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

Lecture 16: Deep Learning for human genetics

Prof. Manolis Kellis

Massachusetts Institute of Technology

http://mit6874.github.io

Slides credit: Mark Daly, David Gifford, et al

Today: Human genetics and GWAS

1. **Genetics intro: Mendel, human traits, Linkage**
   – Mendelian traits: linkage, genetic maps, family studies
   – Complex traits: polygenicity, environment, continuous
2. **Genome-wide association studies (GWAS)**
   – Study design: QC, Chi-Sq, mult testing, replication, QQ
   – Fine mapping, linkage vs. association, combine studies
3. **Interpretation: Individual loci vs. global signals**
   – Loci: mechanism, assays, allelic, patients, mice, cells
   – Sys: networks, biases, stats, PPI, expression, diagnostics
4. **Next Gen: exome, genome, medical sequencing**
   – Common, low-freq, rare, private, somatic; diagnostics
   – Autism: polygenic, carriers; case/control, trios; de novo

Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. ‘Disease gene’ hunting (locus, really): Common/rare alleles, Linkage vs. GWAS
3. Evolution/scaling of GWAS power: Sharing, inflection points
4. Challenge of fine-mapping: Co-inheritance, LD, Haplotypes, Recomb.
5. From locus to mechanism - Case study: FTO and Obesity
6. Challenges in disease mechanism inference
7. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA

1. Intro to Human Genetics

Inheritance, Mendel, Fisher, SNPs, STRs, alleles

Inheritance and Genetics: Ancient foreshadowings

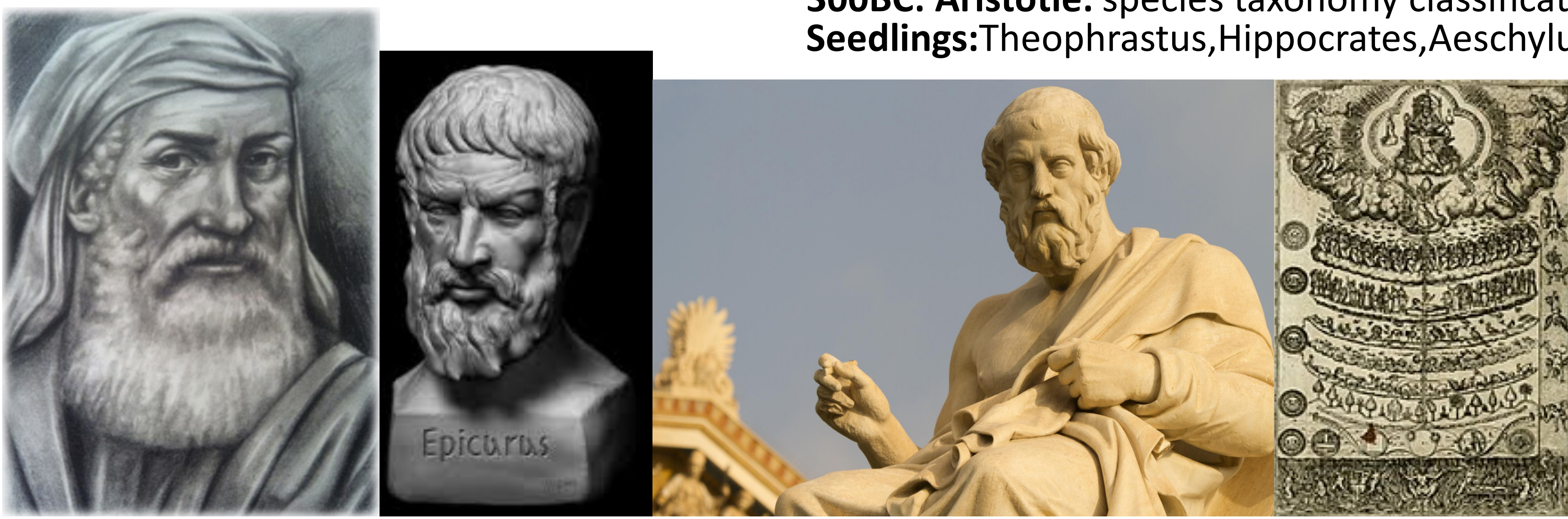
9000BC: Selective breeding of animals/plants

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

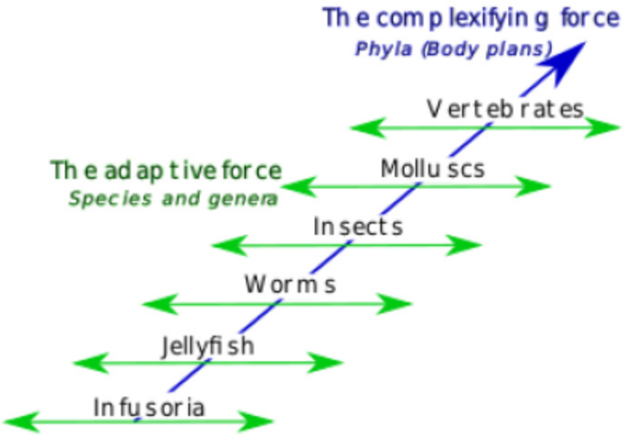
Seedlings: Theophrastus,Hippocrates,Aeschylus



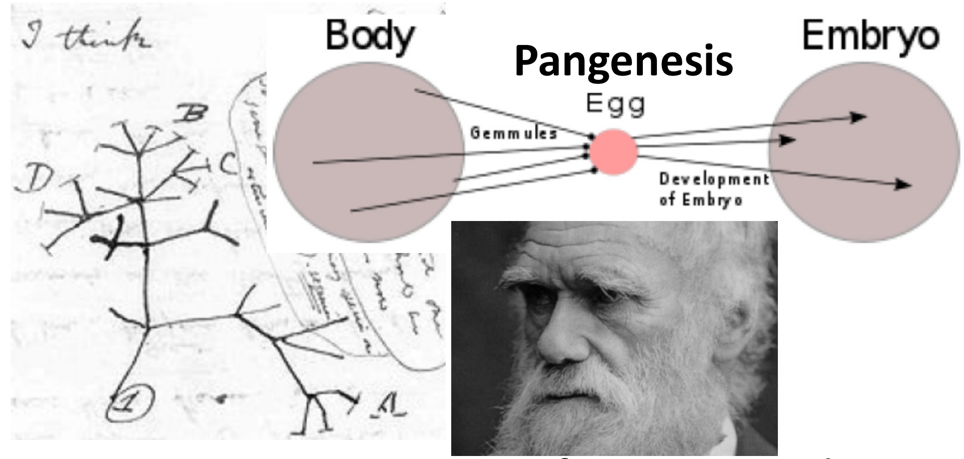
450BC: Empedocles: Random mixing of traits, natural variation, successful ones survive, giving semblance of 'purpose'

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

19th Century: Lamarck, Darwin, Mendel, Biometrics

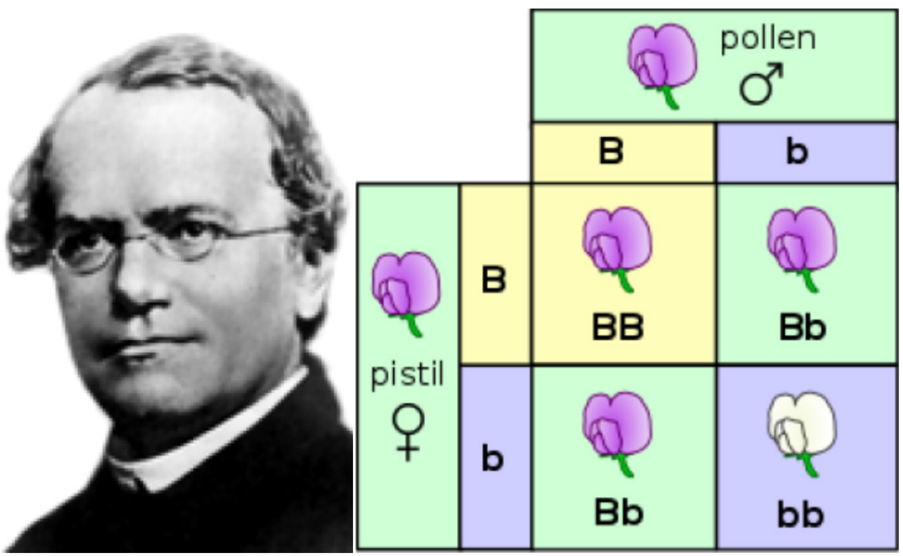


1809: Lamarck: Transmutation, adaptation spontaneous generation of simple life-forms, innate life force drives increased complexity

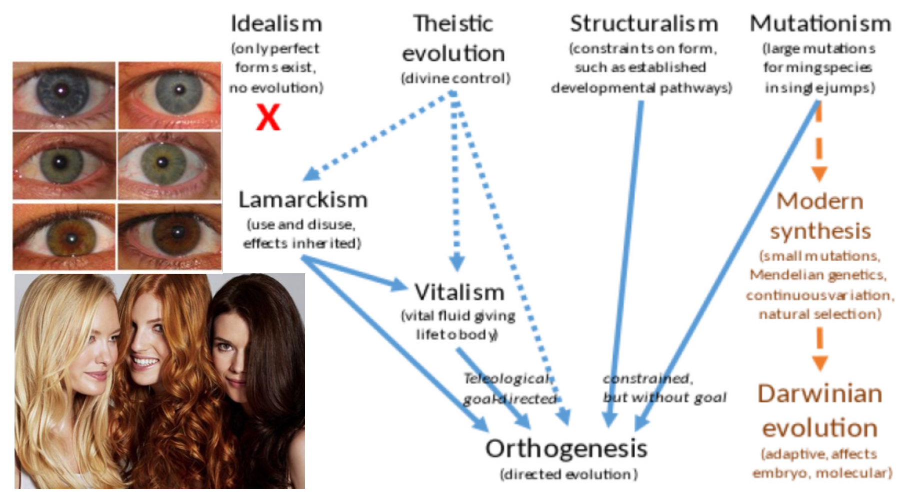


1859: Darwin: Continuum of species, random mutation→diversity, natural selection→fitness.

But: Blending inheritance, gemmules, Lamarckism

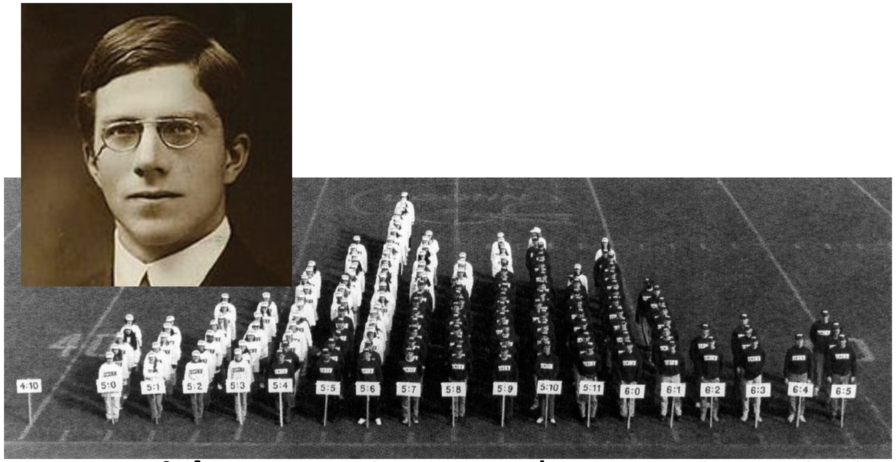


1866: Mendel: Particulate inheritance, no blend. Discrete units=genes. Dominant/recessive alleles Independent assortment. **Digital inheritance.**



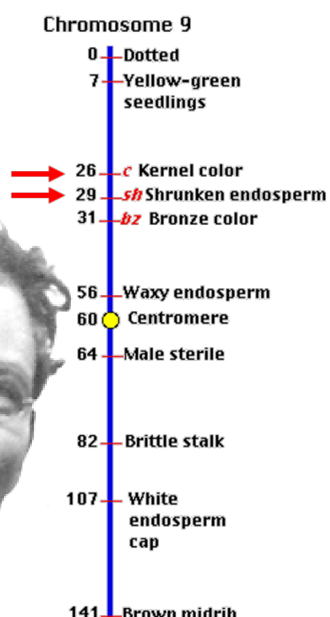
Biometrics: continuous phenotype variation.
Others: Saltationism, orthogenesis, vitalism, neo-Lamarckism, theistic evolution…

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)

CATGGTGCATCTGACTCCTGAG**GAG**AAGTCTGCCGTTACTG

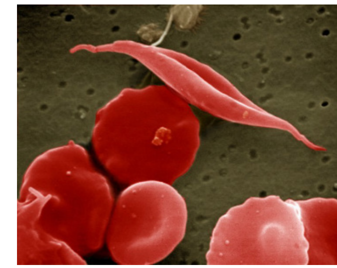
CATGGTGCATCTGACTCCTG**T**G**GAG**AAGTCTGCCGTTACTG

Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

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 **TCACAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT**

10 **TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT**

12 **TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT**

> 30 Huntington's Disease

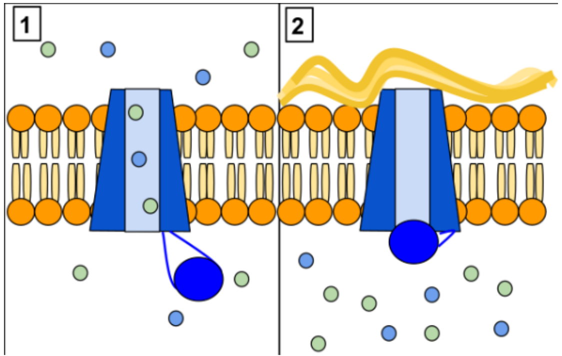
Abnormal protein, damages neurons, brain cell death, mood, coordination, speaking, dementia, etc

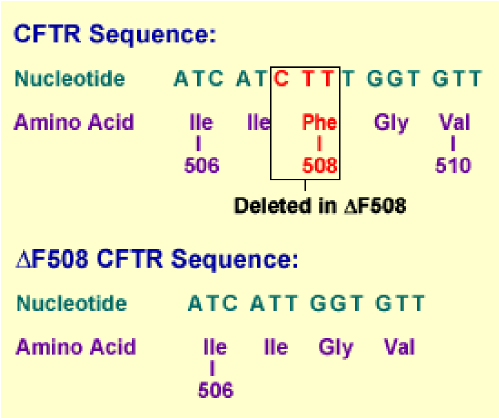
- Insertion/Deletions

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

CATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATA

CATTAAAGAAAATATCATTGGTGTTTCCTATGATGAATA



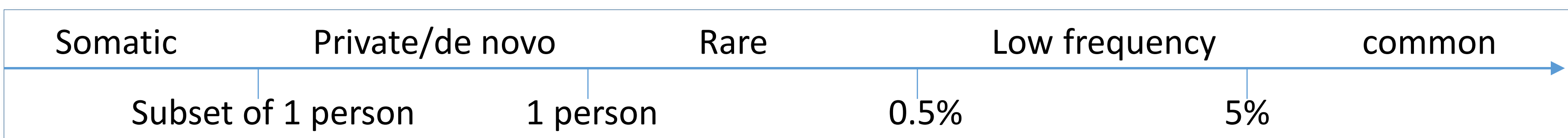


SNP alleles: ref/alt; maj/min; risk/prot; anc/der

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:



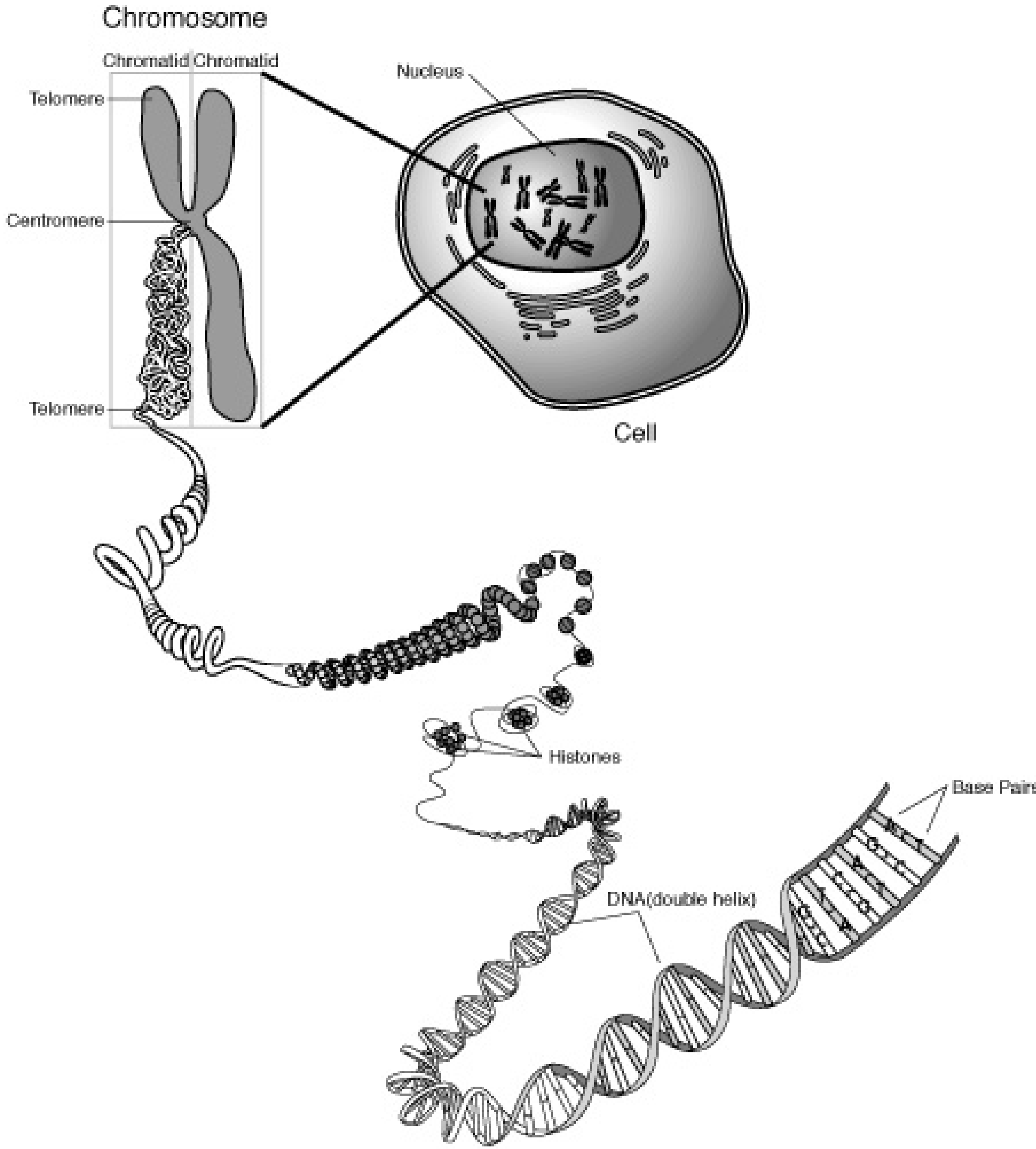
Example: rs189107123

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

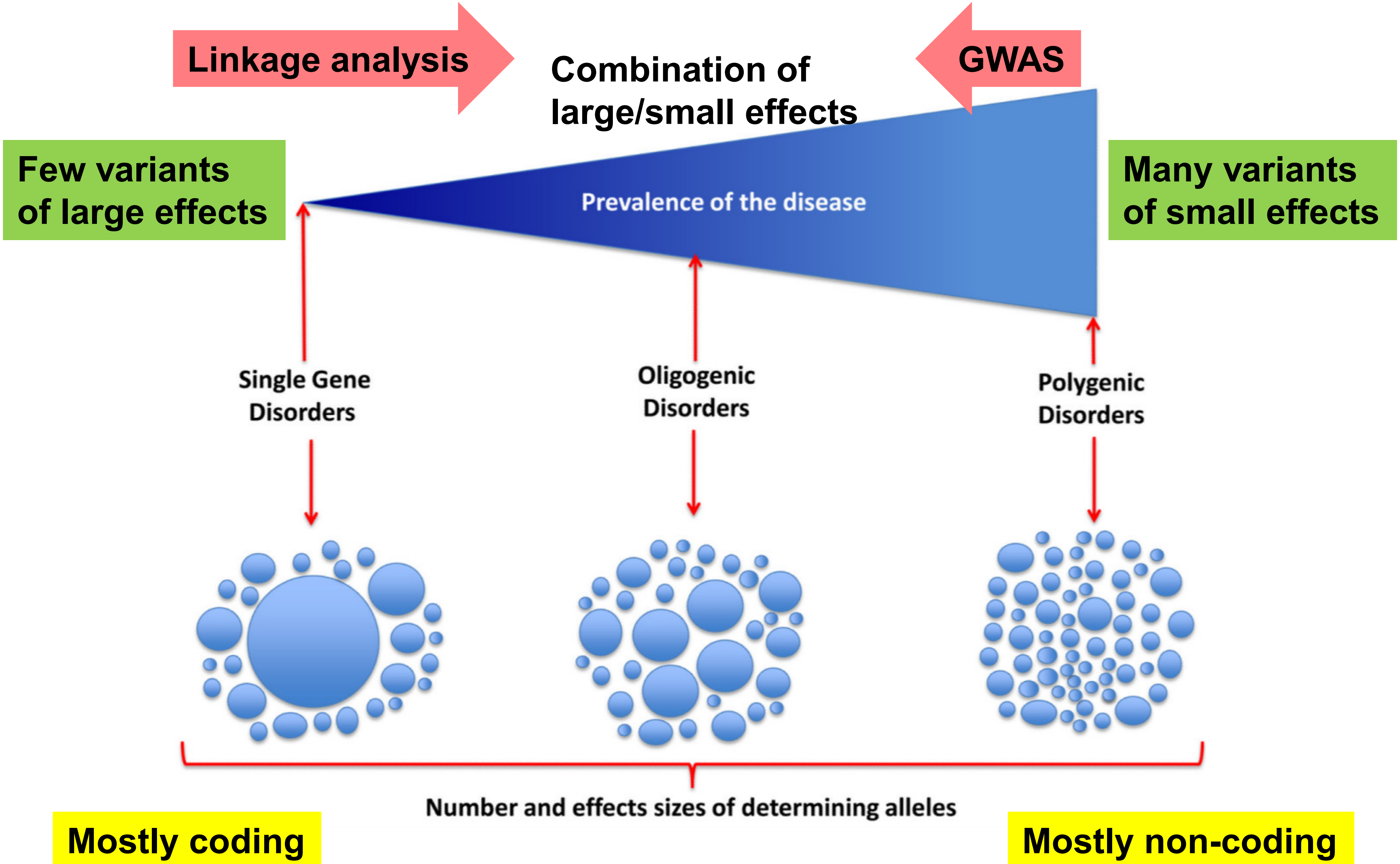
Millions of polymorphic sites

Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. ‘Disease gene’ hunting (locus, really): Common/rare alleles, Linkage vs. GWAS
3. Evolution/scaling of GWAS power: Sharing, inflection points
4. Challenge of fine-mapping: Co-inheritance, LD, Haplotypes, Recomb.
5. From locus to mechanism - Case study: FTO and Obesity
6. Challenges in disease mechanism inference
7. Machine learning tools for variant interpretation
    - Deep variant
    - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
    - DeepSEA

