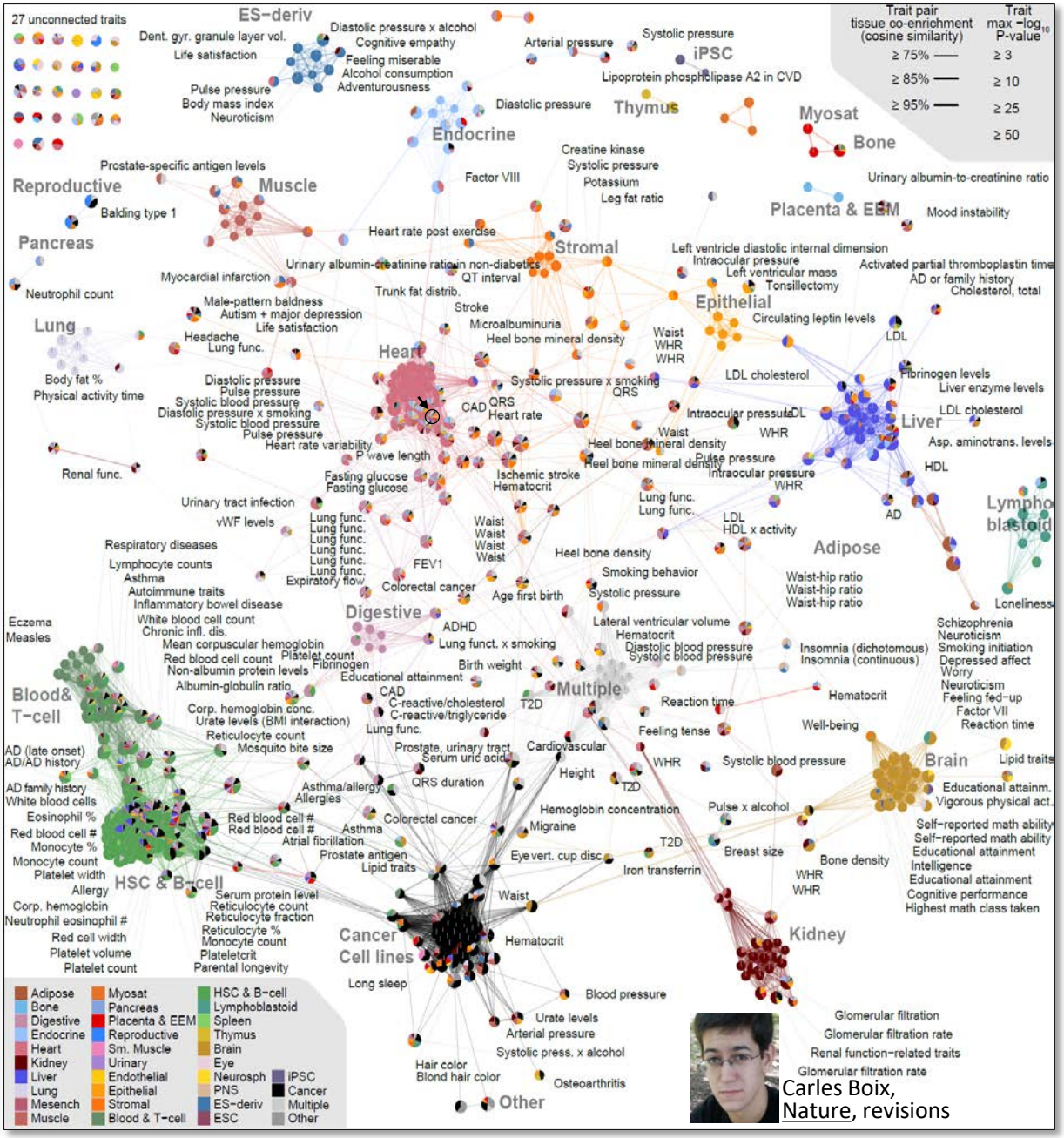
54 enriched GWAS traits (2015)

➔

534 enriched traits

30,247 SNPs in enriched enhancers ➔ Highly-specific associations Emerge ➔ Precise biological hypotheses on mechanistic basis



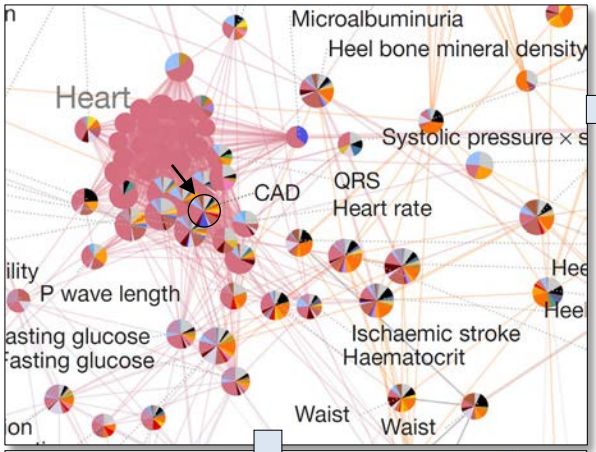


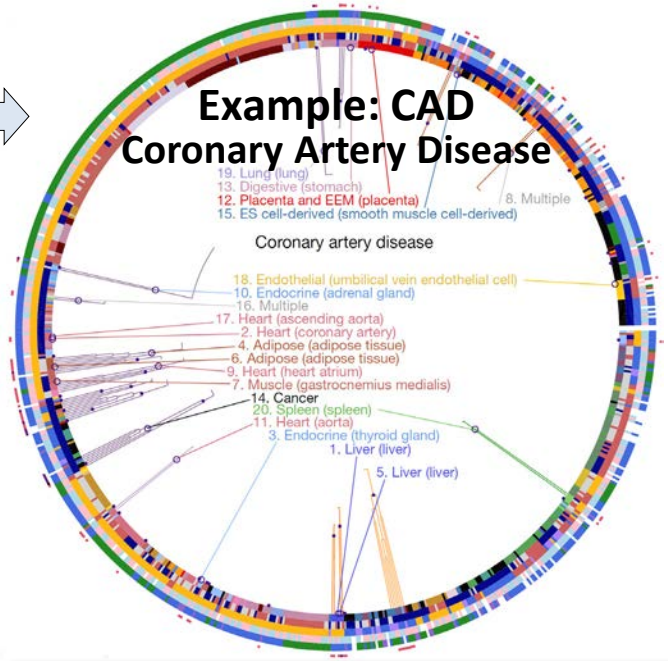
Carles Boix, Nature, revisions

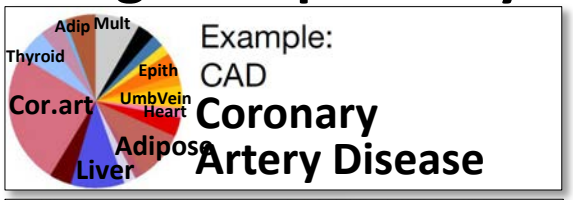
➔ http://compbio.mit.edu/epimap

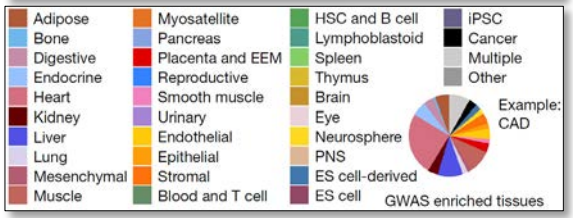
Tissue enrich/co-enrichments ➔ trait clustering, trait-tissue network