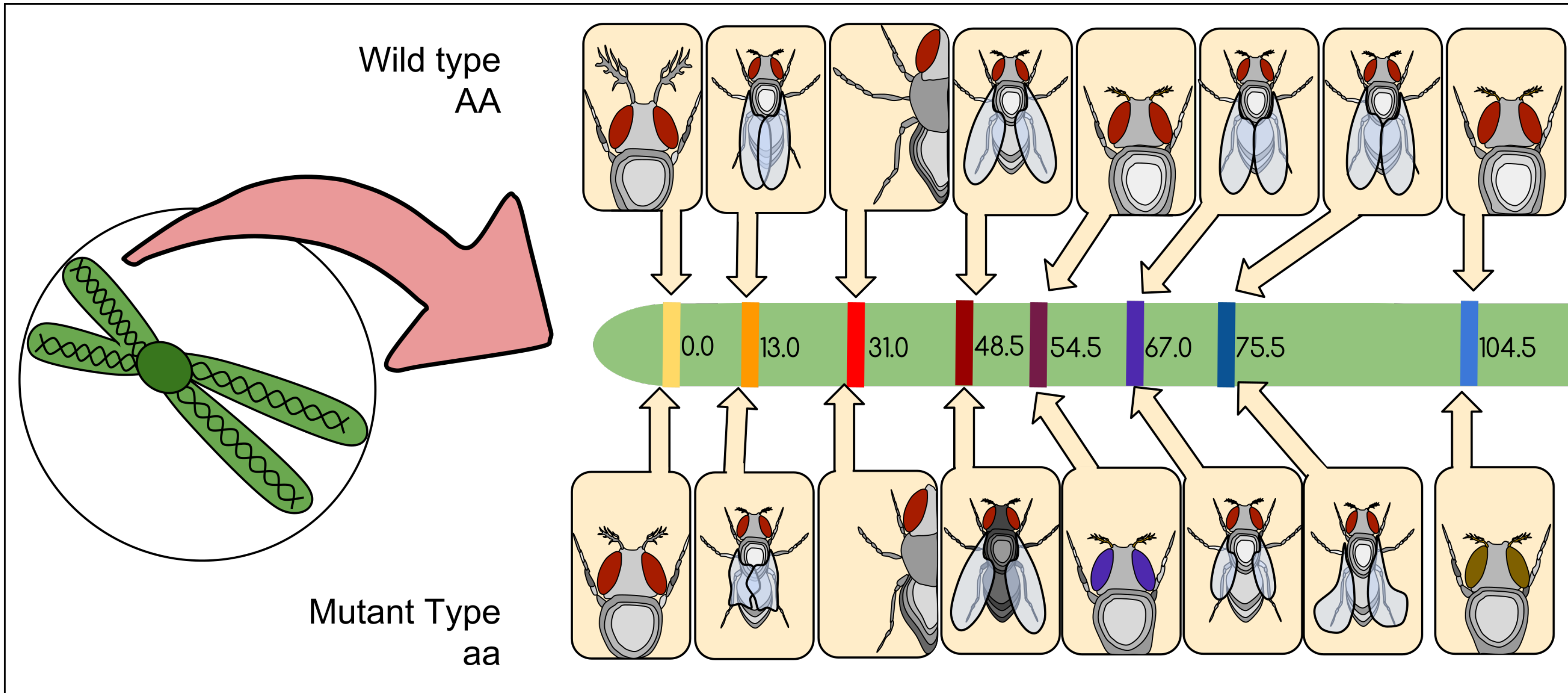
2. ‘Disease gene’ hunting (locus, really):
Common/rare alleles, Linkage vs. GWAS

Monogenic vs. oligogenic vs. polygenic disorders



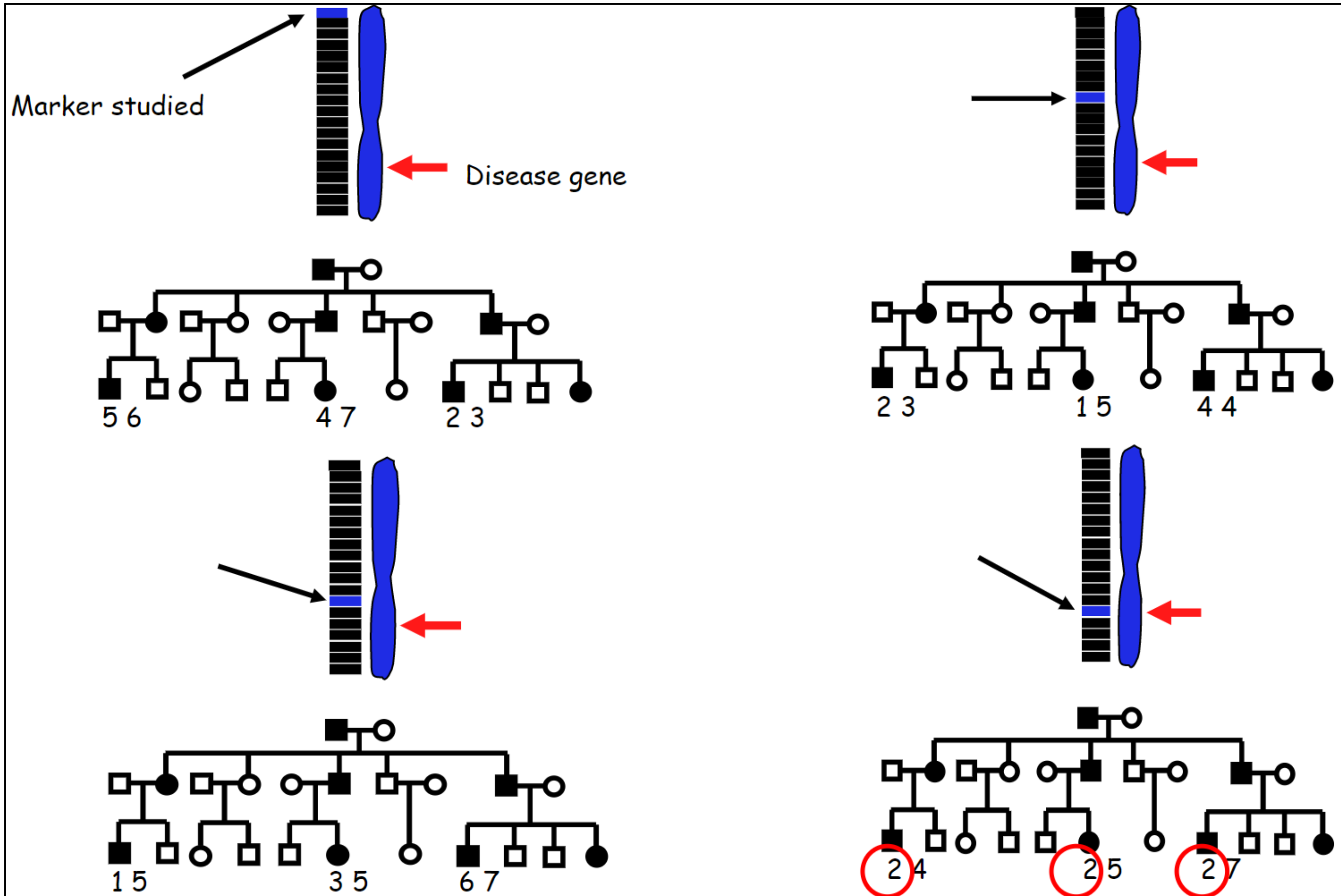
Molecular genetic studies of complex phenotypes. Marian - Translational Research, 2012

Linkage analysis allows mapping of genetic traits



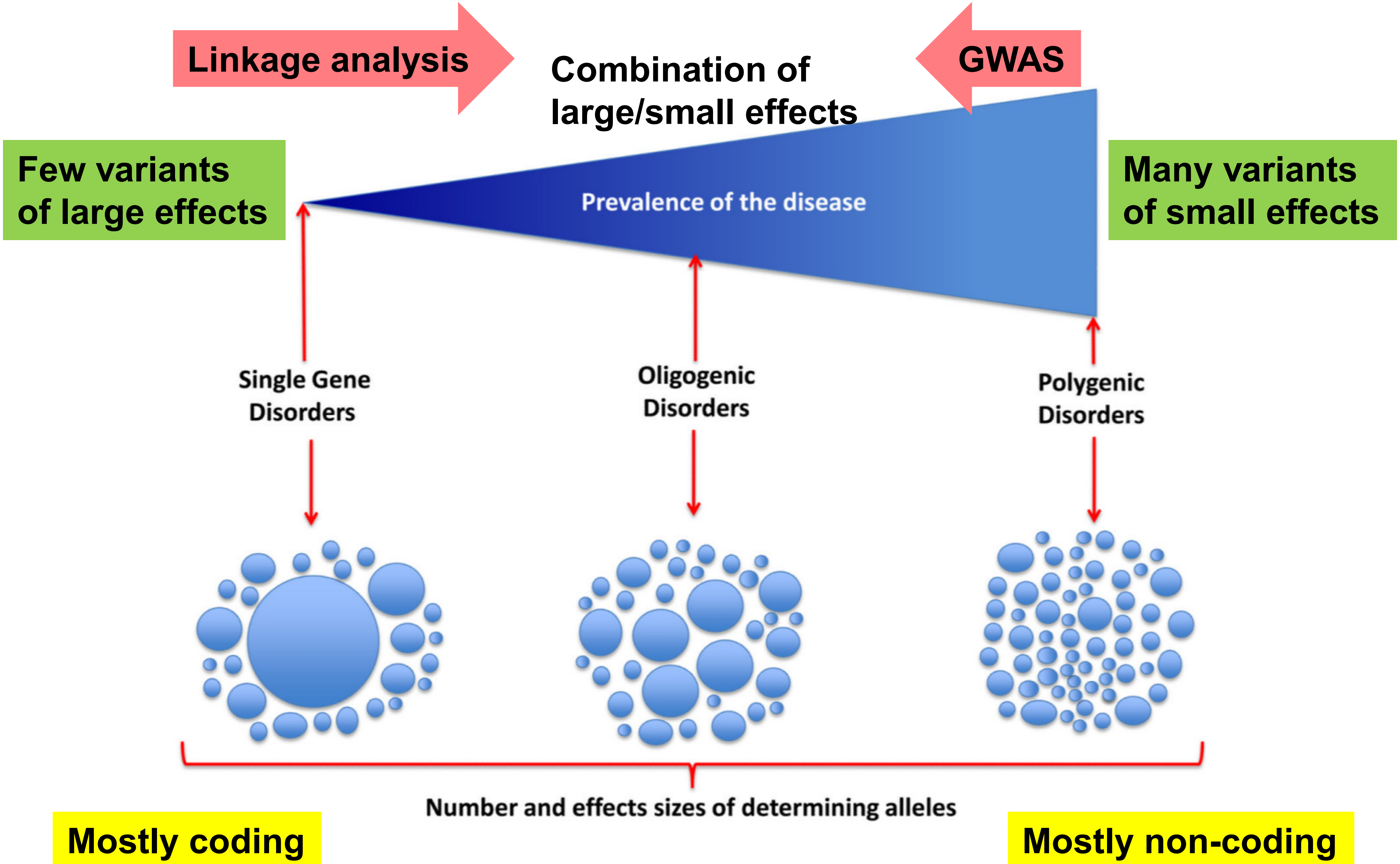
- Frequency of co-inheritance tells you about genetic distance on the chromosome
- Allows making of genetic map of genetic elements (eventually recognized to be genes)
- Used in human to lay out chromosomal maps based on inheritance of STR markers, well before mapping of first genetic trait in human

Linkage analysis allows mapping of disease loci



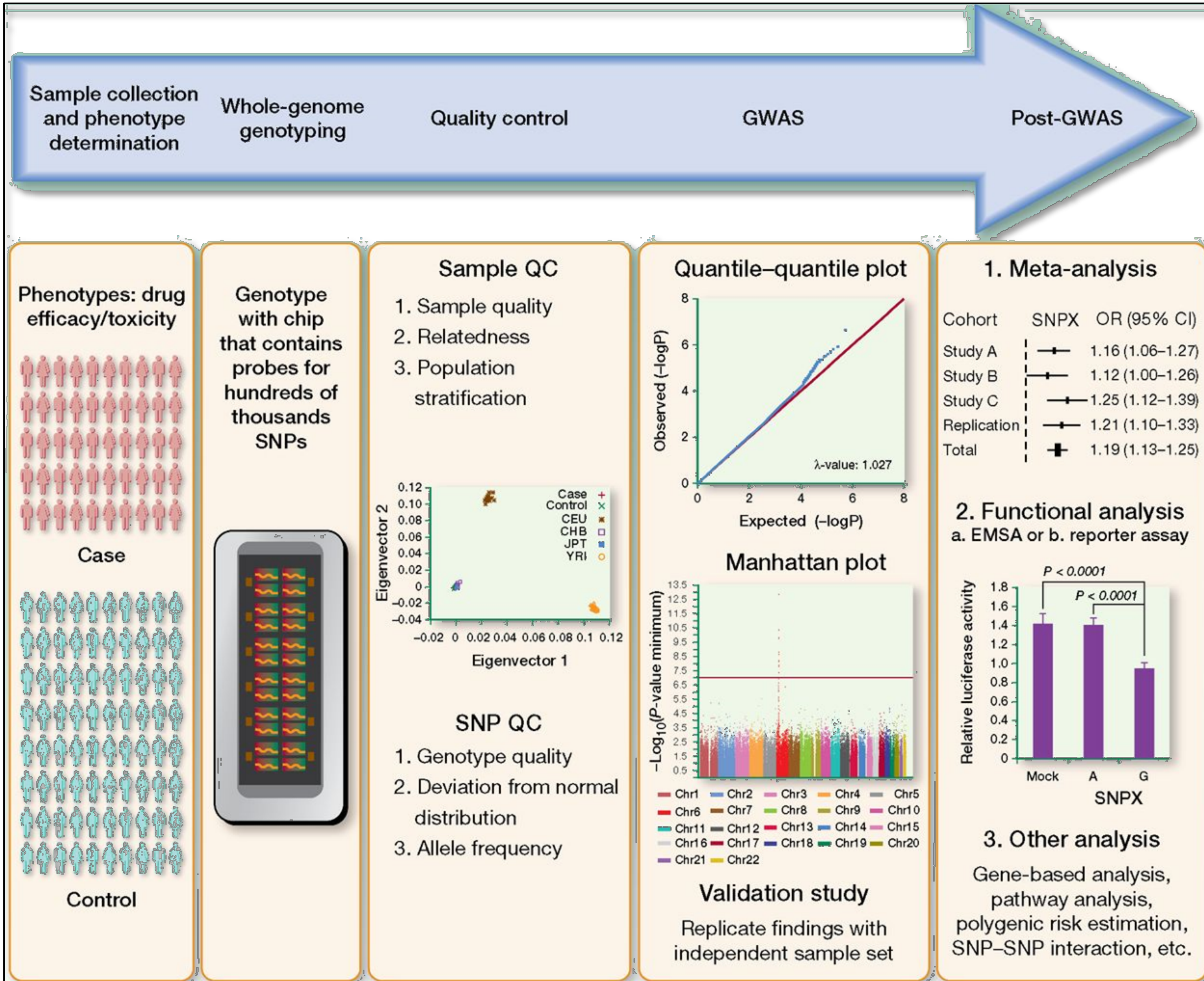
- Exploit human STR marker maps, to search for co-inheritance between STR markers and disease inheritance patterns
- Search for such co-segregation in many independent families (each carries a diff. mutation, but all map to the same region)
- Led to mapping of many Mendelian disease loci in human

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

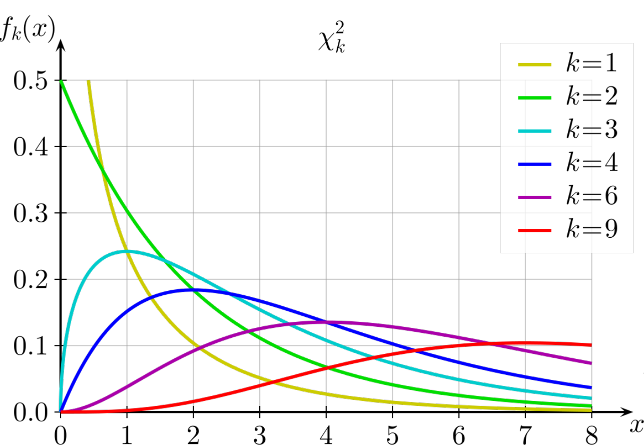
rs11209026	Allele A	Allele G
Cases	22	976
Controls	68	932

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

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



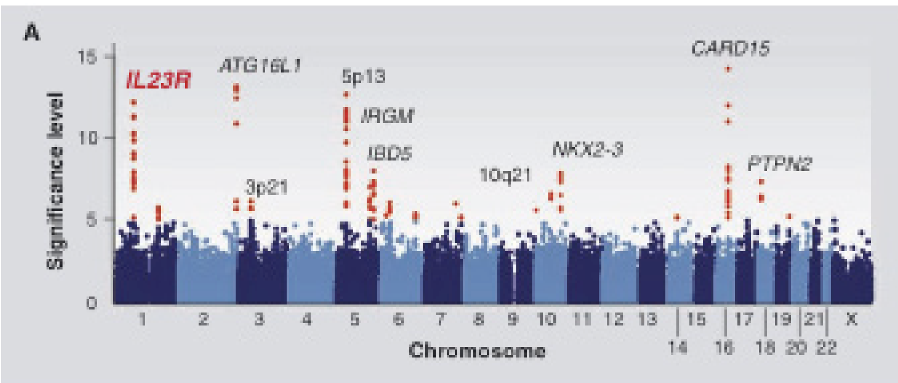
Simplest tests (single marker regression, χ^2) rule the day.
Association results requiring arcane statistics.
Complex multi-marker models are often less reliable

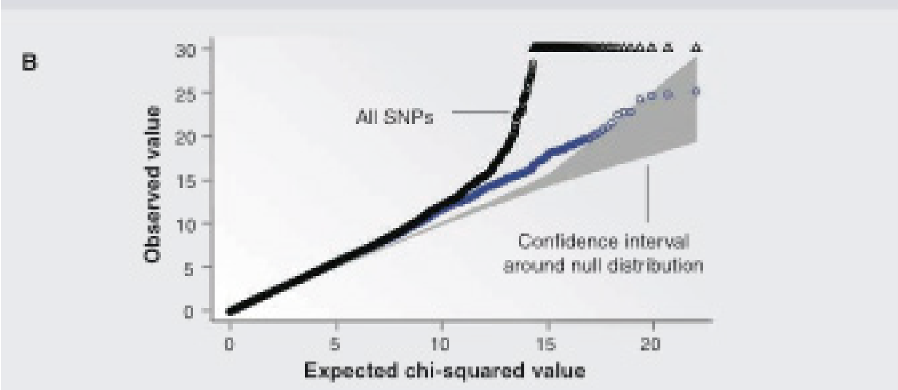
Multiple Testing

- 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 $p=5\times10^{-8}$ (= .05 / 1 million tests)