Dissect circuitry of 30,000 GWAS loci: TF➔Enh➔SNP➔gene➔pathways

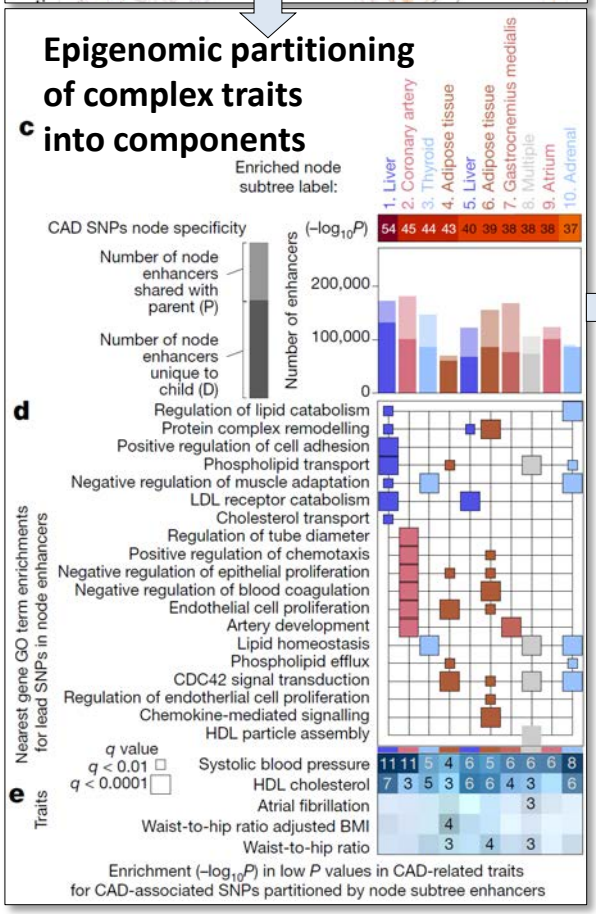


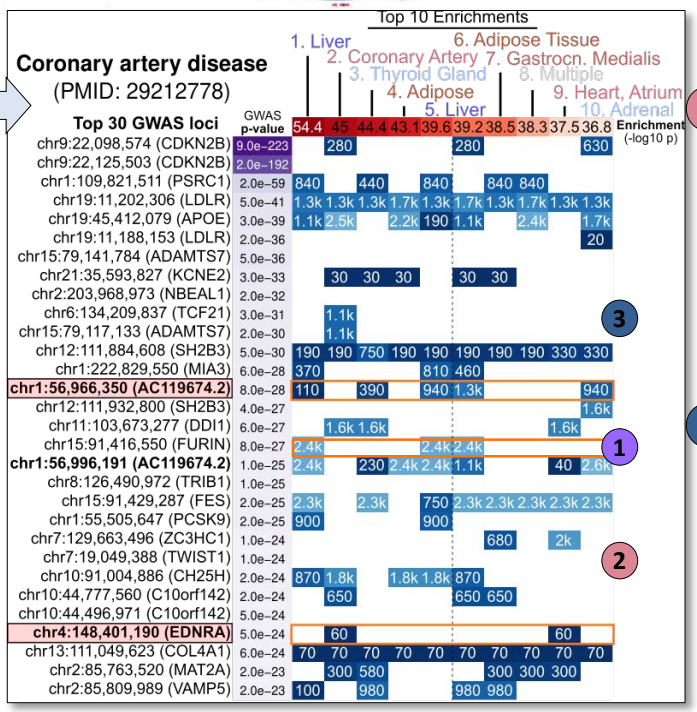




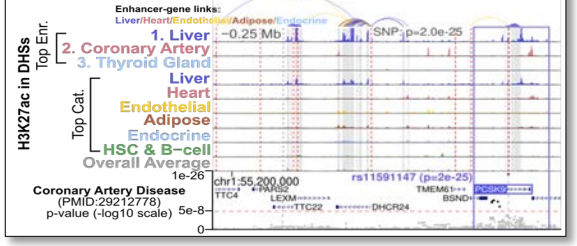


Epigenomic partitioning of complex traits into components

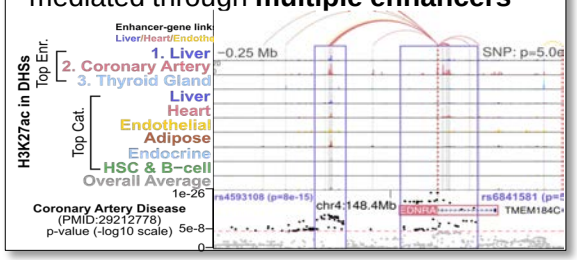




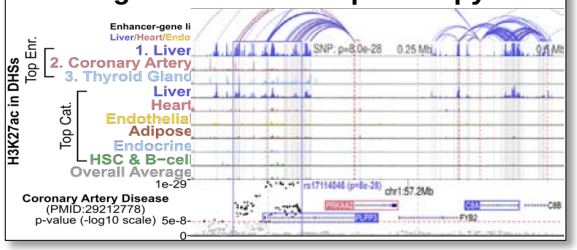
1 **PCSK9: Liver-only** mechanism, mediated through primarily one variant



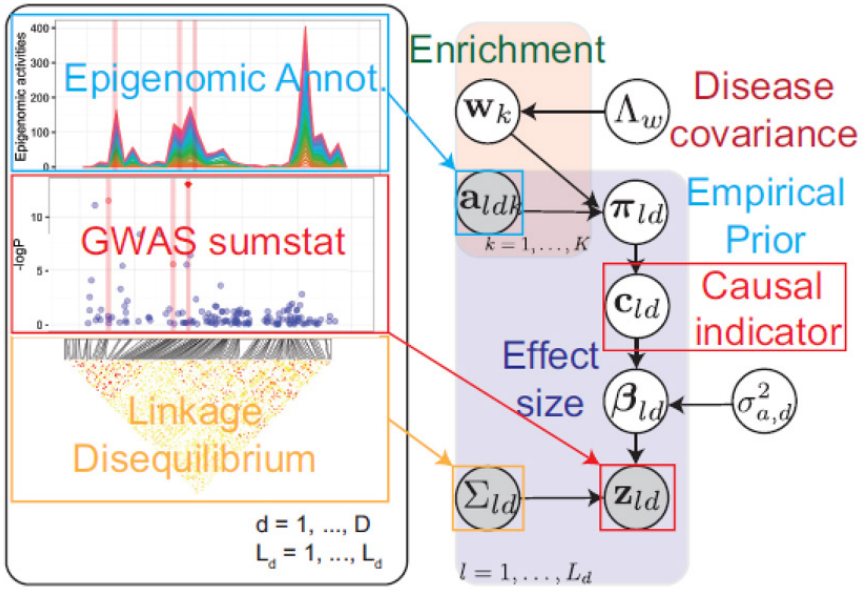
2 **EDNRA Heart/vasculature-only**, mediated through **multiple enhancers**



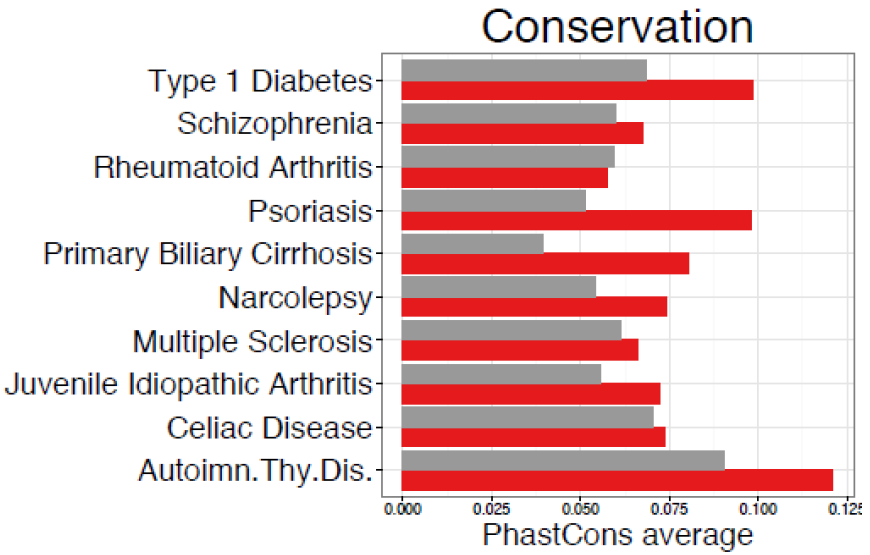
3 **PLPP3: Both liver and coronary artery: multi-gene/multi-tissue pleiotropy**