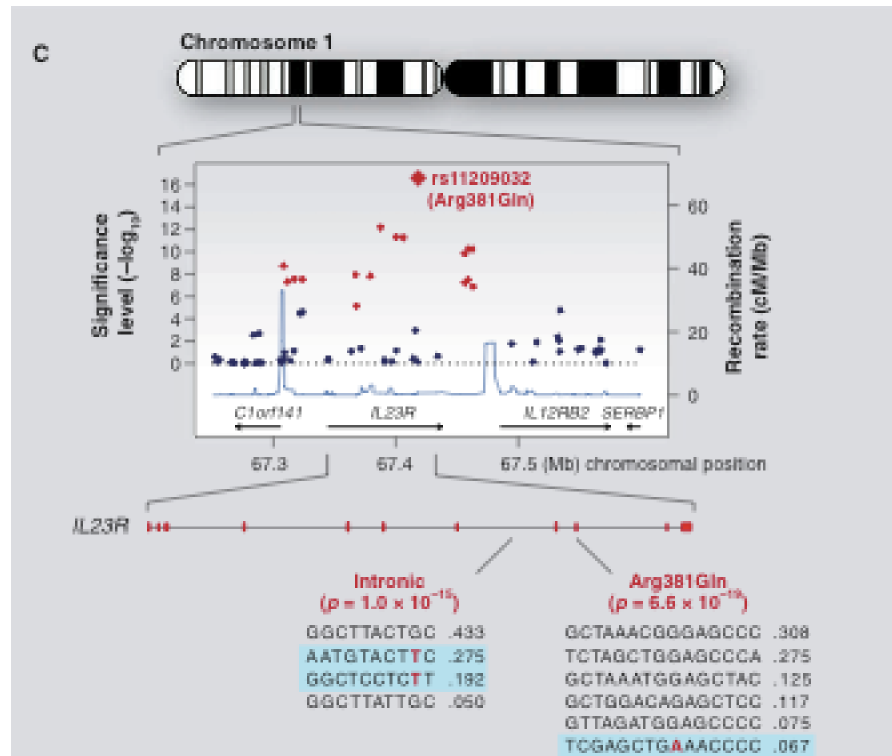
Genome-wide Association

‘Manhattan’ plot

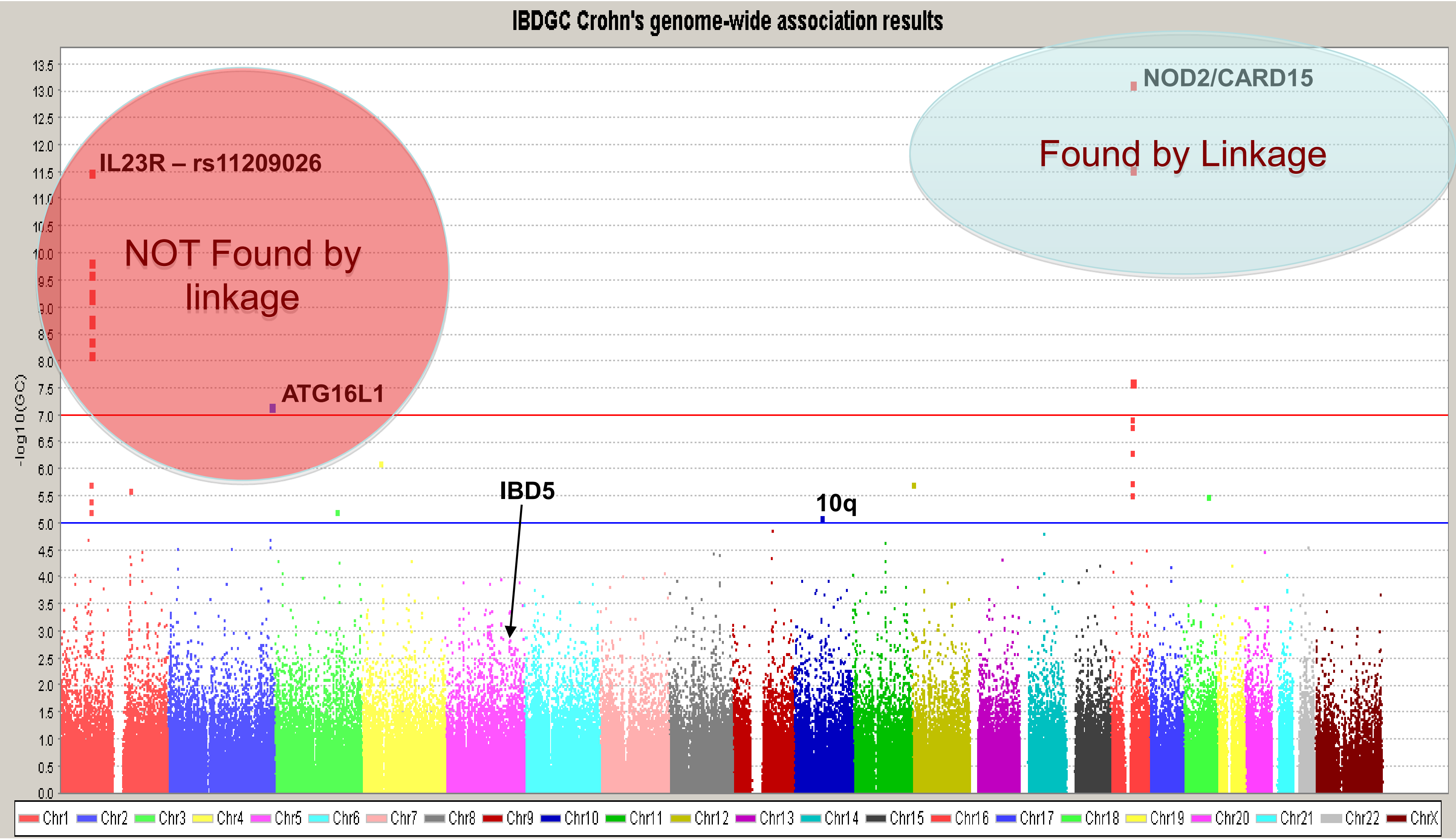




Q-Q plot



Search for gene / mechanism

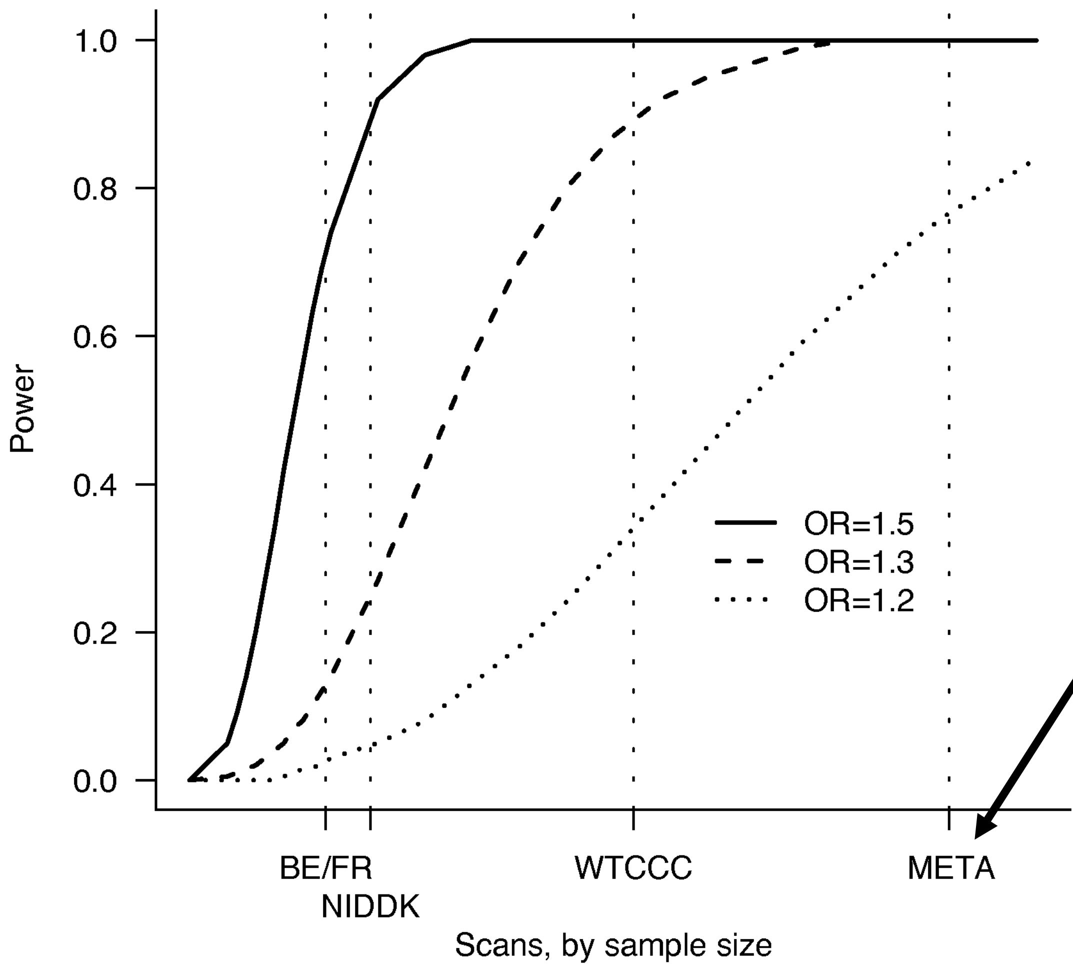

IBDGC Crohn's genome-wide association results
-log10(GC)
13.5
13.0
12.5
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
IL23R – rs11209026
NOT Found by linkage
ATG16L1
NOD2/CARD15
Found by Linkage
IBD5
10q
Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8 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



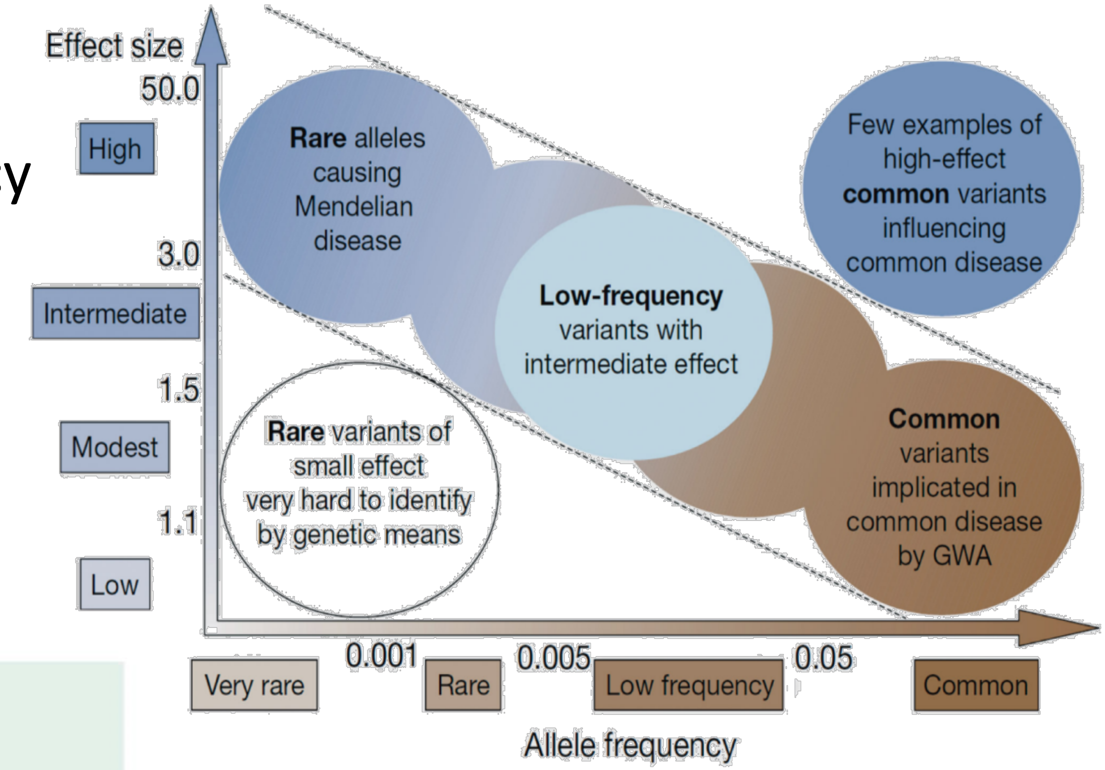
(Example – associated SNP with MAF = 0.20)

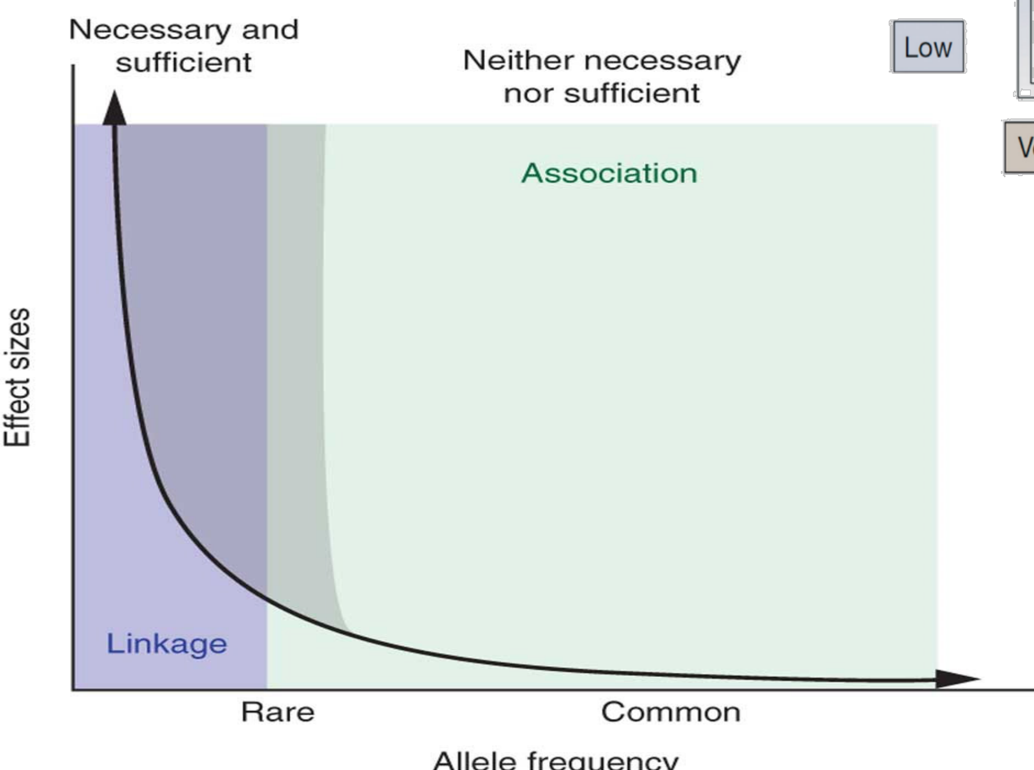
Opportunity: by combining three published studies, we reap the power of an 8000 sample GWAS

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

Common alleles typically have small effects

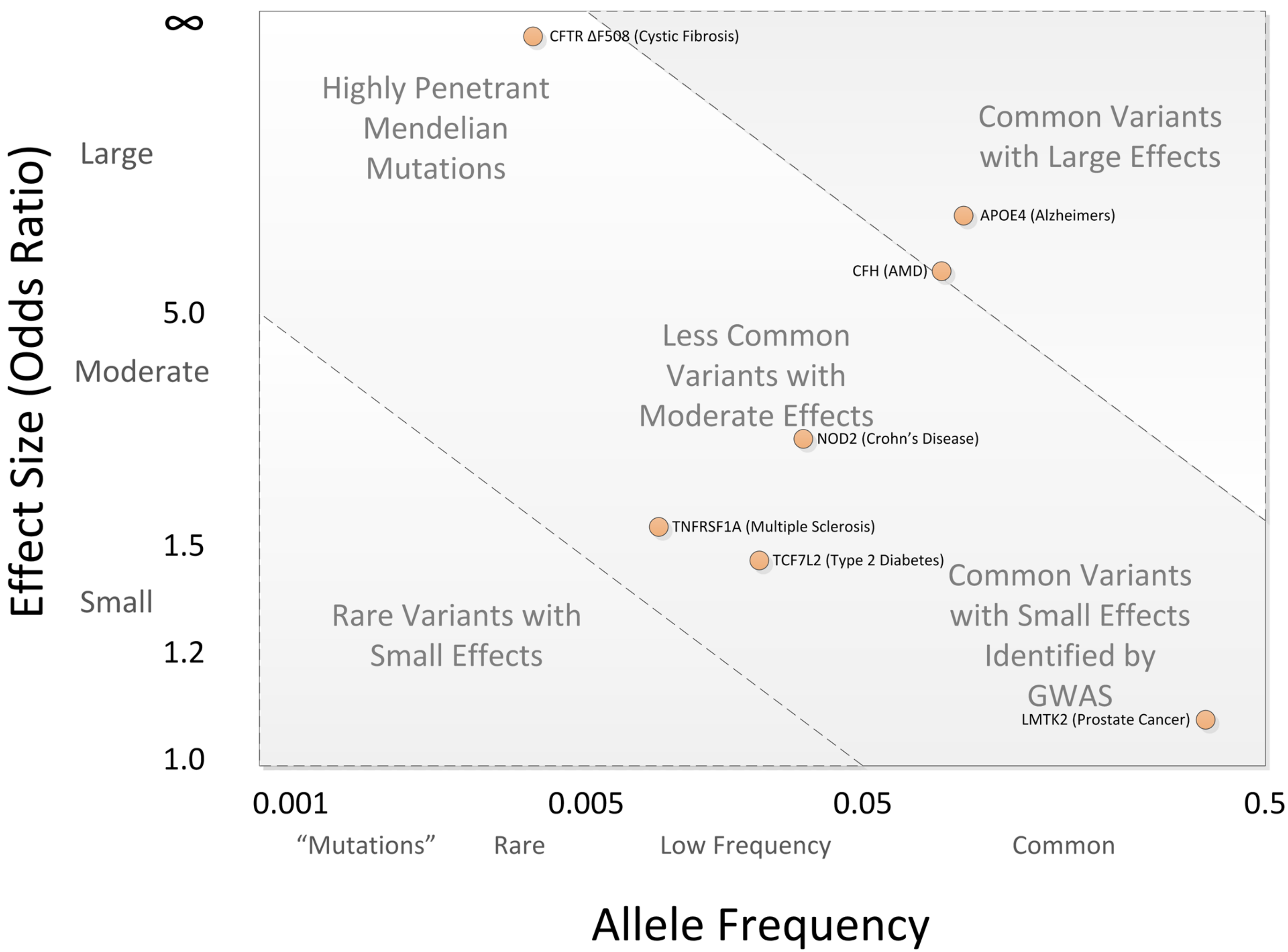
Discovery method tuned to variant effect size/frequency





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

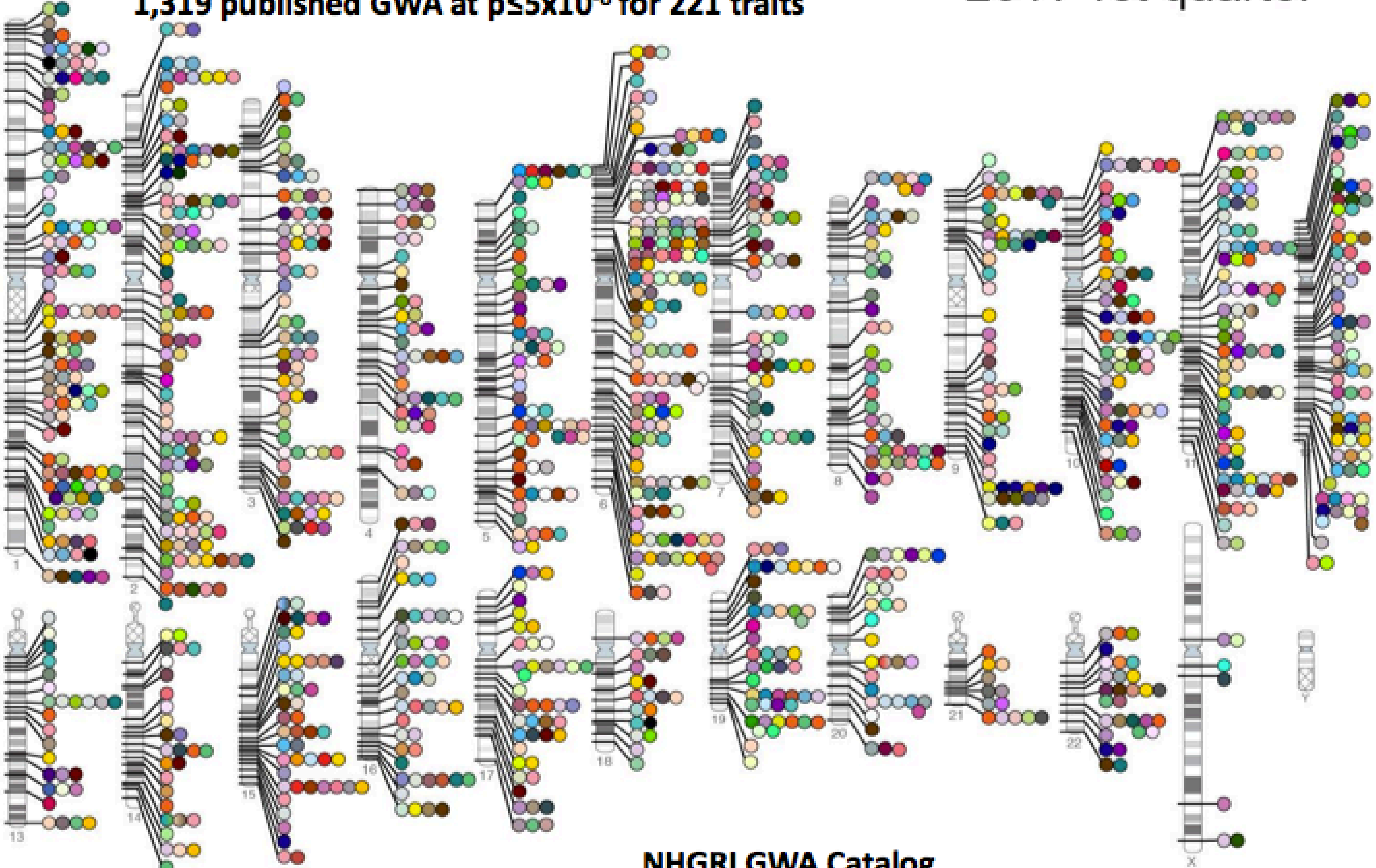
1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting (locus, really): Common/rare alleles, Linkage vs. GWAS
3. Evolution/scaling of GWAS power: Sharing, inflection points
4. Challenge of fine-mapping: Co-inheritance, LD, Haplotypes, Recomb.
5. From locus to mechanism - Case study: FTO and Obesity
6. Challenges in disease mechanism inference
7. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA

3. Evolution and scaling of GWAS power:

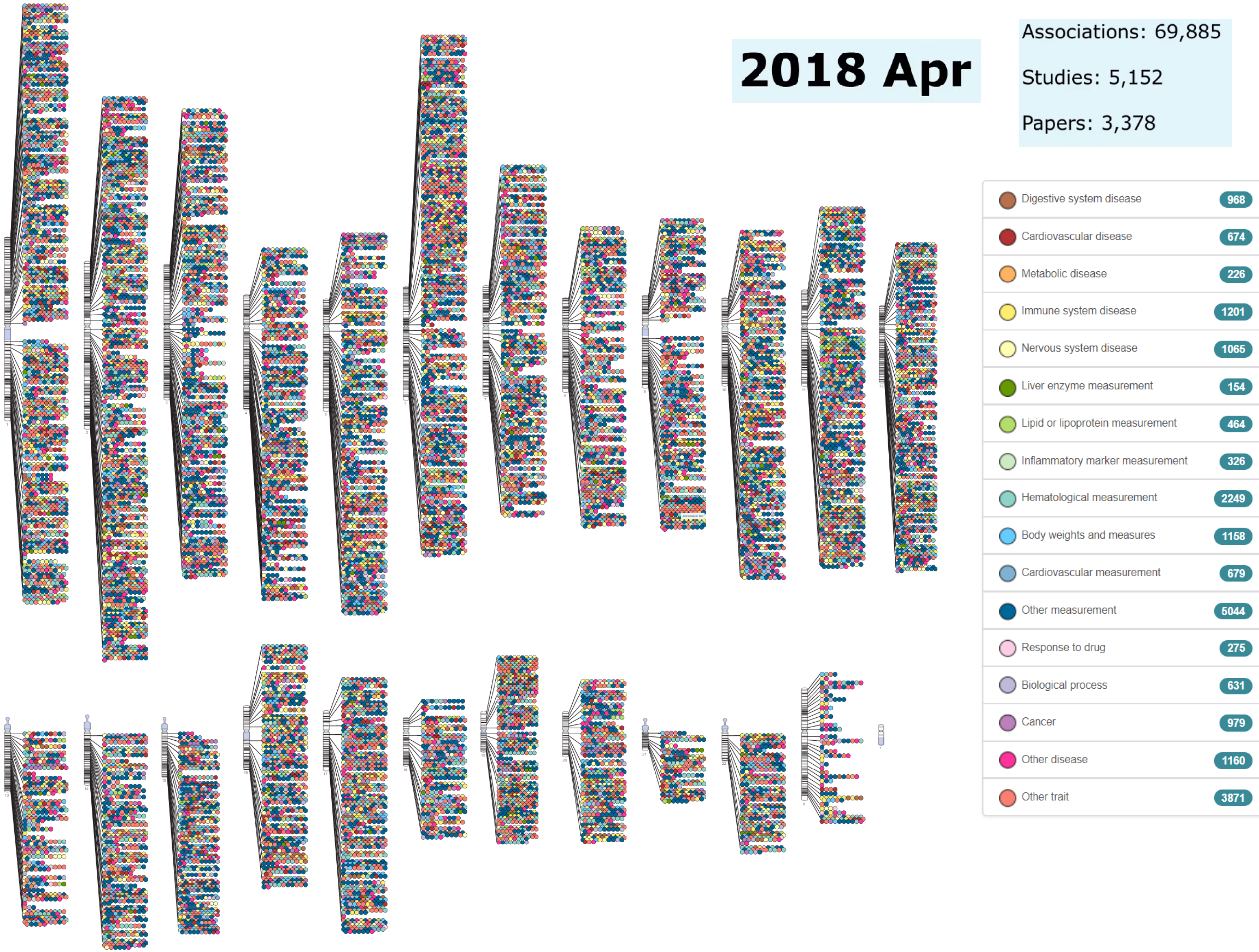
The power of sharing, inflection points

Published Genome-Wide Associations through 03/2011, 1,319 published GWA at $p \leq 5 \times 10^{-8}$ 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

Gene Finding in Crohn's

NIDDK-IBDGC

International IBDGC

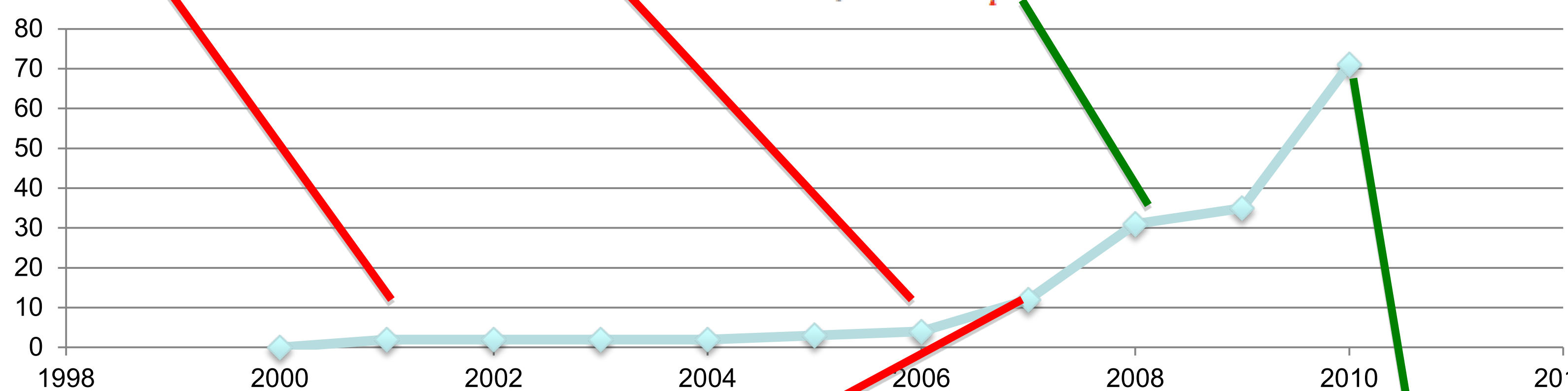
A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene

Richard H. Duerr, Kent D. Taylor, Steven R. Brant, John D. Rioux, Mark S. Silverberg, Mark J. Daly, A. Hillary Steinhart, Clara Abraham, Miguel Regueiro, Anne Griffiths,
Themistocles Dassopoulos, Alain Bitton, Huiying Yang, Stephan Targan,
Lisa Wu Datta, Emily O. Kistner, L. Philip Schumm, Annette T. Lee,
Peter K. Gregersen, M. Michael Barmada, Jerome I. Rotter,
Dan L. Nicolae, Judy H. Cho*

Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease

Jeffrey C Barrett*[1], Sarah Hansoul[2], Dan L Nicolae[3], Judy H Cho[4], Richard H Duerr[5,6], John D Rioux[7,8], Steven R Brant[9,10], Mark S Silverberg[11], Kent D Taylor[12], M Michael Barmada[5], Alain Bitton[13], Themistocles Dassopoulos[9], Lisa Wu Datta[9], Todd Green[8], Anne M Griffiths[14], Emily O Kistner[15], Michael T Murtha[4], Miguel D Regueiro[6], Jerome I Rotter[12], L Philip Schumm[15], A Hillary Steinhart[11], Stephan R Targan[12], Ramnik J Xavier[16], the NIDDK IBD Genetics Consortium[33], Cécile Libioulle[2], Cynthia Sandor[2], Mark Lathrop[17], Jacques Belaiche[18], Olivier Dewit[19], Ivo Gut[17], Simon Heath[17], Debby Laukens[20], Myriam Mni[2], Paul Rutgeerts[21], André Van Gossum[22], Diana Zelenika[17], Denis Franchimont[22], JP Hugot[23], Martine de Vos[20], Severine Vermeire[21], Edouard Louis[18], the Belgian-French IBD Consortium[33], the Wellcome Trust Case Control Consortium[33], Lon R Cardon[1], Carl A Anderson[1], Hazel Drummond[24], Elaine Nimmo[24], Tariq Ahmad[25], Natalie J Prescott[26], Clive M Onnie[26], Sheila A Fisher[26], Jonathan Marchini[27], Jilur Ghori[28], Suzannah Bumpstead[28], Rhian Gwillam[28], Mark Tremelling[29], Panos Deloukas[28], John Mansfield[30], Derek Jewell[31], Jack Satsangi[24], Christopher G Mathew[26], Miles Parkes[29], Michel Georges[2] & Mark J Daly[8,32]

Pre-GWAS: NOD2, 5q31 identified by consortium members



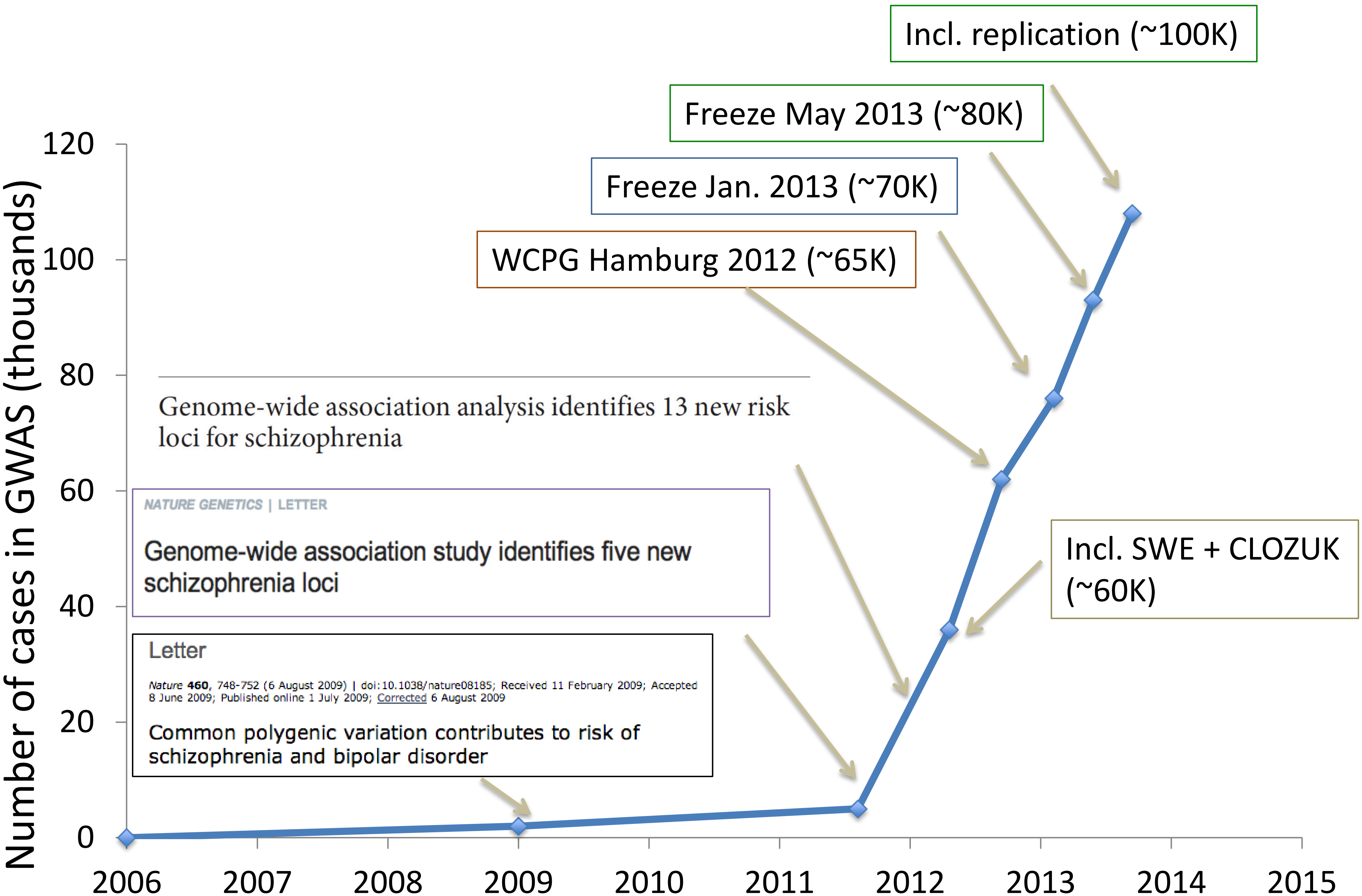
Genome-wide association study identifies new susceptibility loci for Crohn's disease and implicates autophagy in disease pathogenesis

John D Rioux, Ramnik J Xavier, Kent D Taylor, Mark S Silverberg, et al.

Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci

Andre Franke[1,70*], Dermot P B McGovern[2,3,70], Jeffrey C Barrett[4,70], Kai Wang[5], Graham L Radford-Smith[6], Tariq Ahmad[7], Charlie W Lees[8], Tobias Balschun[9], James Lee[10], Rebecca Roberts[11], Carl A Anderson[4], Joshua C Bis[12], Suzanne Bumpstead[4], David Ellinghaus[1], Eleonora M Festen[13], Michel Georges[14], Todd Green[15], Talin Haritunians[3], Luke Jostins[4], Anna Latiano[16], Christopher G Mathew[17], Grant W Montgomery[18], Natalie J Prescott[17], Soumya Raychaudhuri[15], Jerome I Rotter[3], Philip Schumm[19], Yashoda Sharma[20], Lisa A Simms[6], Kent D Taylor[3], David Whiteman[18], Cisca Wijmenga[13], Robert N Baldassano[21]

Inflection point in complex trait GWAS



Schizophrenia GWAS: Number of significant loci

65,000 cases ⇔ 265 loci!

Similar inflection point found in every complex trait!

	Adult height	Crohn's	Schizophrenia
	(per 5000/5000)	(per 1000/1000)	(per 3000/3000)
1x	0	2	1
2x	2	4	2
3x	7	5	6
9x	68	51	62
18x	180	-	-

Significantly associated regions ($p < 5e^{-08}$)

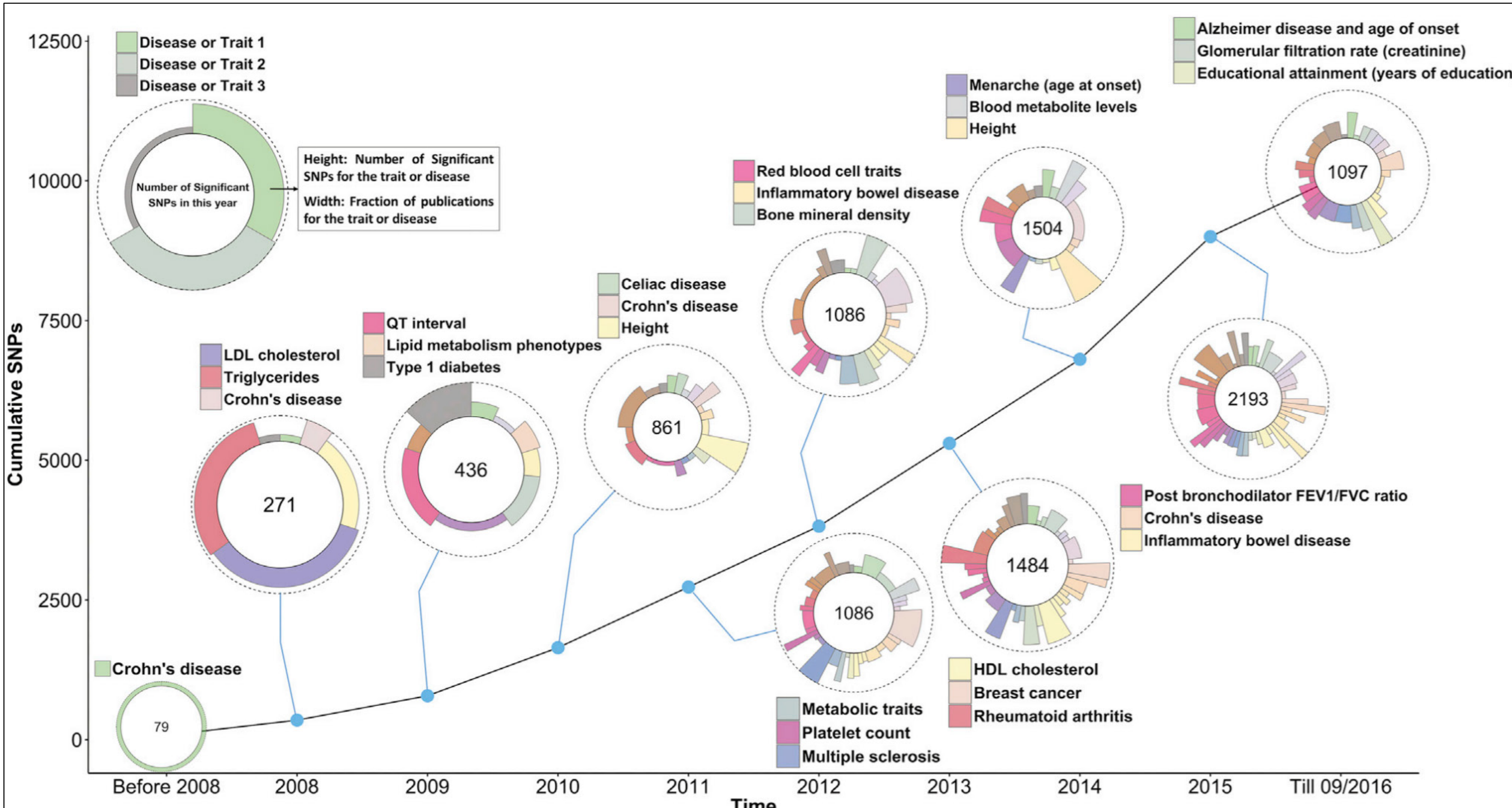
Same story in:

- **Type 1 diabetes**
- **Type 2 diabetes**
- **Serum cholesterol level**
- **Every common chronic disease**

<u>Larger samples lead to new biological insights</u>

- Proof that Schizophrenia is a **heritable, medical disorder**
- **Genetic architecture** similar to non-brain diseases and traits
- Many genes ➔ recognition of **key pathways and processes**
  - Voltage-gated <u>calcium channels</u> (CACNA1C, CACNA1D, CACNA1I, CACNB2)
  - Proteins interacting with FMRP, <u>fragile X</u> gene
  - <u>Neuron</u> organization: Postsynaptic density, dendritic spine heads
  - <u>Enhancers</u>: brain (angular gyrus, inferior temporal lobe), immune

GWAS progress: More traits, more publications, more SNPs



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

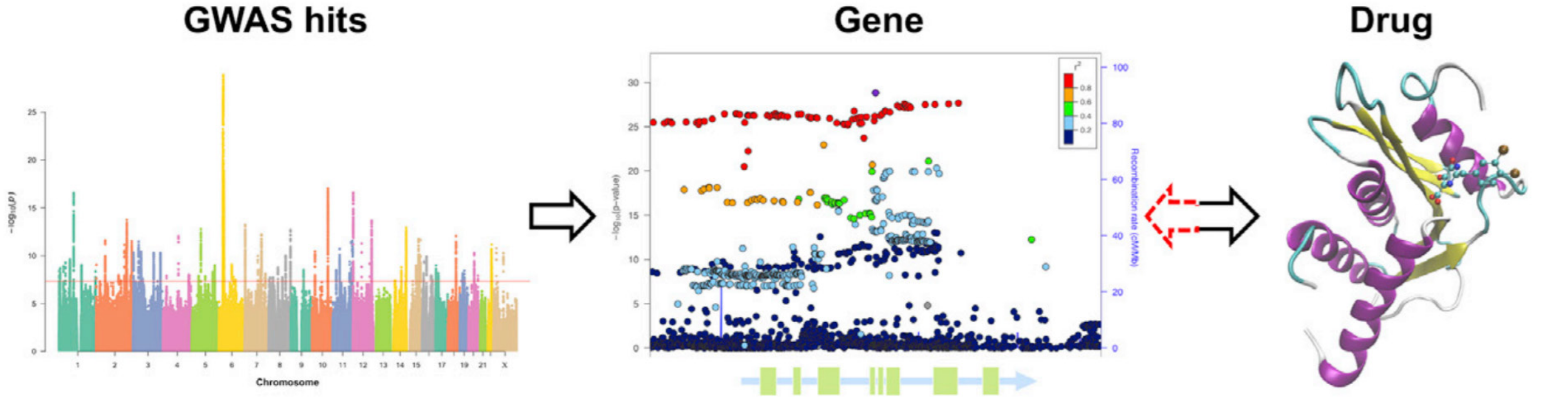
Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. ‘Disease gene’ hunting (locus, really): Common/rare alleles, Linkage vs. GWAS
3. Evolution/scaling of GWAS power: Sharing, inflection points
4. Challenge of fine-mapping: Co-inheritance, LD, Haplotypes, Recomb.
5. From locus to mechanism - Case study: FTO and Obesity
6. Challenges in disease mechanism inference
7. Machine learning tools for variant interpretation
    - Deep variant
    - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
    - DeepSEA

4. The challenge of fine-mapping:

Co-inheritance, LD, Haplotypes, Recombination

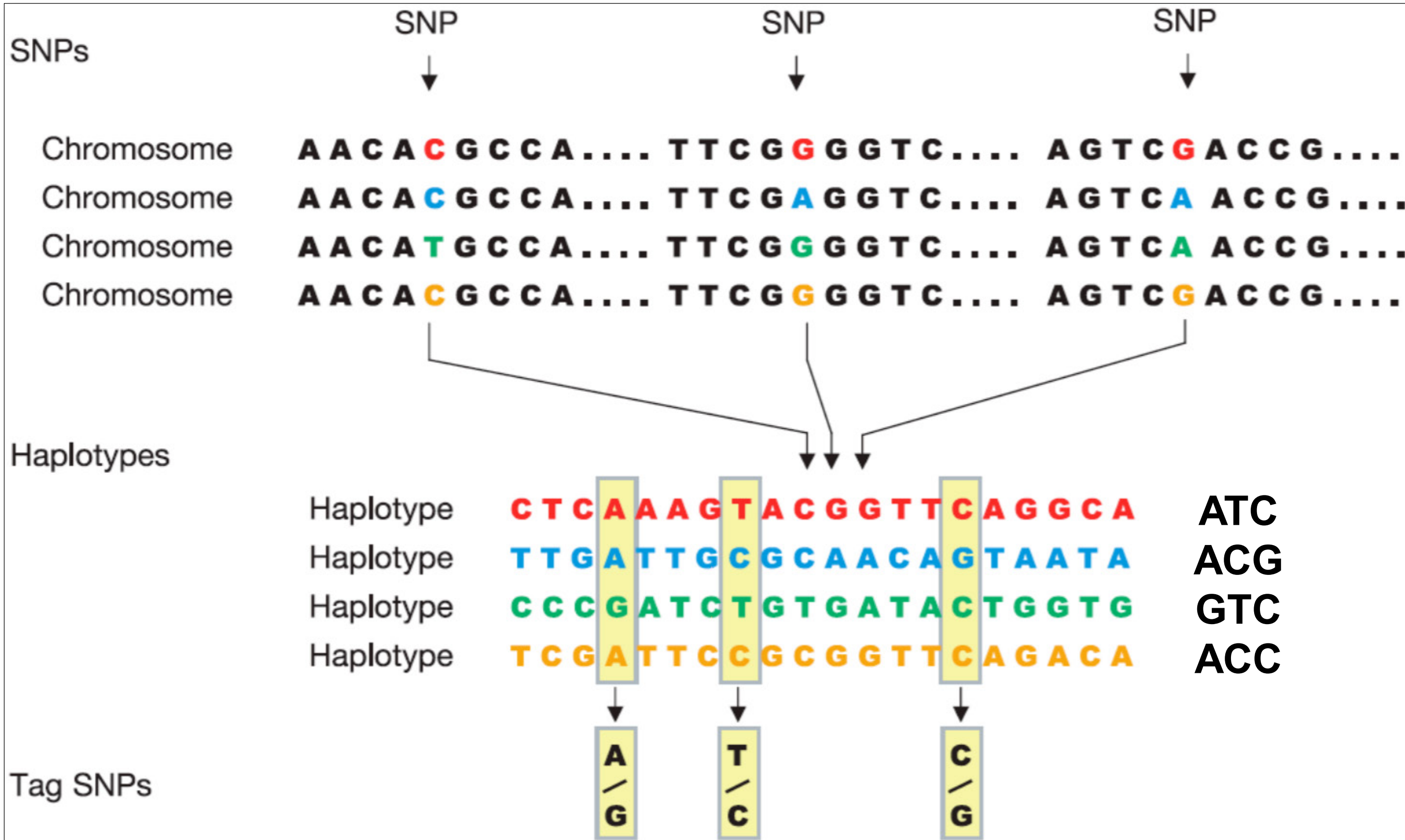
From GWAS hits → to genes → to therapeutics

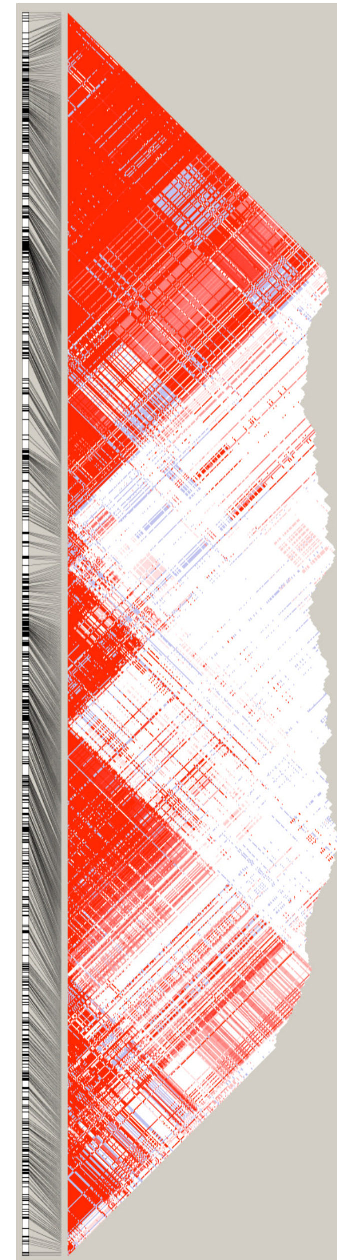


Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*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

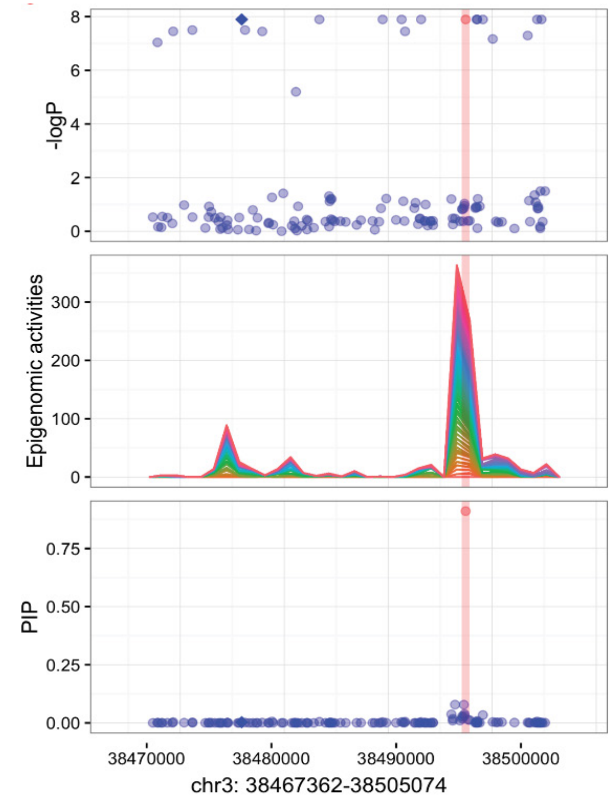


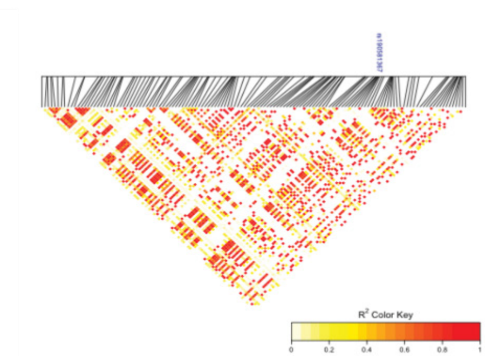


- Common SNPs only once every 1000 nucleotides or so
- These are co-inherited, so only need to profile a subset
- Markers selected for haplotype profiling are “tag” SNPs

Fine-mapping disease associations: Epigenomics / functional data

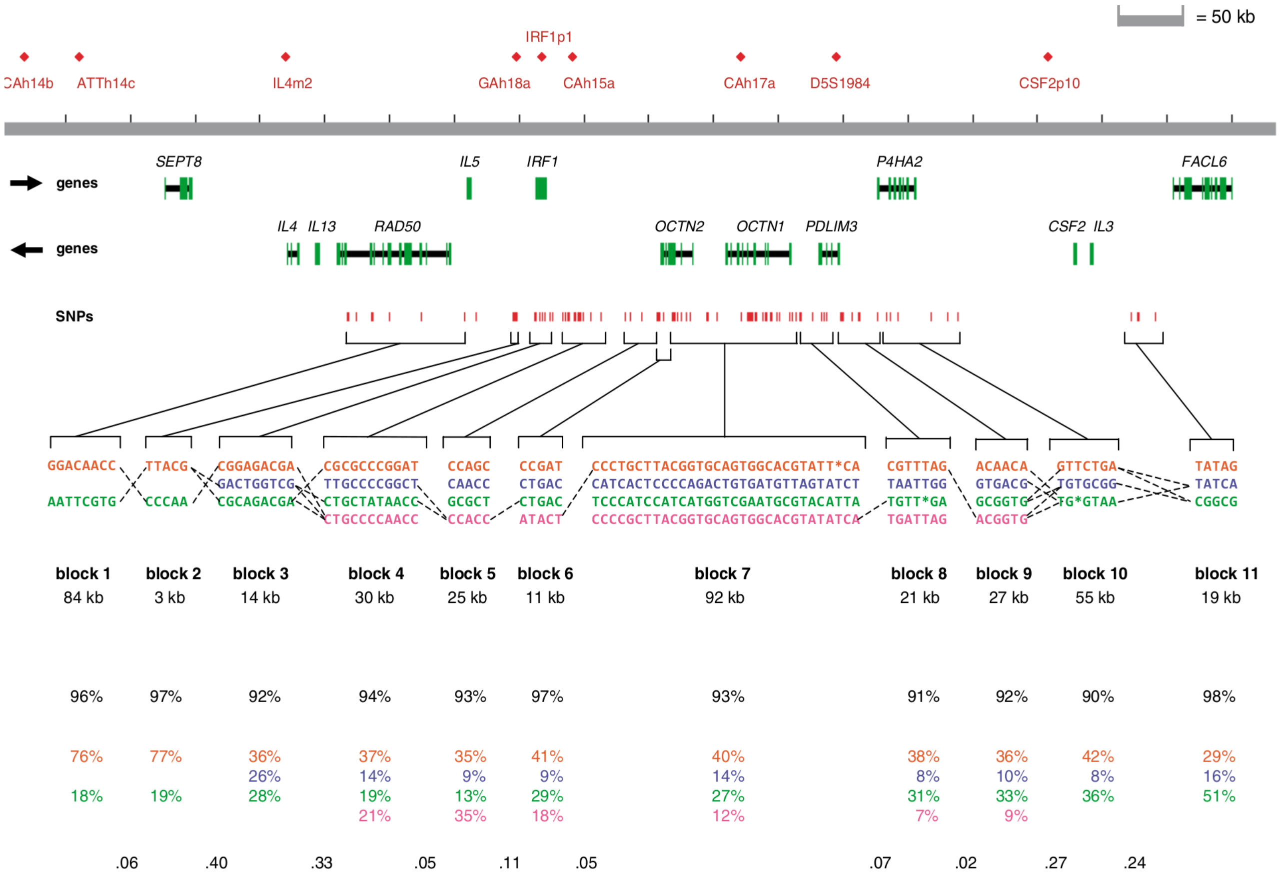
- LD is a **blessing for mapping loci** to disease, as it enables genotyping of just a handful of tag variants
  ➔ enabled the GWAS revolution
- LD is a **curse for fine-mapping** loci into their causal variants
  ➔ many variants are strongly correlated to the true causal variant(s)
  ➔ often indistinguishable 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

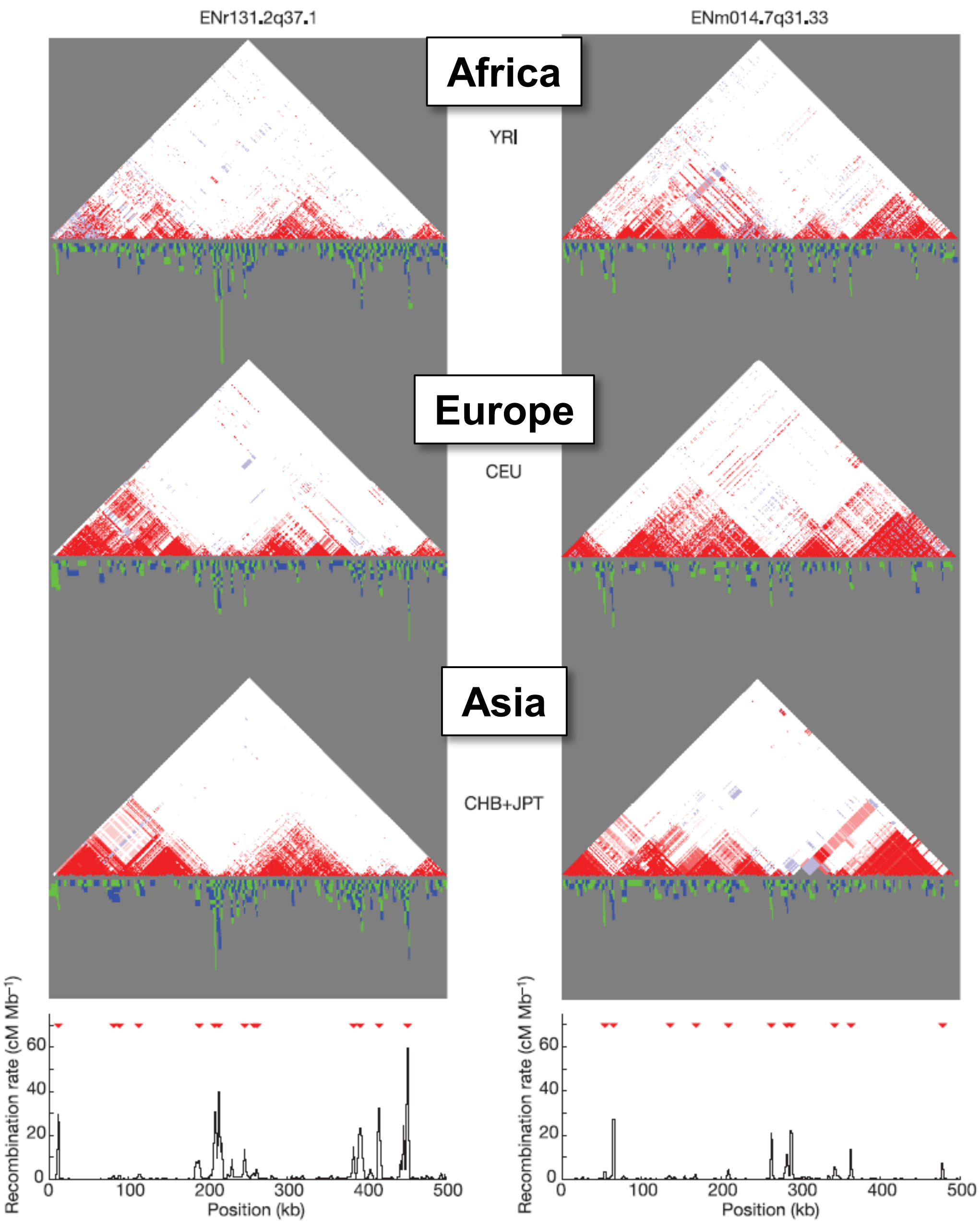
Long-range threading of haplotype blocks



- 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



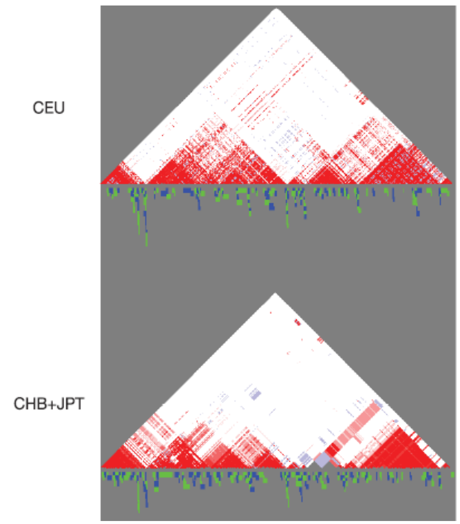
- Recurrent recombination events occur at hotspots

- r^2 correlations between SNPs depend on **historical order** in which they arose

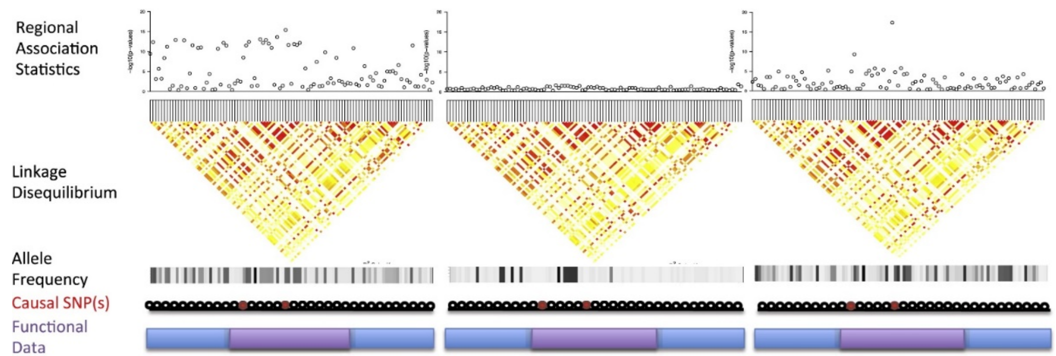
(not in their physical order on the chromosome)

Multi-ethnic analysis can be used for fine-mapping

Case 1: LD boundaries differ



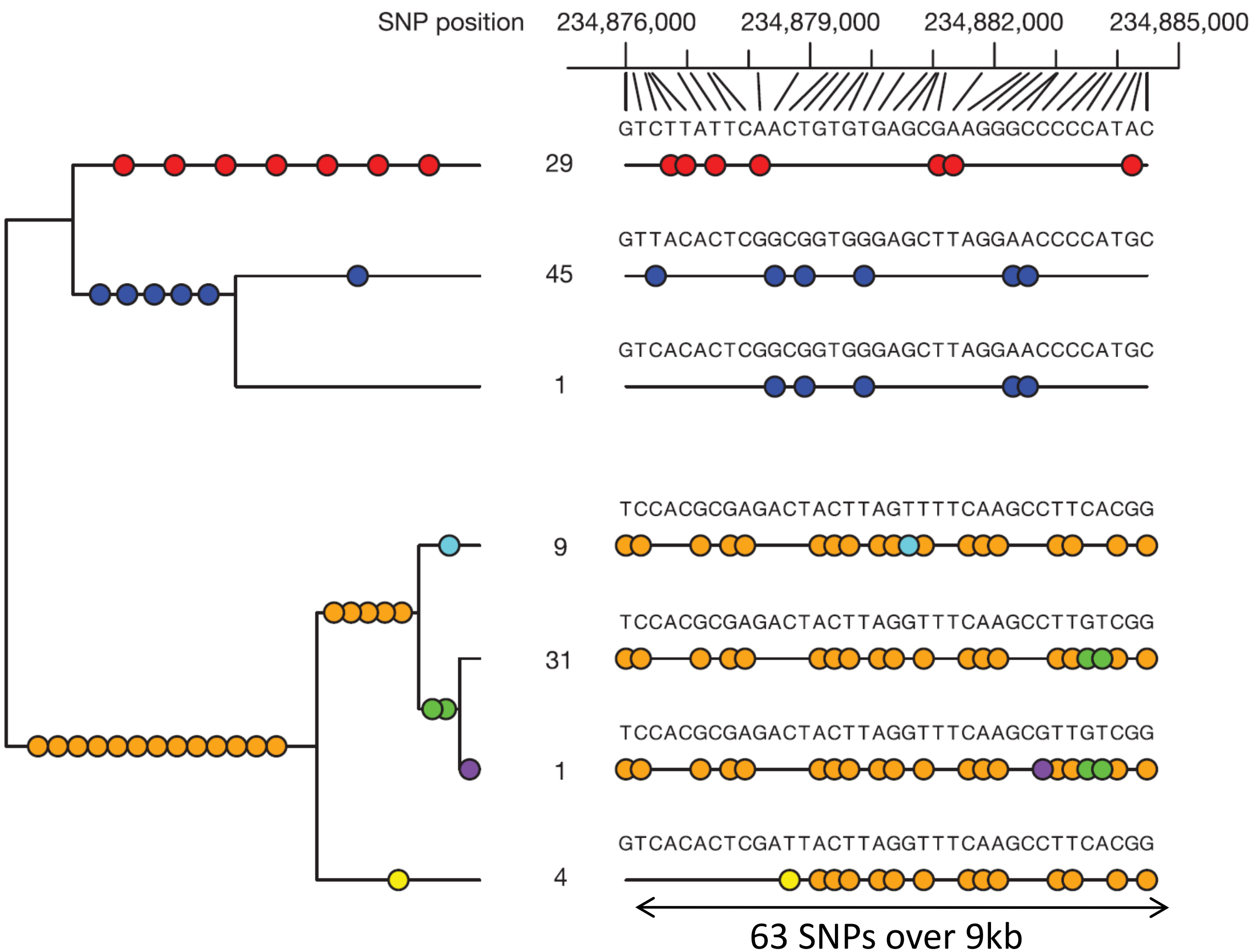
Case 2: allele frequencies differ



- 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

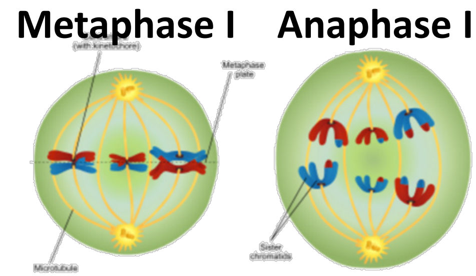
Haplotypes evolve, accumulate mutations

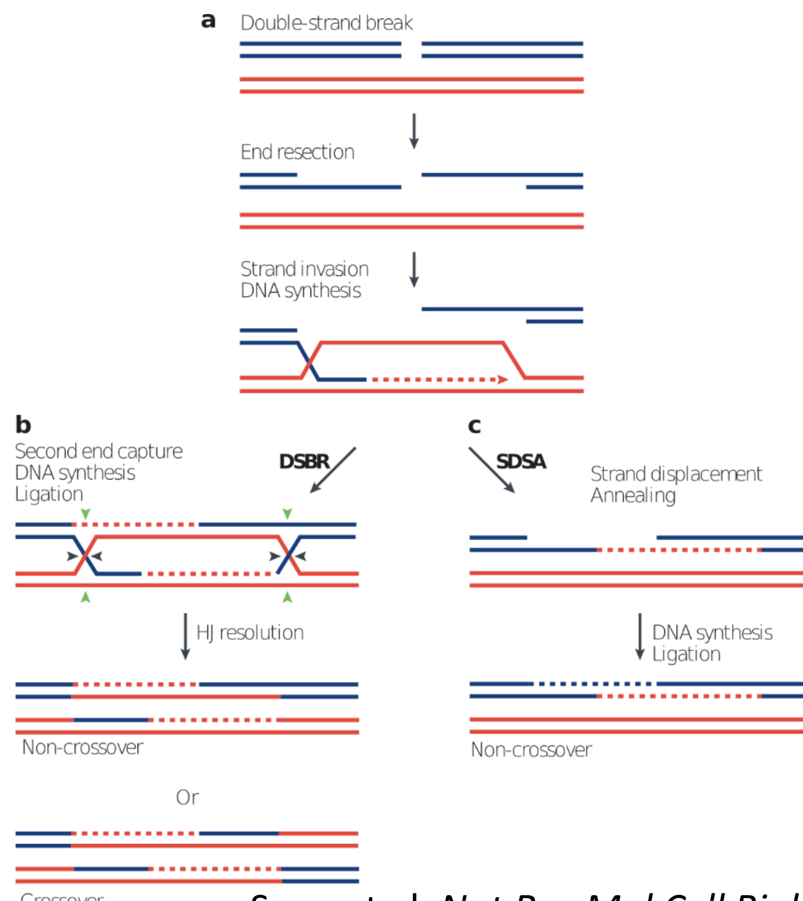


- Example region: 36 SNPs spanning 9kb
- In principle: 2^36 possible allele combinations (haplotypes)
- Sample 120 parental European chromosomes.
- In practice: only 5 recurrent haplotypes seen (and 2 singleton haplotypes)