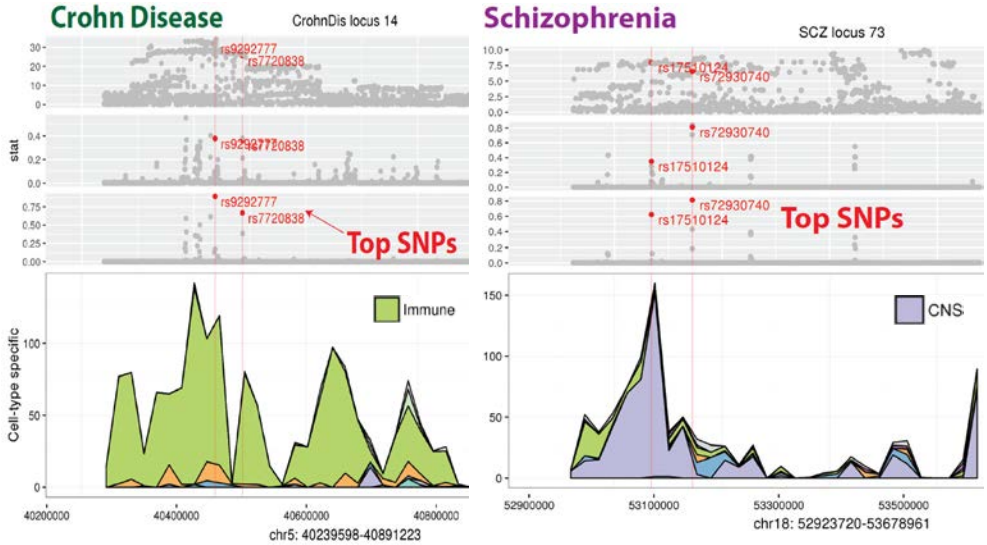
Bayesian fine-mapping: Predict causal variant and cell type



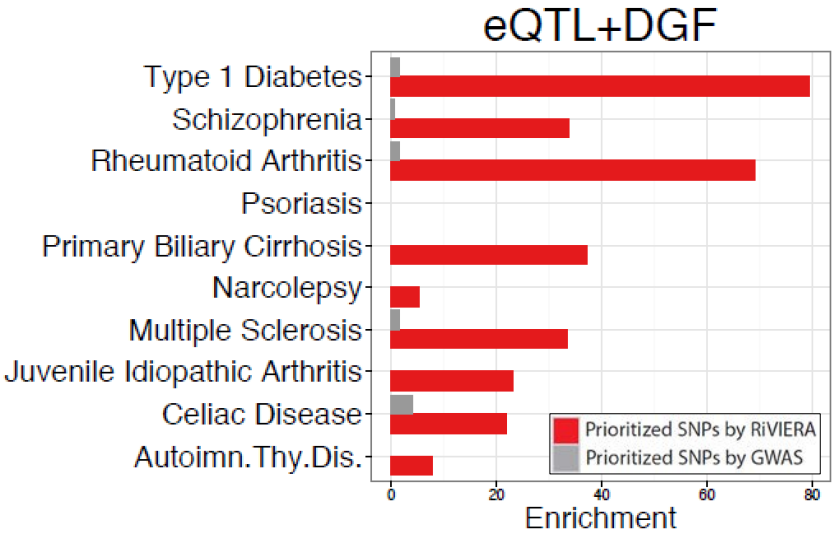
RiVIERA: multi-trait GWAS integration



Capture conserved elements

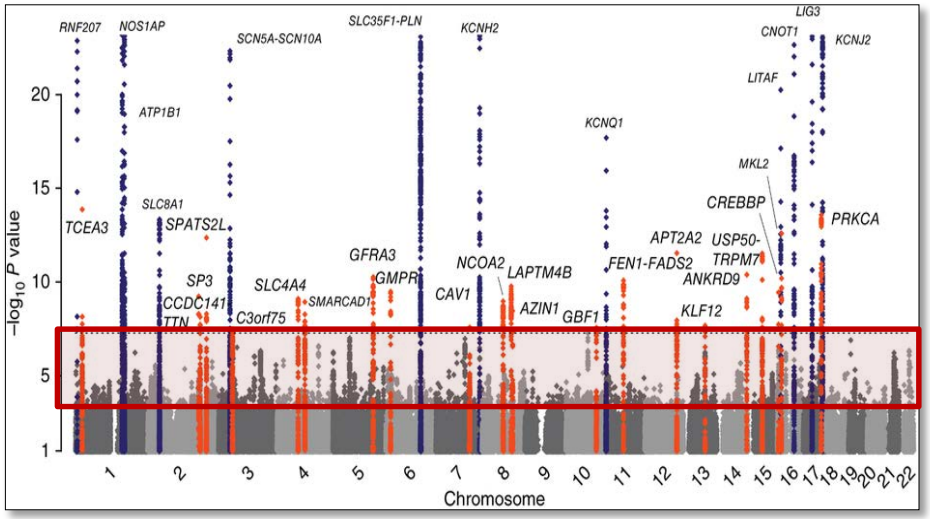


Predict causal variants and cell types



Capture eQTLs from GTEx

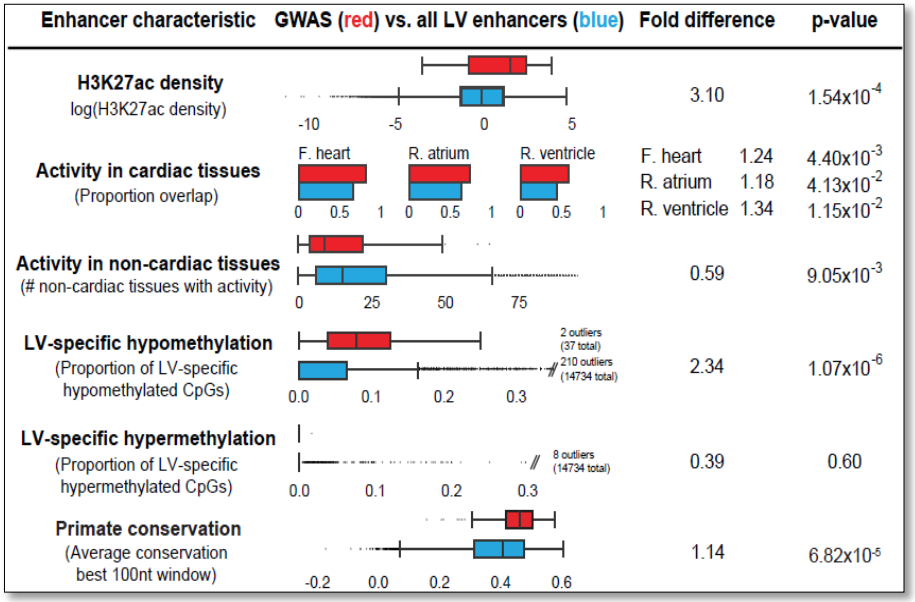
Combine GWAS+Epig to find new target genes/SNPs



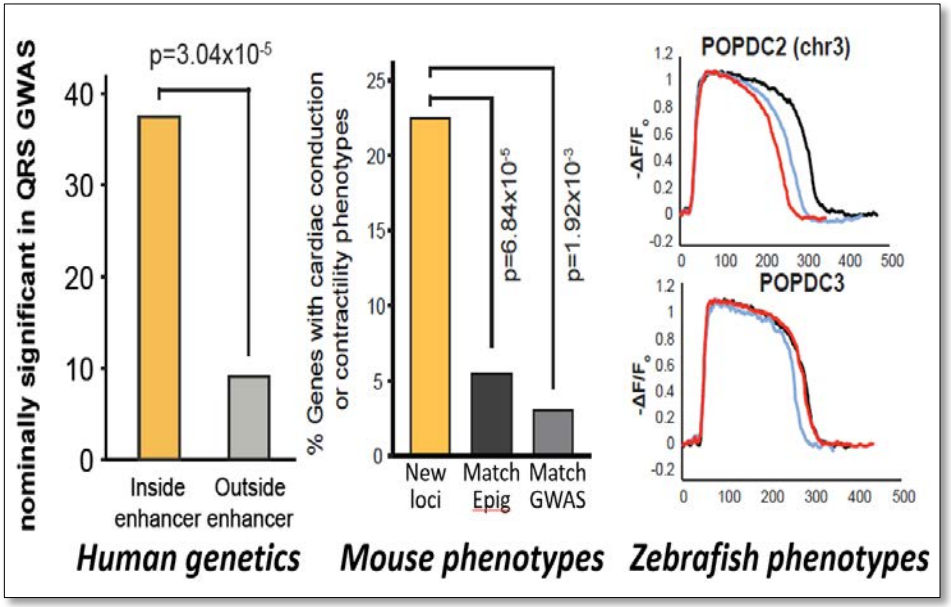
Lead SNP	p-value	Enhancer	1. Luciferase reporter	2. 4C-seq interactions
rs1886512	4.30×10^{-8}	chr13:74,520,000-74,520,400	0.015	No interactions
rs1044503	5.13×10^{-7}	chr14:102,965,400-102,972,000	4.70×10^{-9}	CINP, RCOR1
rs10030238	6.21×10^{-7}	chr4:141,807,800-141,809,600	1.35×10^{-14}	RNF150
		chr4:141,900,800-141,908,000	-	RNF150
rs6565060	1.52×10^{-5}	chr16:82,746,400-82,750,800	5.00×10^{-3}	No interactions
rs3772570	1.73×10^{-5}	chr3:148,733,200-148,738,600	0.67	-
rs3734637	2.23×10^{-5}	chr6:126,081,200-126,081,800	1.06×10^{-4}	HDDC2
rs1743292	6.48×10^{-5}	chr6:105,706,600-105,710,200	3.20×10^{-4}	BVES, POPDC3
		chr6:105,720,200-105,723,000	-	BVES, POPDC3
rs11263841	6.87×10^{-5}	chr1:35,307,600-35,312,200	0.22	GJA4, DLGAP3
rs11119843	7.14×10^{-5}	chr1:212,247,600-212,248,600	0.031	-
rs6750499	7.37×10^{-5}	chr2:11,559,600-11,563,000 (split into two 2kb fragments)	0.54	ROCK2
			3.26×10^{-7}	
rs17779853	7.73×10^{-5}	chr17:30,063,800-30,066,800	4.33×10^{-3}	No interactions

Prioritize sub-threshold loci (<10^{-4})