Haplotypes result from non-uniform recombination

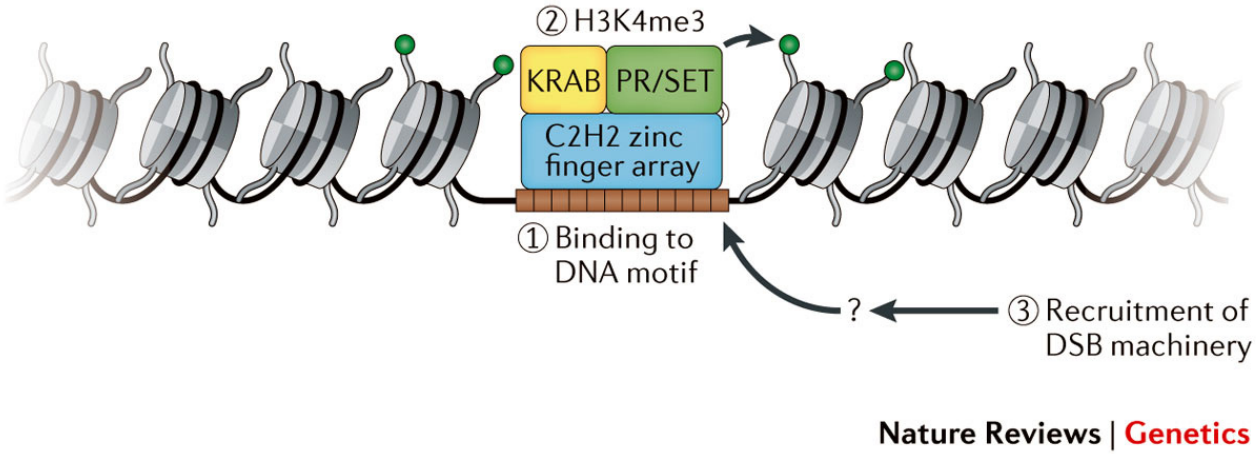
- **Recombination** is crucial for lining up chromosomes during **meiosis** for gamete formation.
- Recombination starts with a **double-stranded break (DSB)**, which is then repaired by strand invasion of the homologous chromosome.
- Repair can lead to either:
  - **Gene conversion**, via strand displacement 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

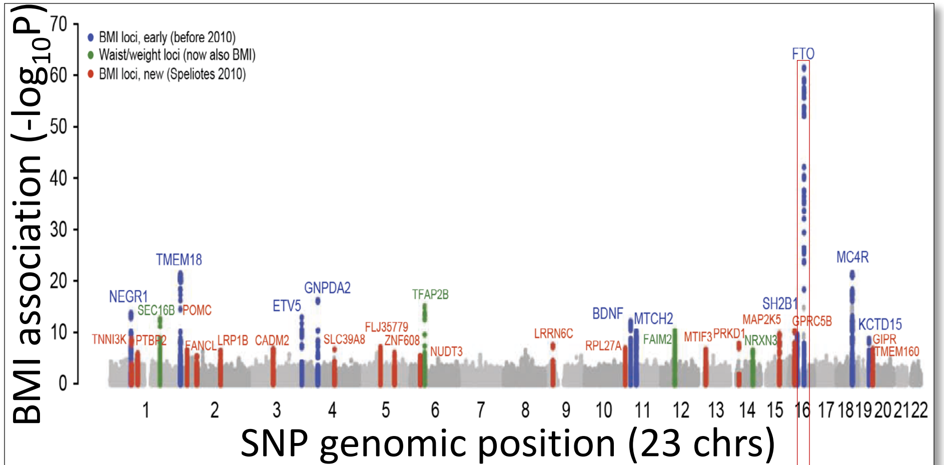
1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting (locus, really): Common/rare alleles, Linkage vs. GWAS
3. Evolution/scaling of GWAS power: Sharing, inflection points
4. Challenge of fine-mapping: Co-inheritance, LD, Haplotypes, Recomb.
5. From locus to mechanism - Case study: FTO and Obesity
6. Challenges in disease mechanism inference
7. Machine learning tools for variant interpretation
    - Deep variant
    - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
    - DeepSEA

5. From locus to mechanism

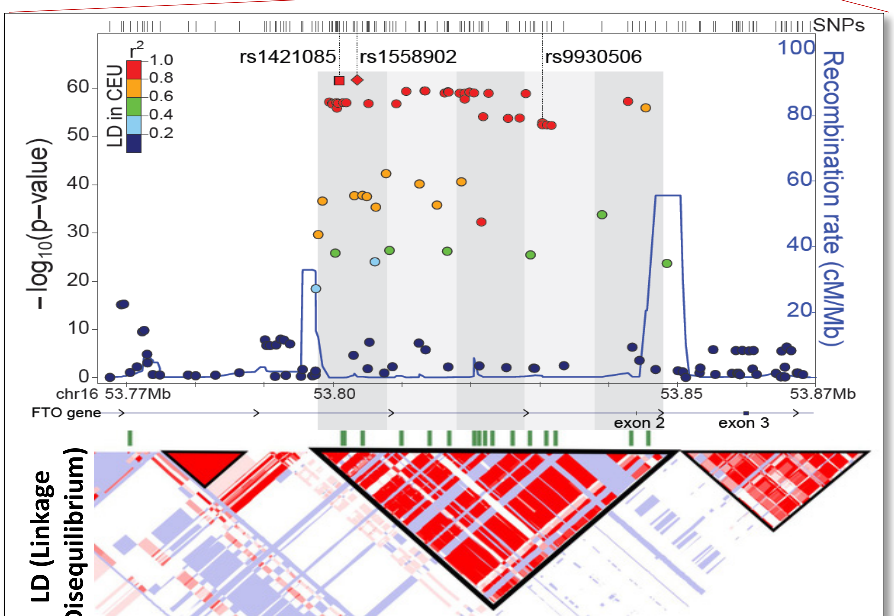
Case study: FTO and Obesity

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

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.

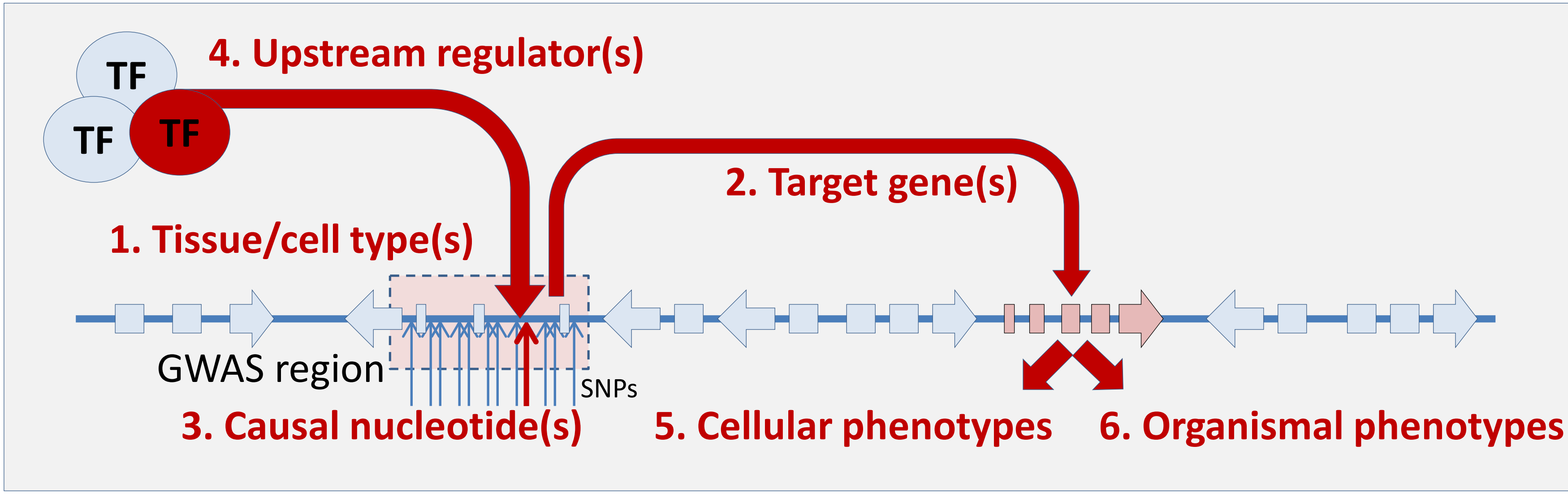
SEPTEMBER 3, 2015 VOL. 373 NO. 10 N Engl J Med 2015;373:895-907.

2. 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

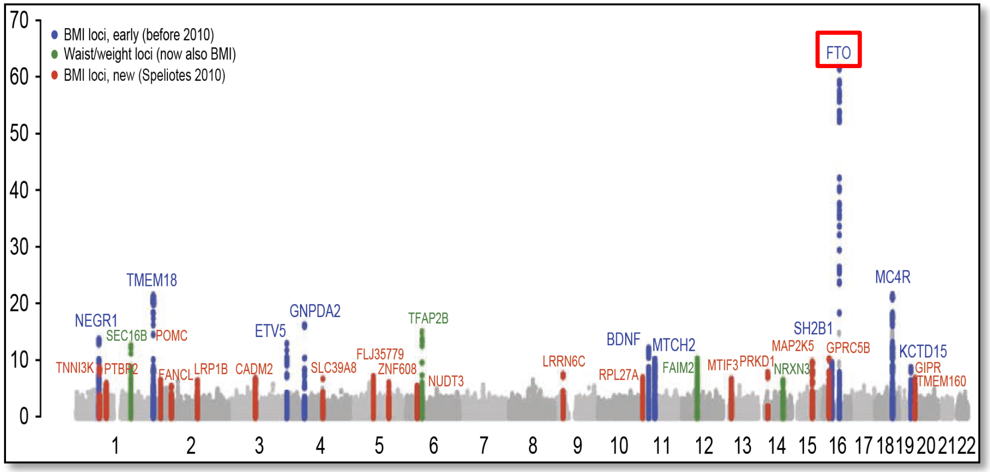


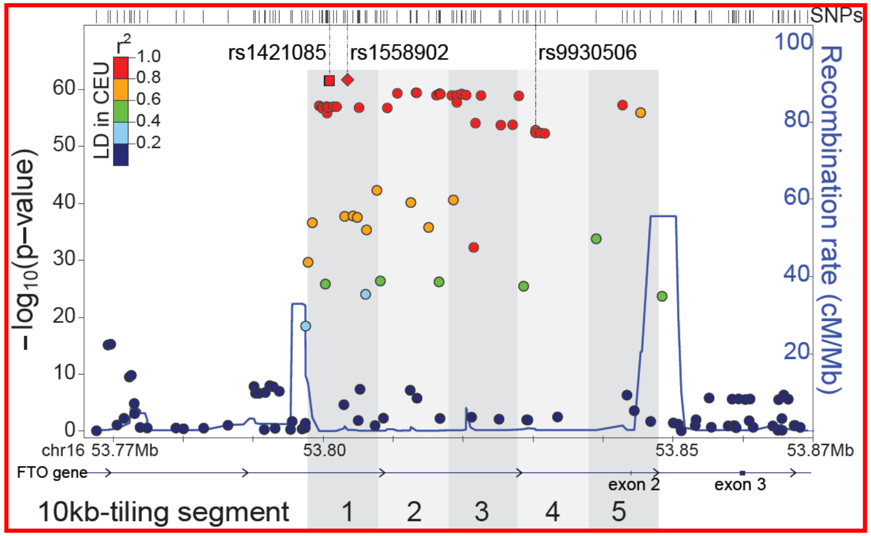
1. Establish relevant **tissue/cell type**
2. Establish downstream **target** gene(s)
3. Establishing **causal** nucleotide variant
4. Establish upstream **regulator** causality
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

<u>Goal:</u>

Apply these to the FTO locus in obesity

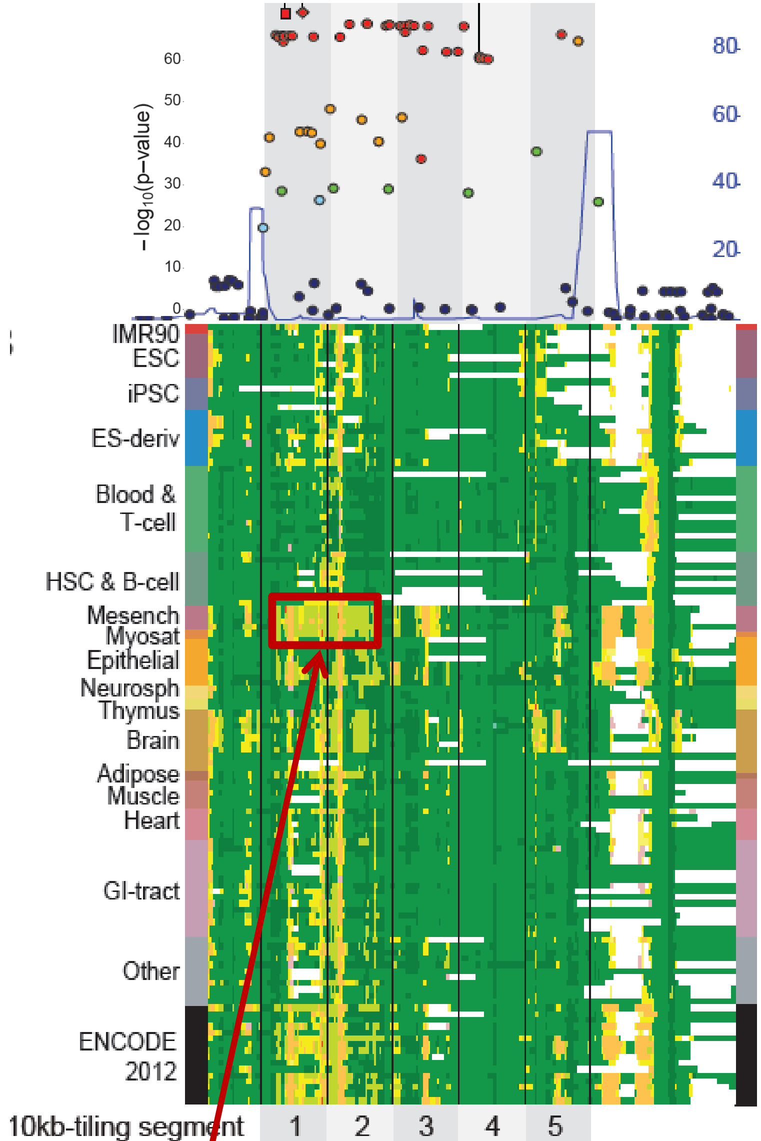
FTO region: strongest association with obesity



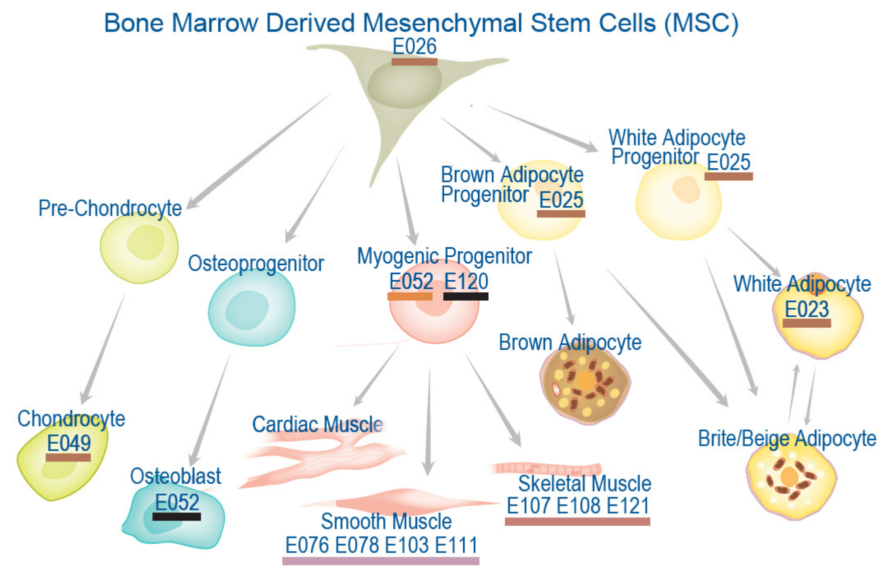


- First and strongest association with obesity (not just ‘your fault’)
- Associated with **obesity**, Type 2 Diabetes, Cardiovascular traits
- 89 variants in LD, spanning 47kb, intron 1 of FTO gene
- No protein-altering variants: regulatory role? Target gene, tissue?

1. Tissue: Chromatin states predict adipocyte function

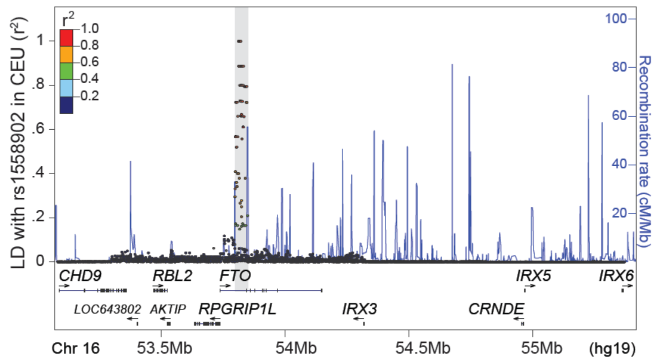


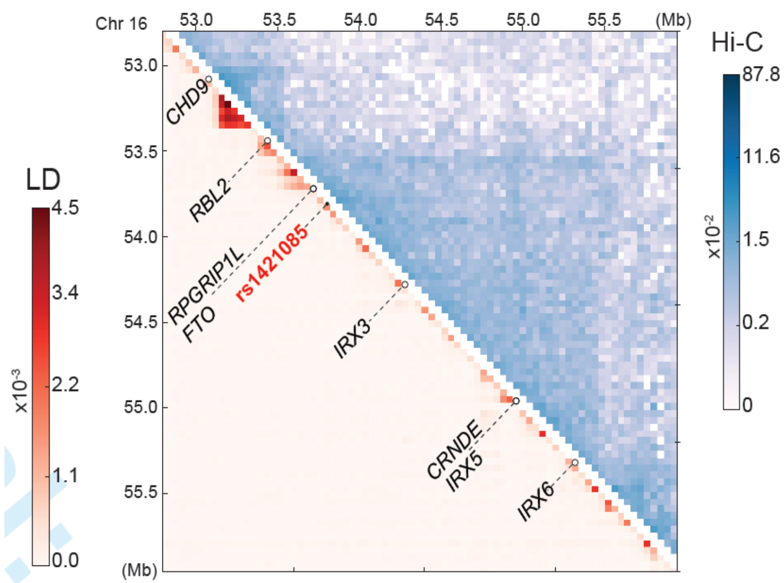
12 kb super-enhancer



Progenitors of white/beige adipocytes

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

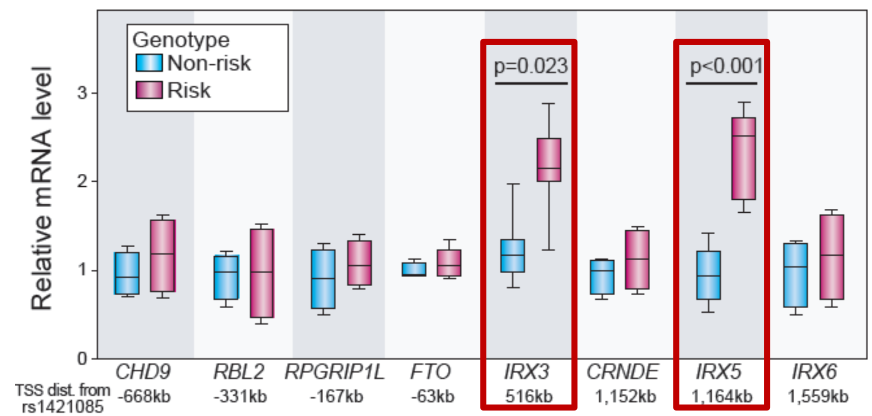




Dixon, Nature 2012

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

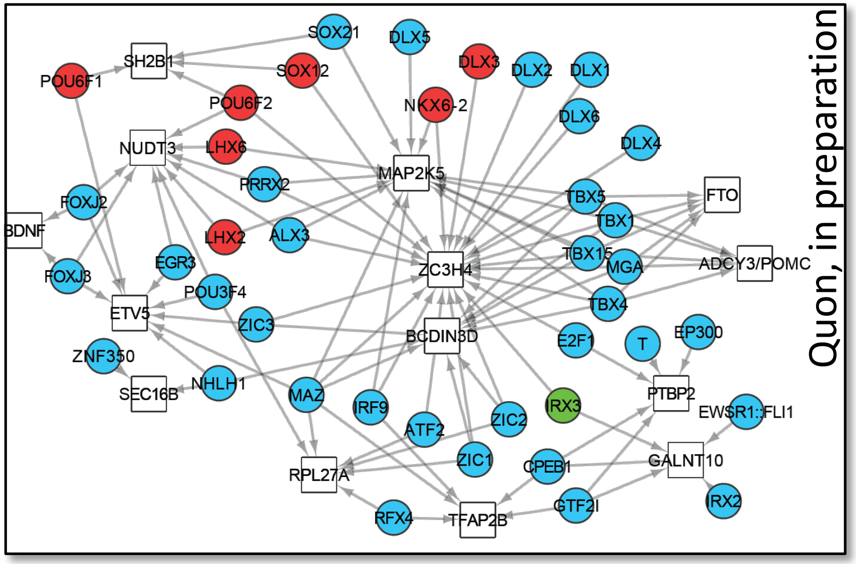
Cohort of **20 homozygous risk** and **18 homozygous non-risk** individuals:
Genotype-dependent expression?