Validate new enhancers: allelic activity, enh-prom looping



Machine learning predictive features



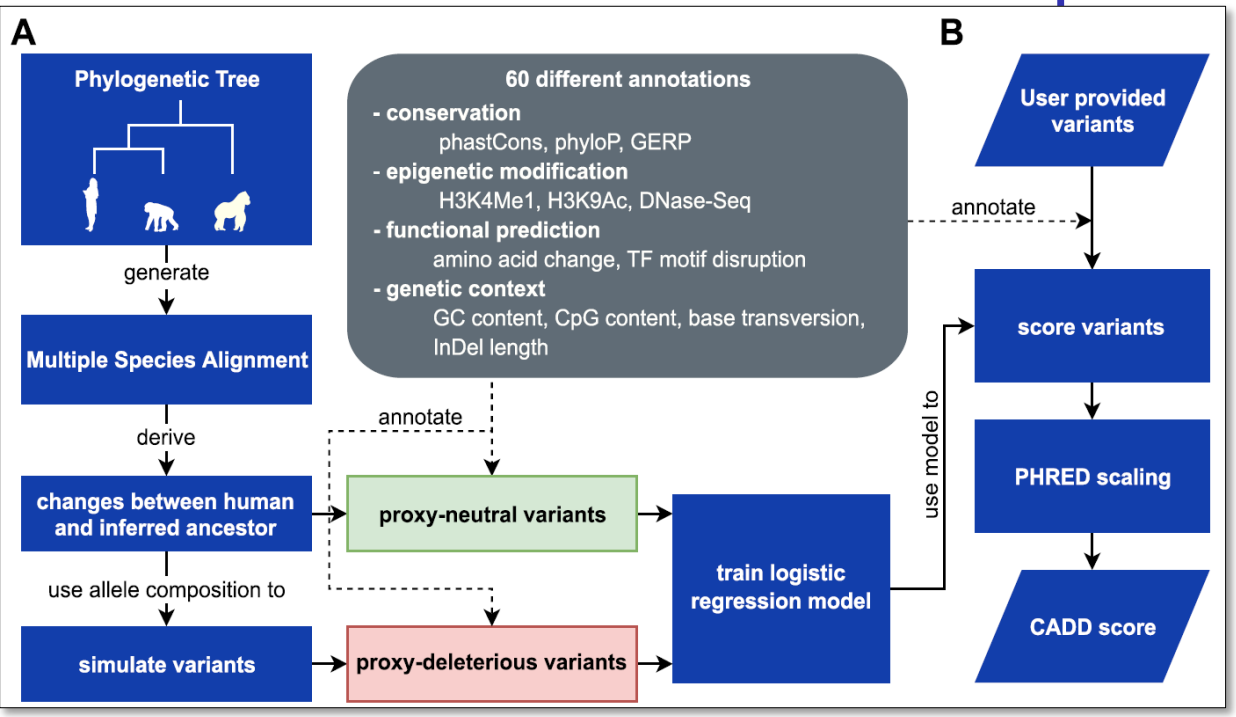
Validate new genes in hum/mou/zb

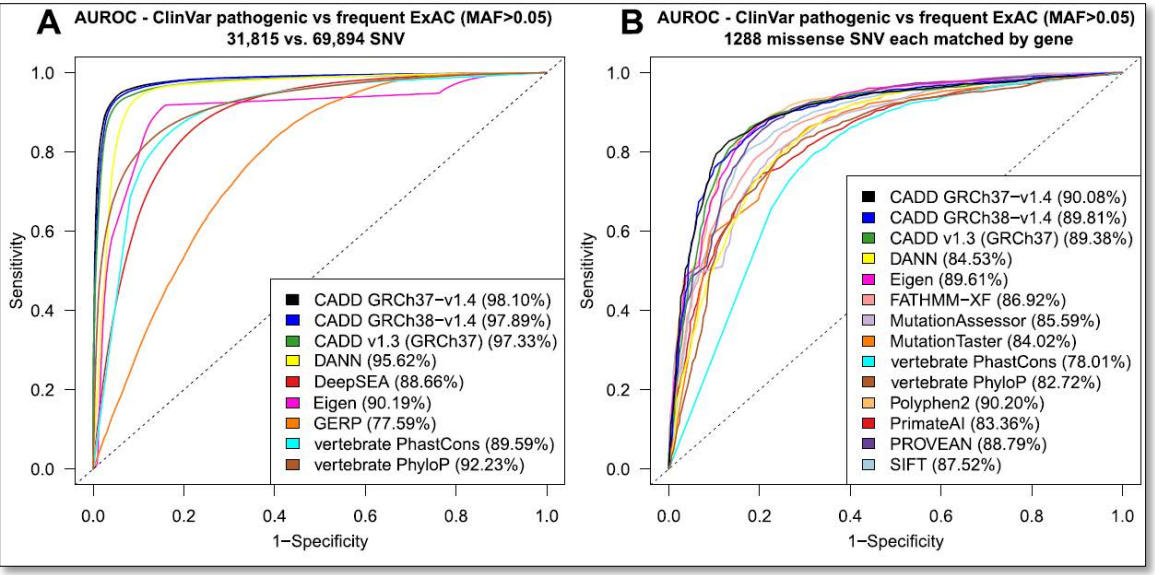
Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting: Common/rare alleles, Linkage vs. GWAS
3. LD, Haplotypes, Co-inheritance, and the challenge of fine-mapping
4. From locus to mechanism - Case study: FTO and Obesity
5. Epigenomics-GWAS integration: ENCODE, Roadmap, EpiMap
6. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA
7. Interpreting non-coding variation: DeepSEA (Jian Zhu guest lecture)

6. Machine Learning methods in genetics

CADD: combine evidence to predict variant function





Nucleic Acids Research, 2018 1
doi: 10.1093/nar/gky1016

CADD: predicting the deleteriousness of variants throughout the human genome

Philipp Rentzsch[1,2], Daniela Witten[3], Gregory M. Cooper[4], Jay Shendure[5,6,*] and Martin Kircher[1,2,5,*]

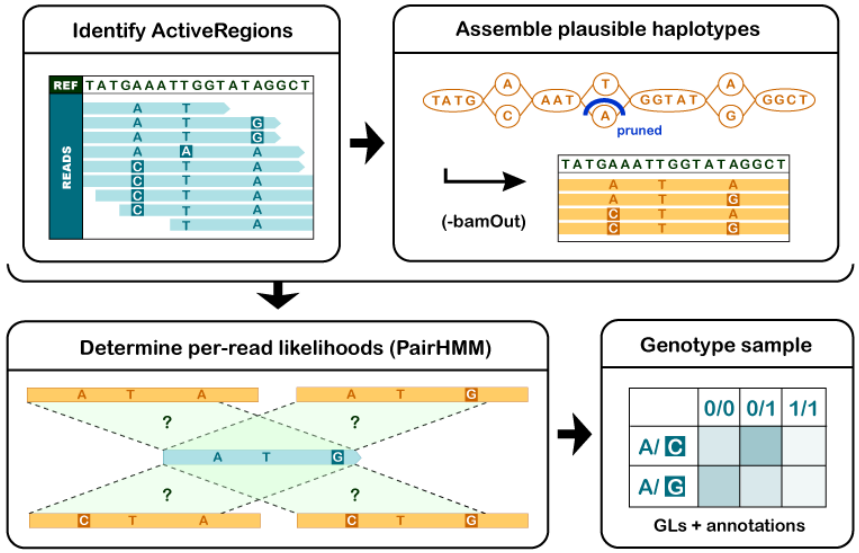
Large number of methods for variant prioritization

Score	Data sources	Approach	Refe
Eigen	• Uses data from the ENCODE and Roadmap Epigenomics projects	• Weighted linear combination of individual annotations • Unsupervised learning method	(14)
FunSeq2	• Inter- and Intra-species conservation • Loss- and gain-of-function events for transcription factor binding • Enhancer–gene linkage	• Weighted scoring system	(15)
LINSIGHT	• Conservation scores (phastCons, phylopP), predicted binding sites (TFBS, RNA), regional annotations (ChIP-seq, RNA-seq)	• Graphical model • Selection parameter fitting using generalized linear model based on 48 genomic features	(16)
CADD	• Ensembl variant effect predictor • Protein-level scores: Grantham, SIFT, PolyPhen • DNase hypersensitivity, TFBS, transcript information • GC content, CpG content, histone methylation	• Support vector machine	(11)
FATHMM	• 46-way sequence conservation • ChIP-seq, TFBS, DNase-seq • FAIRE, footprints, GC content	• Hidden Markov models	(17)
ReMM	• Predict potential of non-coding variant to cause a Mendelian disease if mutated • 26 features: PhastCons, PhyloP, CpG, GC, regulation annotations	• Random forest classifier	(18)
Orion	• Predict potential of non-coding variant to cause a Mendelian disease if mutated • Independent from annotation and features	• Expected and observed site-frequency spectrum of a given stretch of sequence	(19)
CDTS	• Identify constrained non-coding regions in the human genome and deleteriousness of variants • Independent from annotation and features. Uses *k*-mers	• Expected and observed site-frequency spectrum of a given heptamer	(8)

DePristo, Human Molecular Genetics, 2018, 10.1093/hmg/ddy115

Whole genome variant calling: GATK HaplotypeCaller

1. Use heuristic to find mismatches not explained by noise
2. Use assembly graph to identify possible haplotypes
3. For each haplotype, estimate:
   P(read | haplotype)
   using *probabilistic sequence alignment*
   - Hidden Markov Model
   - States: insertion, deletion, substitution
   - Emissions: pairs of aligned nucleotides/gaps
   - Transitions: equivalent to insertion/deletion/gap penalties from Smith-Waterman algorithm (DP alignment)
   - Get **P(read | haplotype)** using forward-backward algorithm
4. Use Bayes rule to get **P(haplotype | read)**
5. Assign genotypes to each sample based on the max a posteriori haplotypes

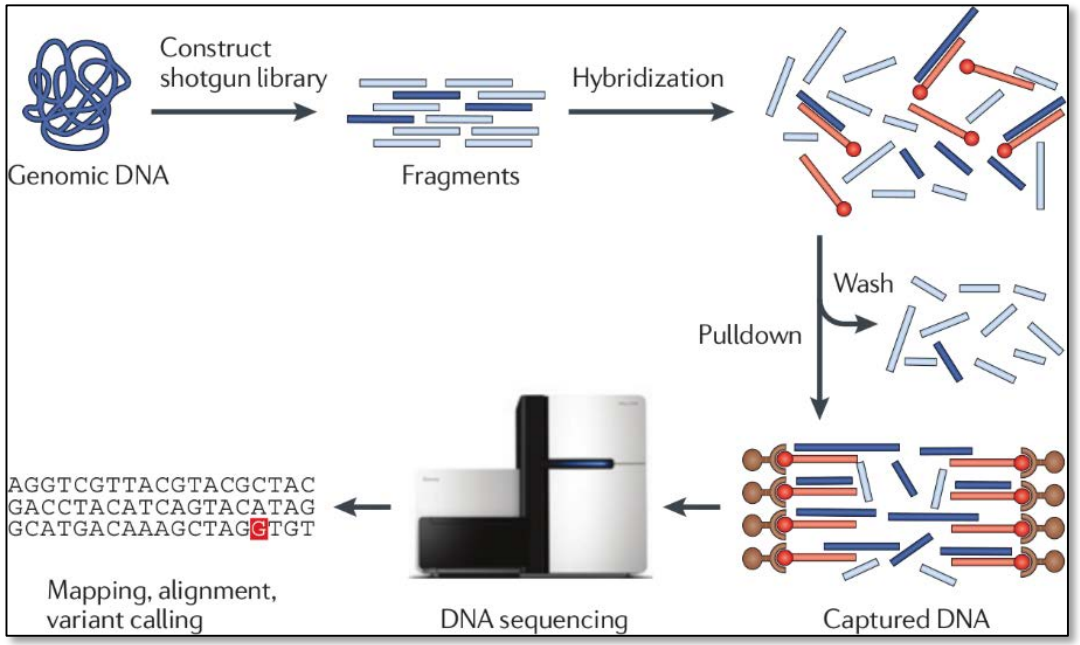


Tour de Force, combining many methods:

- **Logistic regression** to model base errors
- **Hidden Markov models** to compute read likelihoods
- **Naive Bayes** classification to identify variants
- **Gaussian mixture model** with hand-crafted features to filter likely false positive variants, capturing common error modes

http://gatkforums.broadinstitute.org/gatk/discussion/4148/hc-overview-how-the-haplotypecaller-works

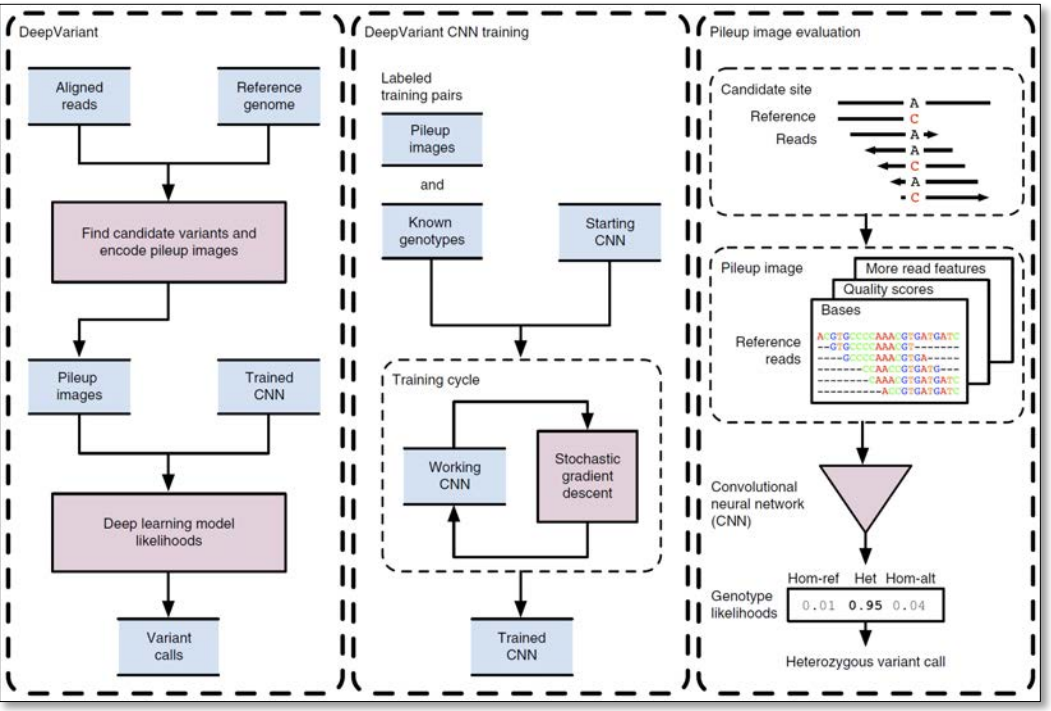
Exome variant calling: atlas2



- Motivation: the exome has different sequence properties than the rest of the genome (e.g., substitution rates, GC content).
- Train **logistic regression classifier** to predict which mismatches are errors and which are variants
  - Training data: 1KG Exome project sequencing reads where >2 reads align with a mismatch
  - True positives: Reads where mismatch is also discovered in 1KG Exon pilot project
  - True negatives: Remaining reads
  - Features: mismatch quality score, flanking quality score, whether neighboring nucleotides were swapped, normalized distance to 3' end of the read
- Much faster than full Bayesian model (e.g. HaplotypeCaller), lower false positive rate in validation data

Bamshad et al. *Nat Rev Genet* 2011

DeepVariant: Combine evidence to call variants

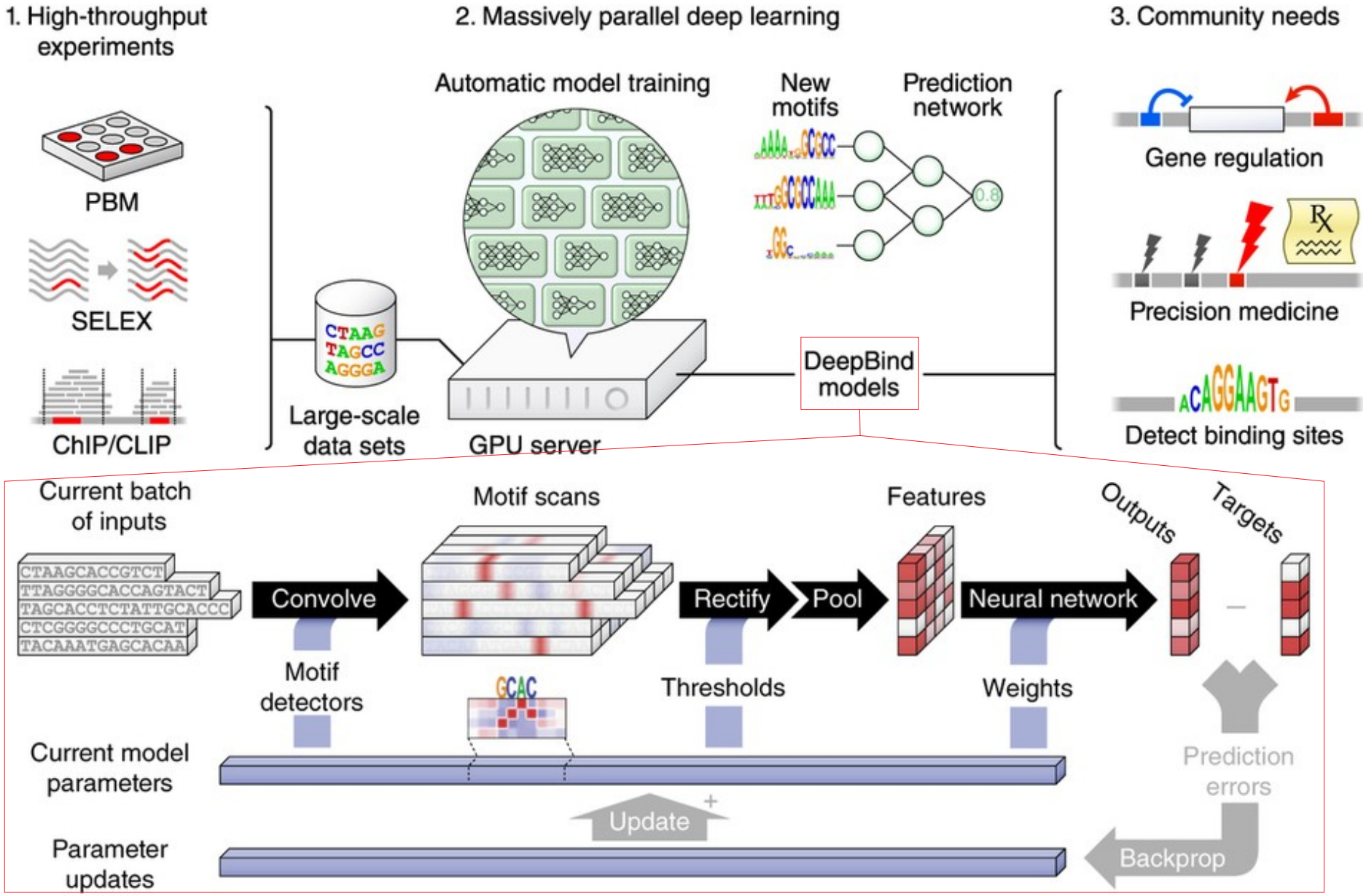


A universal SNP and small-indel variant caller using deep neural networks

Ryan Poplin[1,2], Pi-Chuan Chang[2], David Alexander[2], Scott Schwartz[2], Thomas Colthurst[2], Alexander Ku[2], Dan Newburger[1], Jojo Dijamco[1], Nam Nguyen[1], Pegah T Afshar[1], Sam S Gross[1], Lizzie Dorfman[1,2], Cory Y McLean[1,2] & Mark A DePristo[1,2]