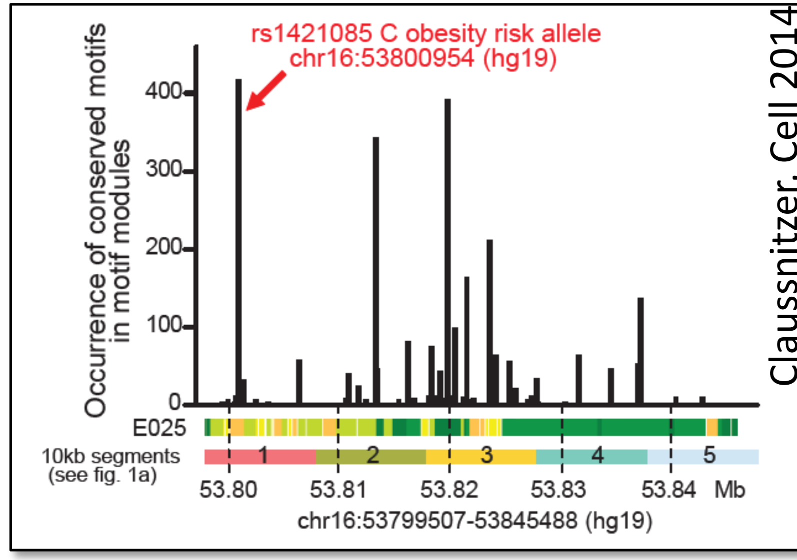


eQTL targets: IRX3 and IRX5

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

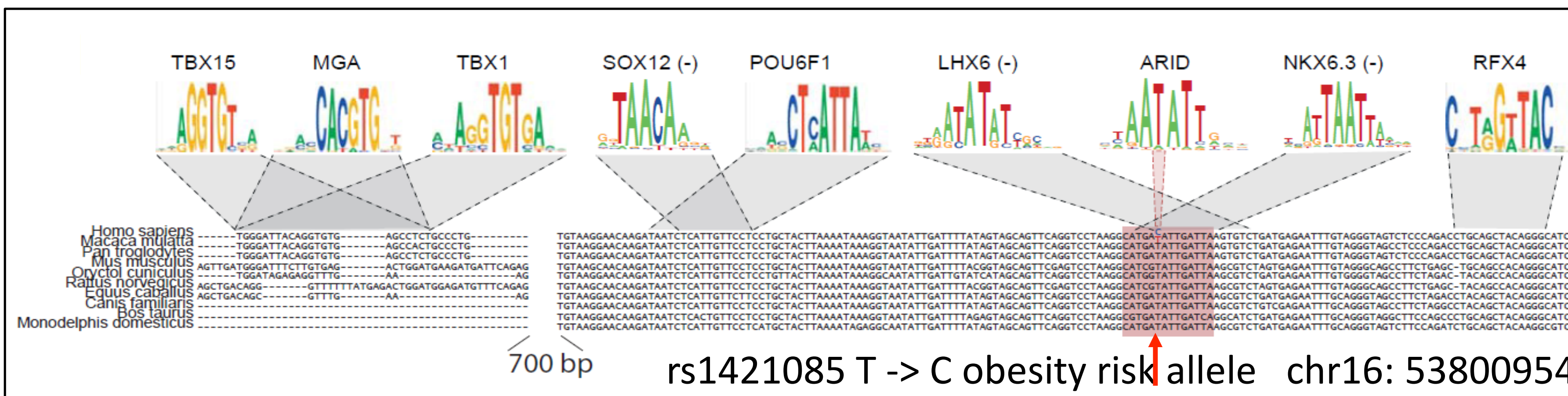
3. Causal SNP: motif enrichment + conservation: rs1421085





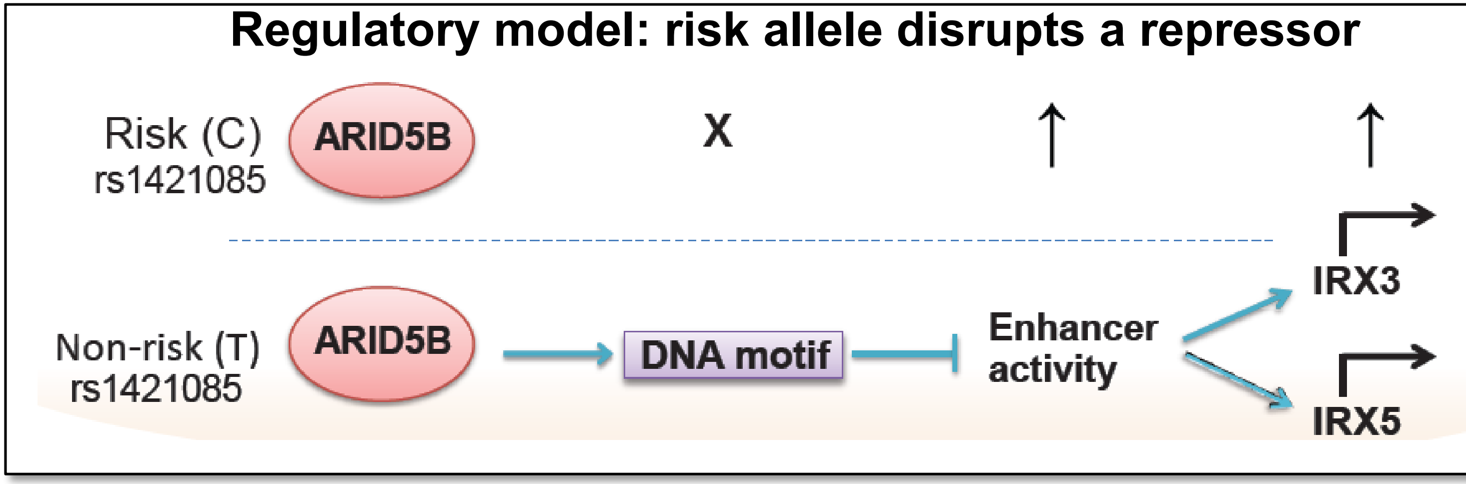
Regulatory motifs enriched in BMI GWAS hits

Regulatory motif combinations conserved across mammals



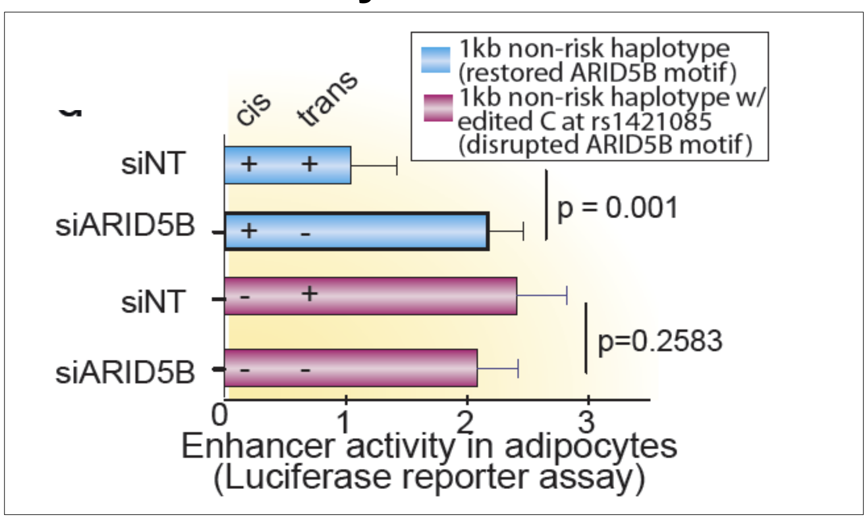
Causal nucleotide rs1421085: risk alters T to C, abolishes AT-rich motif

4. Regulator: Causality and epistasis of ARID5B repressor

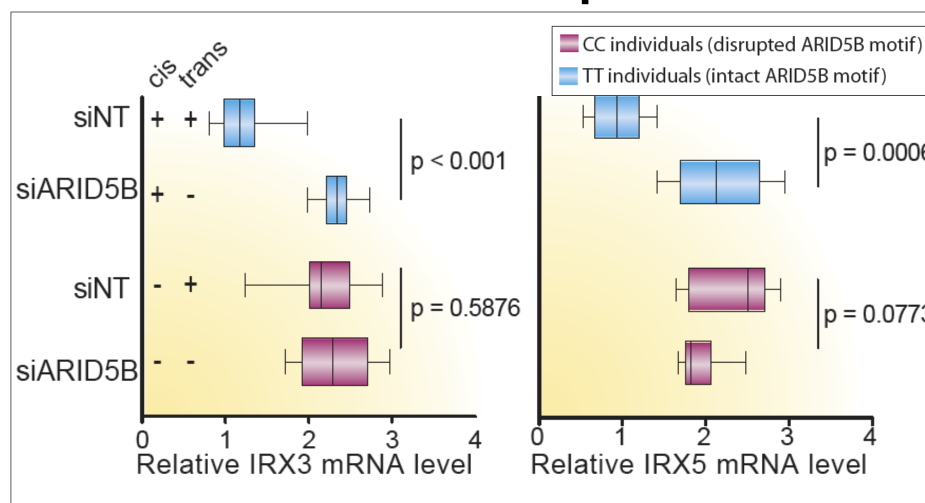


Cis/trans conditional analysis

Enhancer activity

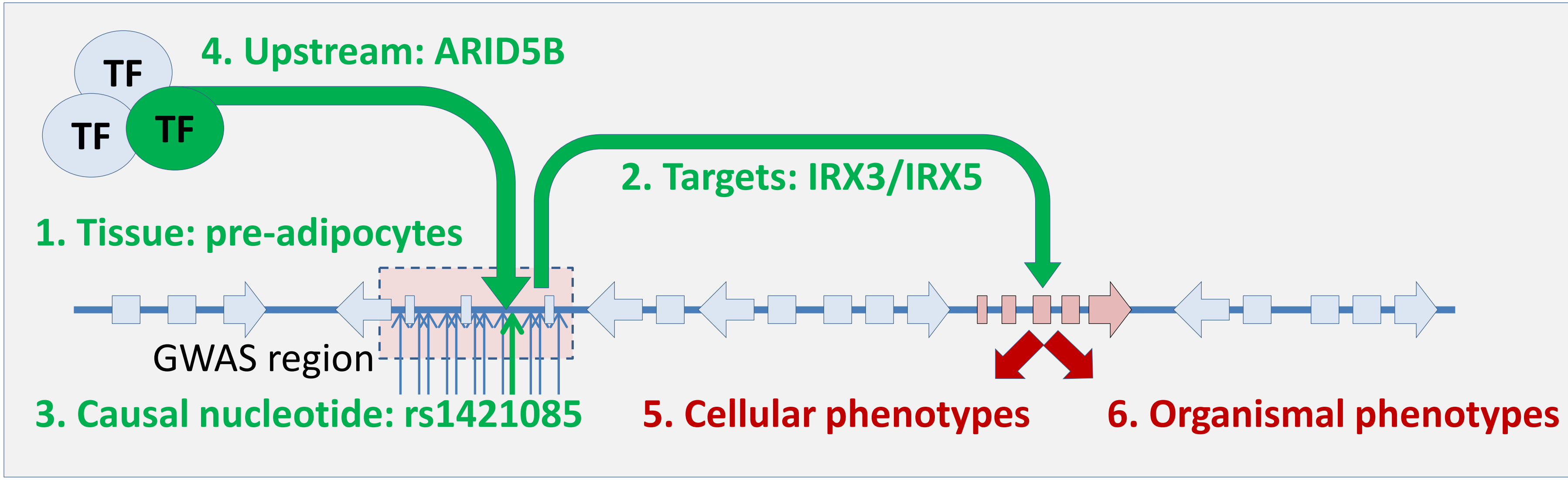


IRX3/5 expression



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

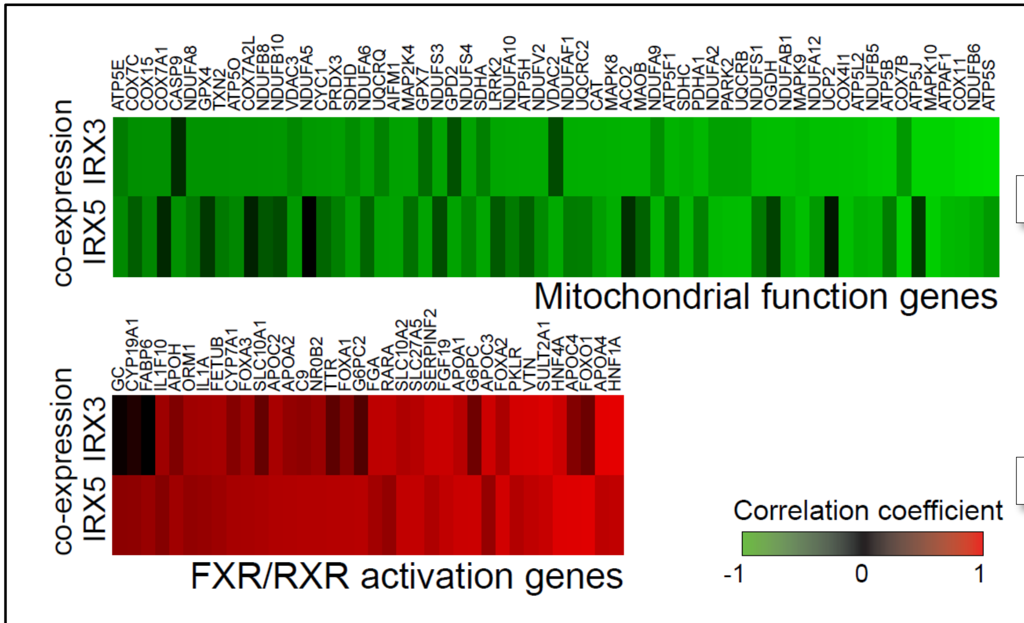
Steps 5-6. Does this circuitry actually lead to obesity?



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

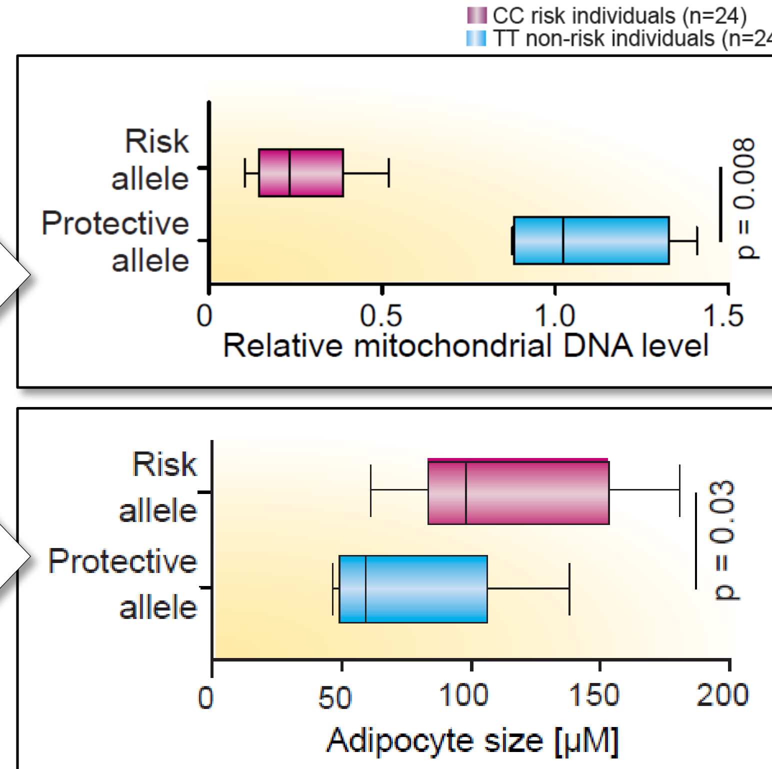
Expression analysis to recognize target processes

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



Negative correlation: mitochondria
Positive correlation: lipid storage

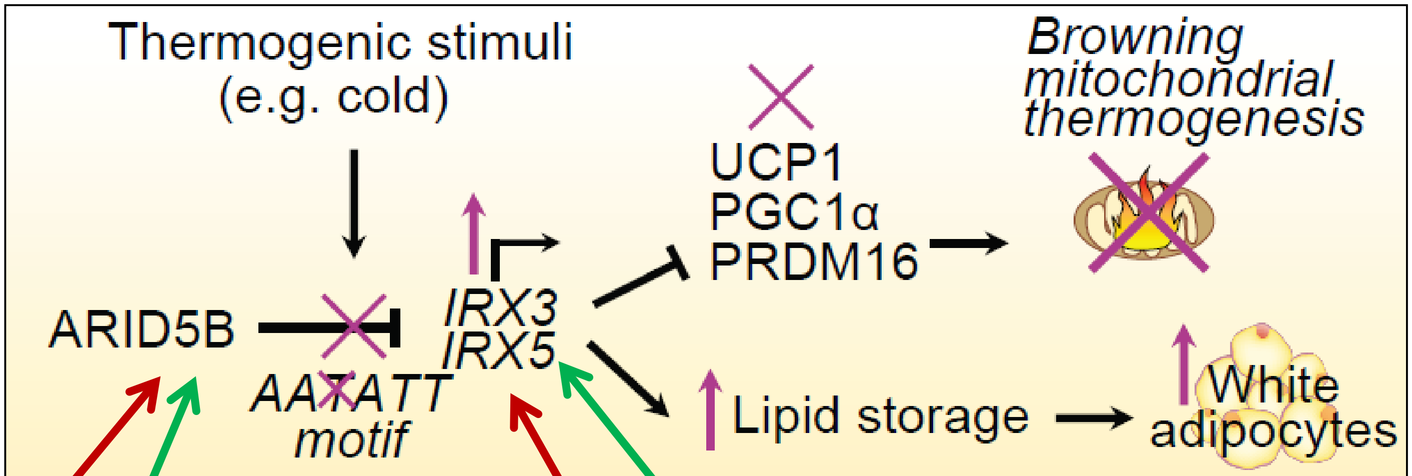
Reflected in cellular phenotypes



Risk carriers: increased mito
Non-risk: increased adipocytes

Risk allele: shift from dissipation to storage

Test model by systematic perturbations



Lean

Obese

ARID5B KD
(obesity)

ARID5B OE
(anti-obesity)

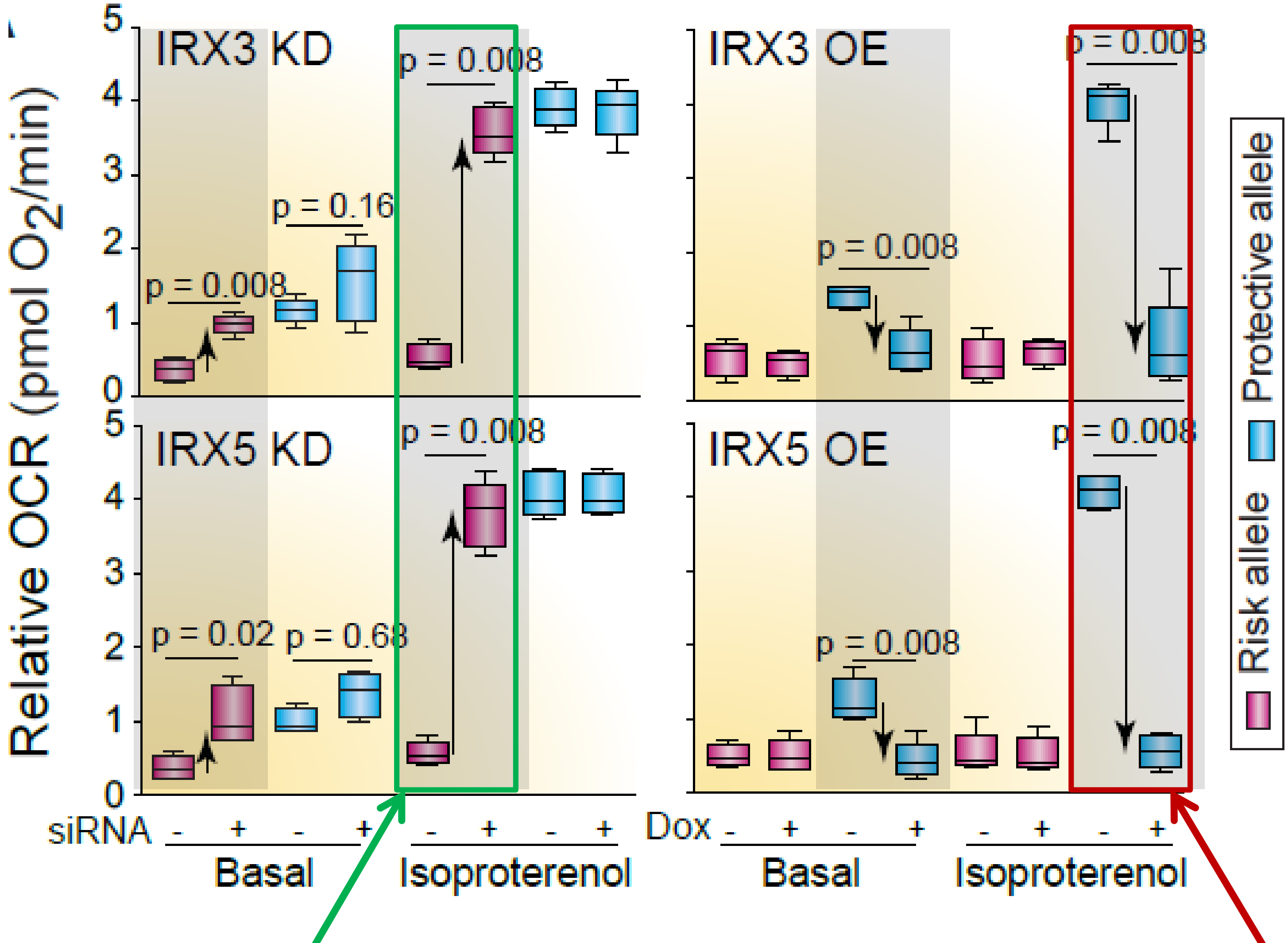
IRX3, IRX5 knock-down ★
(anti-obesity phenotypes)