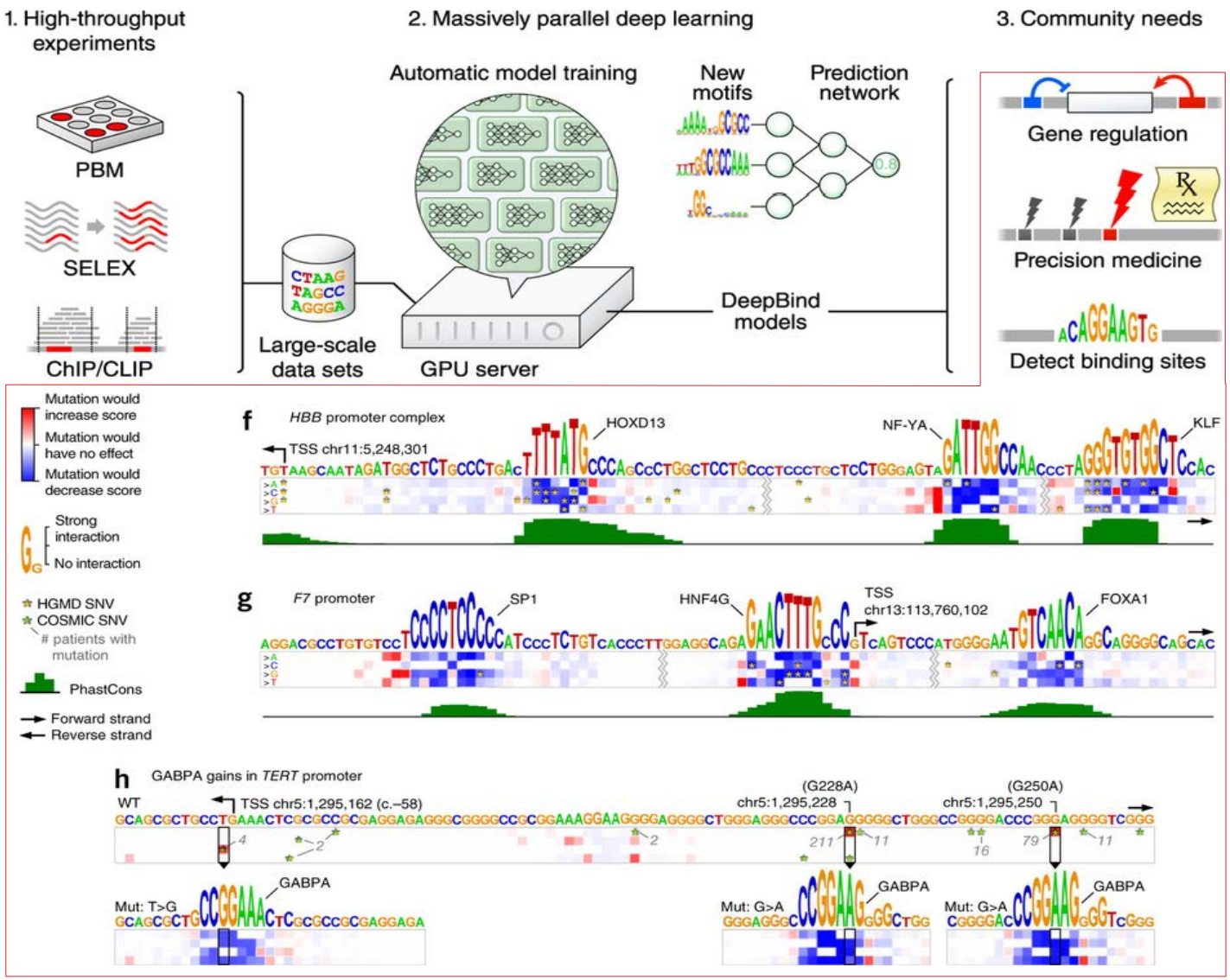
Method	Type	F1	Recall	Precision	TP	FN	FP	FP.gt	FP.al	Version
DeepVariant (live GitHub)	Indel	0.99507	0.99347	0.99666	357,641	2350	1,198	217	840	Latest GitHub v0.4.1-b4e8d37d
GATK (raw)	Indel	0.99366	0.99219	0.99512	357,181	2810	1,752	377	995	3.8-0-ge9d806836
Strelka	Indel	0.99227	0.98829	0.99628	355,777	4214	1,329	221	855	2.8.4-3-gbe58942
DeepVariant (pFDA)	Indel	0.99112	0.98776	0.99450	355,586	4405	1,968	846	1,027	pFDA submission May 2016
GATK (VQSR)	Indel	0.99010	0.98454	0.99573	354,425	5566	1,522	343	909	3.8-0-ge9d806836
GATK (flt)	Indel	0.98229	0.96881	0.99615	348,764	11227	1,349	370	916	3.8-0-ge9d806836
FreeBayes	Indel	0.94091	0.91917	0.96372	330,891	29,100	12,569	9,149	3,347	v1.1.0-54-g49413aa
16GT	Indel	0.92732	0.91102	0.94422	327,960	32,031	19,364	10,700	7,745	v1.0-34e8f934
SAMtools	Indel	0.87951	0.83369	0.93066	300,120	59,871	22,682	2,302	20,282	1.6
DeepVariant (live GitHub)	SNP	0.99982	0.99975	0.99989	3,054,552	754	350	157	38	Latest GitHub v0.4.1-b4e8d37d
DeepVariant (pFDA)	SNP	0.99958	0.99944	0.99973	3,053,579	1,727	837	409	78	pFDA submission May 2016
Strelka	SNP	0.99935	0.99893	0.99976	3,052,050	3,256	732	87	136	2.8.4-3-gbe58942
GATK (raw)	SNP	0.99914	0.99973	0.99854	3,054,494	812	4,469	176	257	3.8-0-ge9d806836
16GT	SNP	0.99583	0.99850	0.99318	3,050,725	4,581	20,947	3,476	3,899	v1.0-34e8f934
GATK (VQSR)	SNP	0.99436	0.98940	0.99937	3,022,917	32,389	1,920	80	170	3.8-0-ge9d806836
FreeBayes	SNP	0.99124	0.98342	0.99919	3,004,641	50,665	2,434	351	1,232	v1.1.0-54-g49413aa
SAMtools	SNP	0.99021	0.98114	0.99945	2,997,677	57,629	1,651	1,040	200	1.6
GATK (flt)	SNP	0.98958	0.97953	0.99983	2,992,764	62,542	509	168	26	3.8-0-ge9d806836

DeepBind



[Alipanahi et al., 2015]

Predicting disease mutations



[Alipanahi et al., 2015]

DeepBind summary

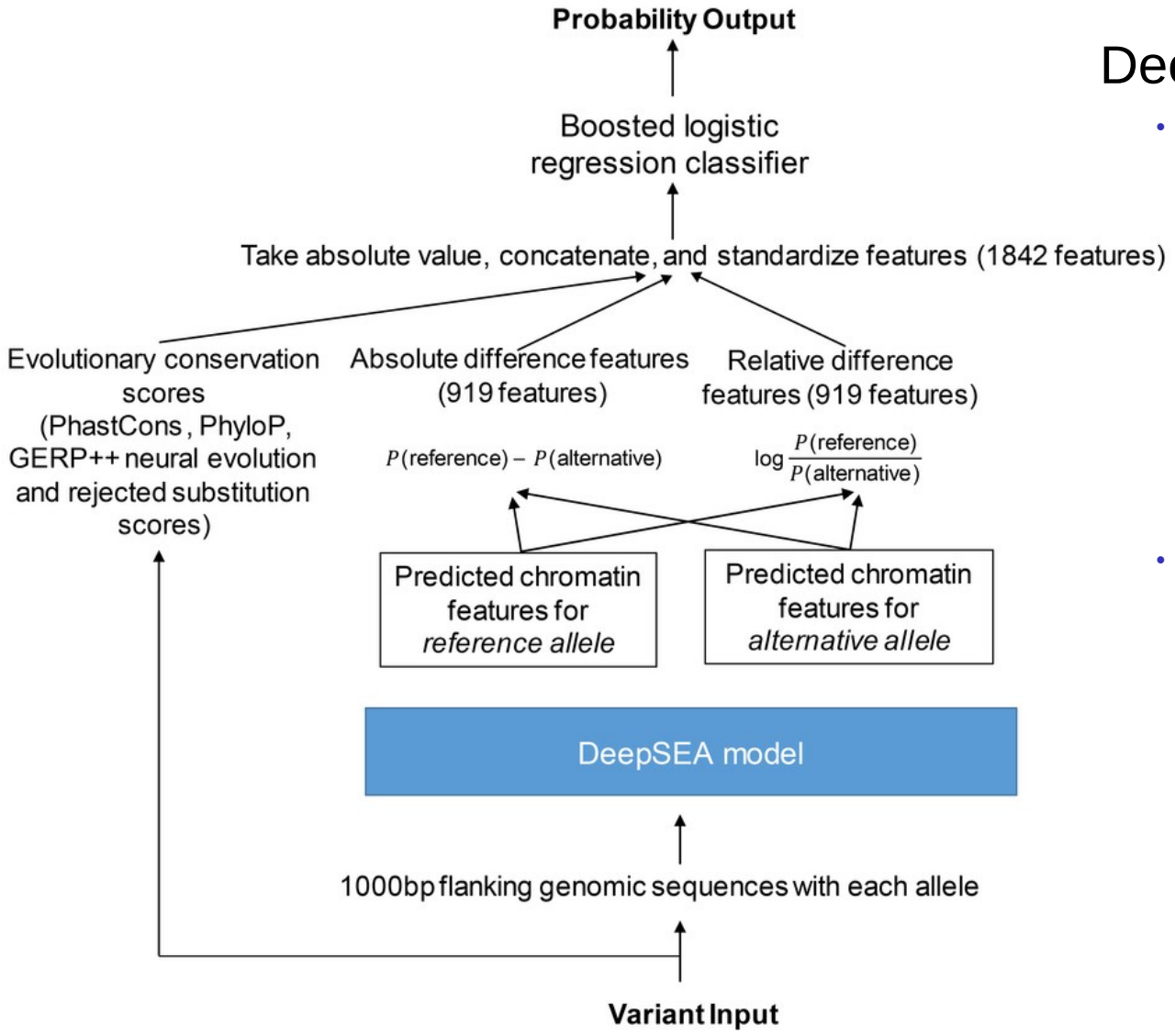
The key deep learning techniques:

- Convolutional learning
- Representational learning
- Back-propagation and stochastic gradient
- Regularization and dropout
- Parallel GPU computing especially useful for hyperparameter search

Limitations in DeepBind:

- Require defining negative training examples, which is often arbitrary
- Using observed mutation data only as post-hoc evaluation
- Modeling each regulatory dataset separately

DeepSea



DeepSea:

- Similar as DeepBind but trained a separate CNN on each of the ENCODE/Roadmap Epigenomic chromatin profiles 919 chromatin features (125 DNase features, 690 TF features, 104 histone features).
- It uses the Δs mutation score as input to train a linear logistic regression to predict GWAS and eQTL SNPs defined from the GRASP database with a P-value cutoff of 1E-10 and GWAS SNPs from the NHGRI GWAS Catalog

[Zhou and Troyanskaya, 2015]

CNNs for DNA-binding prediction from sequence

DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Uses convolution layers to capture regulatory motifs, and a recurrent layer to discover a 'grammar' for how these single motifs work together. Based on Keras/Theano.

Basset—learning the regulatory code of the accessible genome with deep convolutional neural networks. CNN to discover regulatory sequence motifs to predict the accessibility of chromatin. Accounts for cell-type specificity using multi-task learning.

DeepBind and DeeperBind—predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Based on ChIP-seq, ChIP-chip, RIP-seq, protein-binding microarrays and others. Deeperbind adds a recurrent sequence learning module (LSTM) after the convolutional layer(s).

DeepMotif—visualizing genomic sequence classifications. Predicting binding specificities of proteins to DNA motifs. Makes use of a convolutional layers with more layers than the DeepBind network.

Convolutional neural network architectures for predicting DNA–protein binding. Systematic exploration of CNN architectures for predicting DNA sequence binding using a large compendium of transcription factor data sets.

Predicting enhancers, 3d interactions and cis-regulatory regions

PEDLA: predicting enhancers with a deep-learning-based algorithmic framework. Predicting enhancers based on heterogeneous features from (e.g.) the ENCODE project using a deep learning, HMM hybrid model.

DEEP: a general computational framework for predicting enhancers. Predicting enhancers based on data from the ENCODE project.

Genome-wide prediction of cis-regulatory regions using supervised deep-learning methods. toolkit based on the Theano) for applying different deep-learning architectures to cis-regulatory elements.

FIDDLE: an integrative deep-learning framework for functional genomic data inference. Prediction of transcription start site and regulatory regions. FIDDLE stands for Flexible Integration of Data with Deep Learning that models several genomic signals using convolutional networks (DNase-seq, ATAC-seq, ChIP-seq, TSS-seq, RNA-seq signals).

DNA methylation

DeepCpG—predicting DNA methylation in single cells. Neural network for predicting DNA methylation in multiple cells.

Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Uses a stacked autoencoder with a supervised layer on top of it to predict whether CpG islands are methylated.

Variant callers, pathogenicity scores and identification of genomic elements

DeepVariant—a variant caller in germline genomes. Uses a deep neural network architecture (Inception-v3) to identify SNP and small indel variants from next-generation DNA sequencing data.

DeepLNC, a long non-coding RNA prediction tool using deep neural network. Identification of lncRNA-based on *k*-mer profiles.

evoNet—deep learning for population genetic inference [code][paper]. Jointly inferring natural selection and demographic history

DANN. Uses the same feature set and training data as CADD to train a deep neural network

DeepSEA—predicting effects of non-coding variants with deep-learning-based sequence model. Models chromatin accessibility as well as the binding of transcription factors, and histone marks associated with changes in accessibility.

Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting: Common/rare alleles, Linkage vs. GWAS
3. LD, Haplotypes, Co-inheritance, and the challenge of fine-mapping
4. From locus to mechanism - Case study: FTO and Obesity
5. Epigenomics-GWAS integration: ENCODE, Roadmap, EpiMap
6. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA
7. Interpreting non-coding variation: DeepSEA (Jian Zhu guest lecture)

7. Guest Lecture: Jian Zhu on DeepSEA