IRX3, IRX5 overexpression
(pro-obesity phenotypes)

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

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

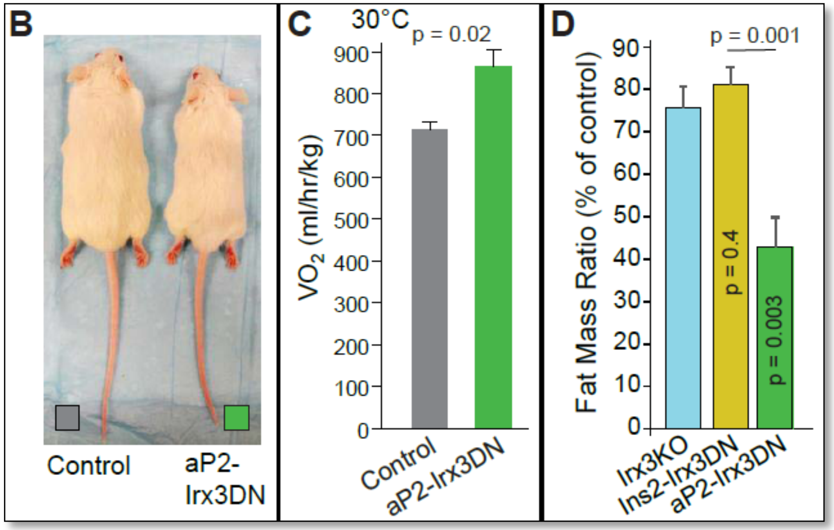
IRX3+IRX5 expression impacts energy utilization

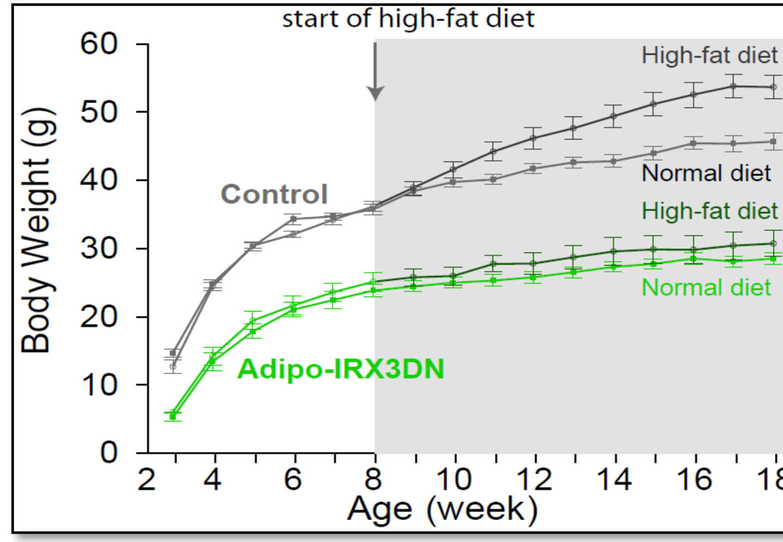


Risk individuals: IRX3/5 repression restores respiration,thermogenesis

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

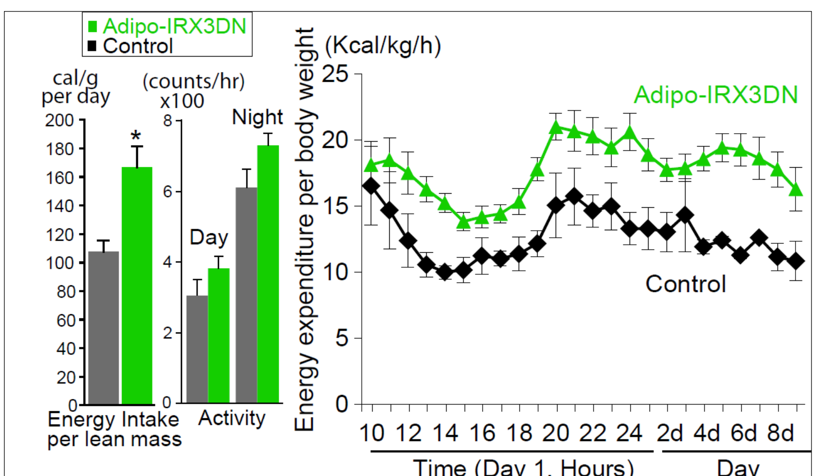
Irx3 adipose repression: anti-obesity phenotypes in mice





54% reduced body weight

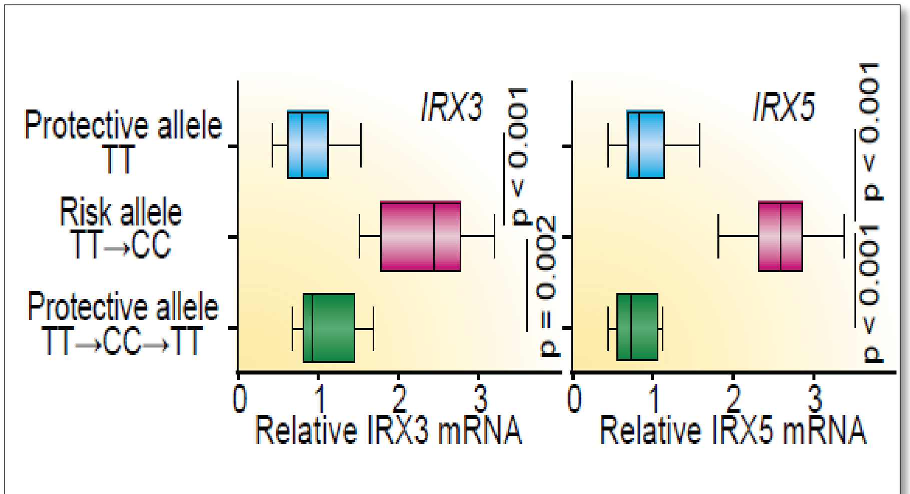
Resistance to high-fat diet



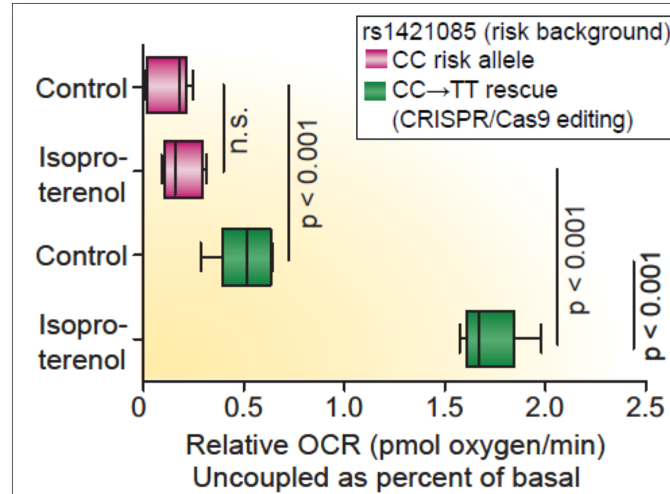
Increased energy dissipation

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

Single-nucleotide editing reverses thermogenesis in humans



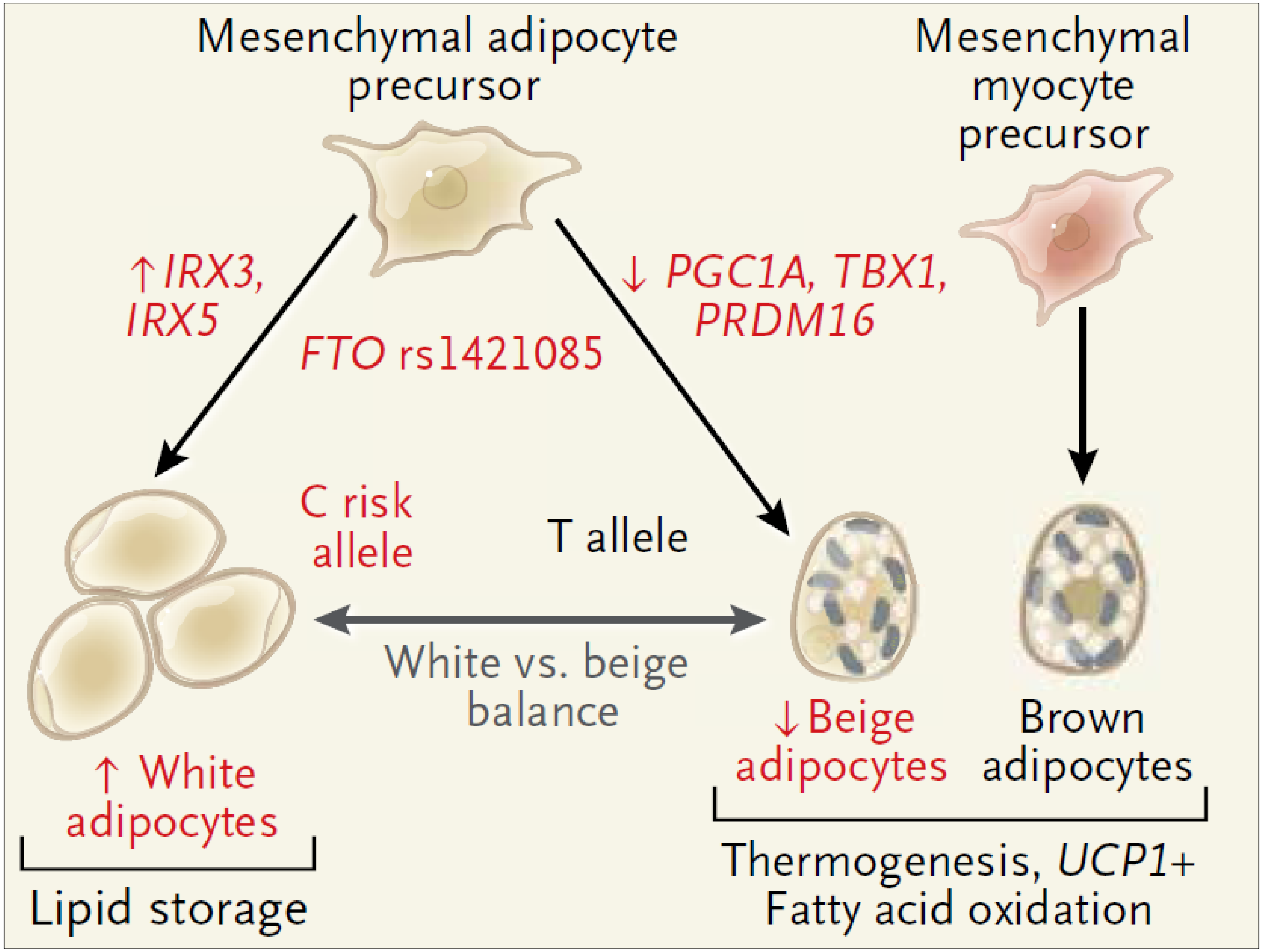
rs1421085 editing alters IRX3+IRX5 expression (500,000 and 1 million nucleotides away!)



rs1421085 editing restores thermogenesis

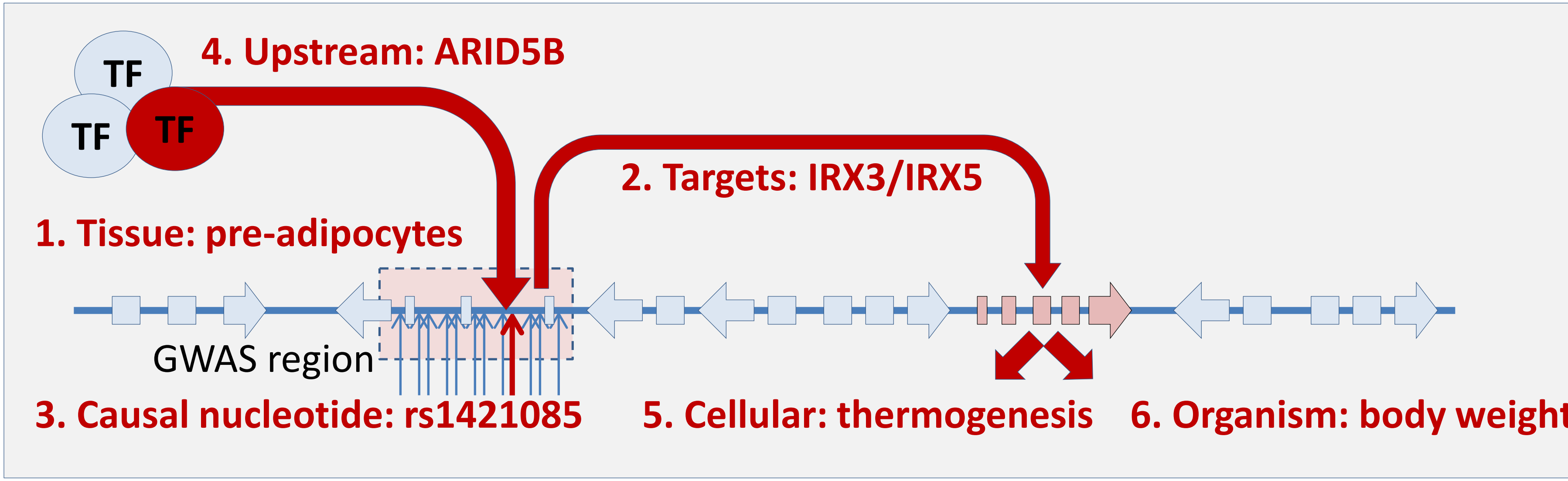
rs1421085 causality: C-to-T editing rescues IRX3/IRX5 expression, ARID5B repression, thermogenesis, developmental expression

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 (locus, really): Common/rare alleles, Linkage vs. GWAS
3. Evolution/scaling of GWAS power: Sharing, inflection points
4. Challenge of fine-mapping: Co-inheritance, LD, Haplotypes, Recomb.
5. From locus to mechanism - Case study: FTO and Obesity
6. Challenges in disease mechanism inference
7. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA

6. Challenges in disease mechanism

and resources for overcoming them

Biological questions using GWAS datasets

Analysis	Purpose	Discoveries
GWAS	detecting trait-SNP associations	~10,000 robust associations with diseases and disorders, quantitative traits, and genomic traits
Genome-wide CNV analysis	detecting trait-CNV associations	hundreds of associations with diseases and disorders
Genome-wide assessment of LD	quantifying genome architecture	large variation in LD in the genome
Estimation of SNP heritability[a]	genetic architecture	large proportion of genetic variation captured by common SNPs
Estimation of genetic correlation[a]	detecting and quantifying pleiotropy	pleiotropy is ubiquitous
Polygenic risk scores[a]	detecting pleiotropy; validating GWAS discoveries	out-of-sample prediction works as expected; detection of novel trait associations
Mendelian randomization[a]	testing causal relationships	replication of known causal relationships; empirical evidence of observational associations that are not causal
Population differences in allele frequencies	reconstructing human population history; detecting selection	genetic structure can mimic geographical structure; evidence of natural selection
Trait GWAS with -omics GWAS[a]	fine-mapping; detecting target genes; function	two-thirds of GWAS-associated loci implicate a gene that is not the nearest gene to the most associated SNP

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

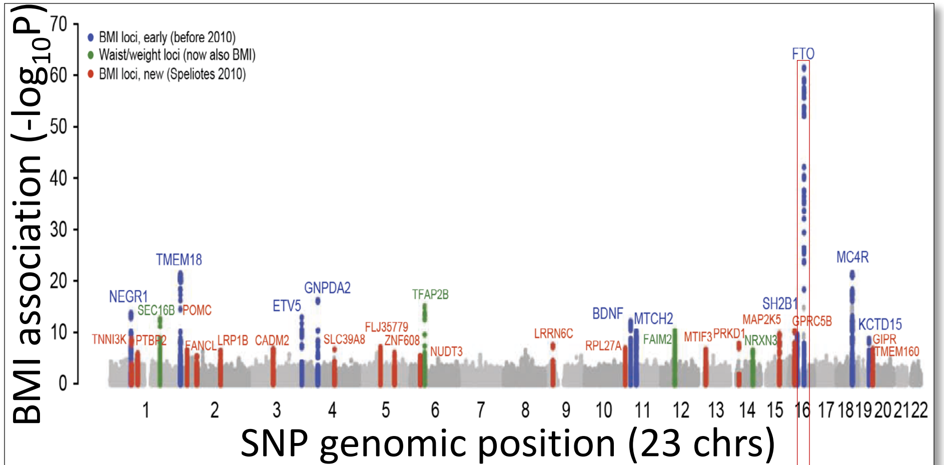
Disease architecture guides study design for discovery

Determinant	Outcome
Sample size	Direct correlation between the sample size and power to detect the causative alleles (within a limit)
Effect sizes of the causative alleles	Inverse correlation between effect size and power to detect an association
Minor allele frequency	GWAS by design detect only common alleles (MAF > 0.05)
Proximity of the phenotype to the genotype	More powerful for detecting an effect on proximal than distal phenotypes
Population characteristics	Presence of other competing factors dilute the power to detect an effect
Population admixture	Increase the risk of spurious results
Phenotype	Phenotypic admixture) and phenocopy conditions dilute the power to detect an effect size
Study design platform	Prospective studies are free of potential confounding differences between cases and controls
Structure of LD	Might not adequately capture information content of the true causative alleles
Density of the genotyping	Low-density arrays may not offer adequate cover for the common haplotypes

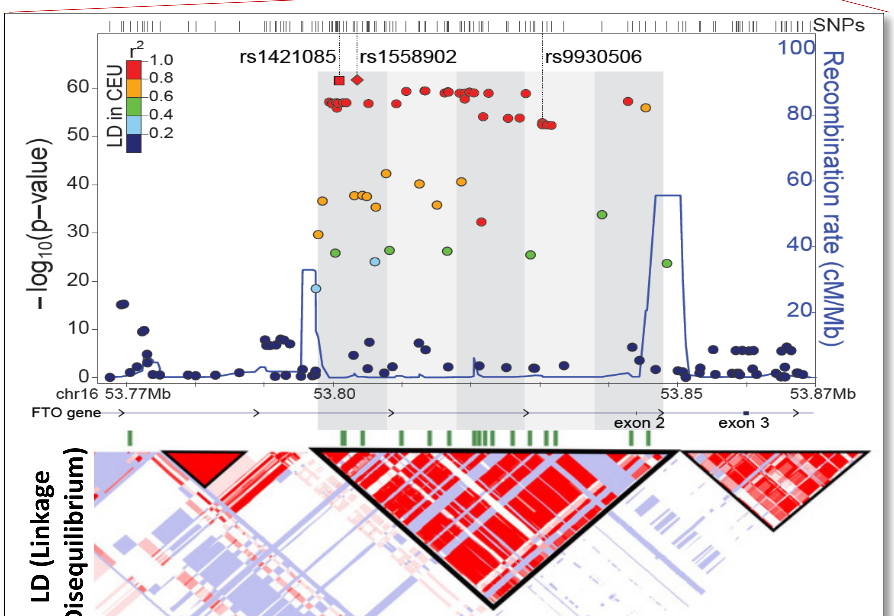
Molecular genetic studies of complex phenotypes. Marian - Translational Research, 2012

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

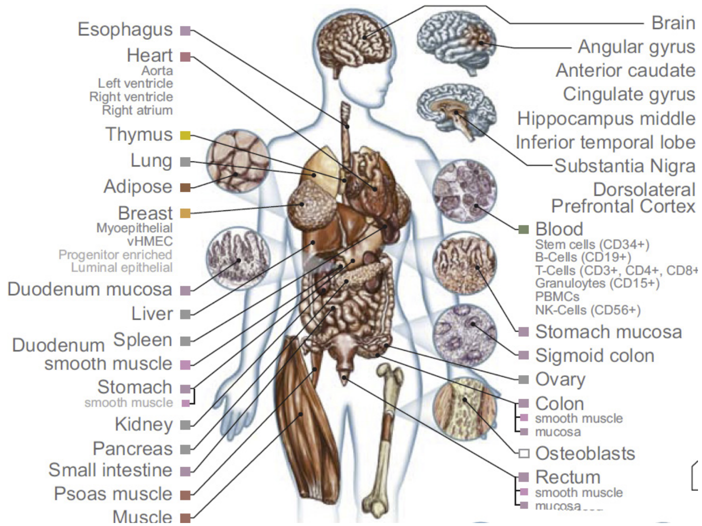
The promise of genetics

- 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

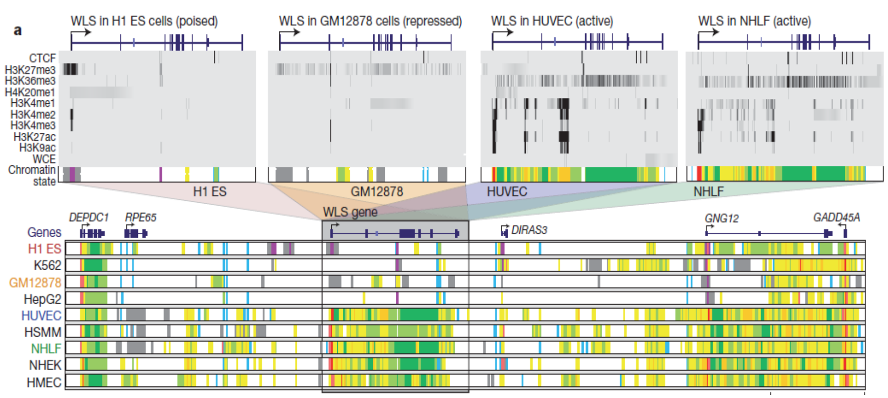
Genomic medicine: challenge and promises



Roadmap Epigenomics, Nature 2015

The remedy

- Annotate non-coding genome (ENCODE/Roadmap)
- Link enhancers to regulators and target genes
- Elucidate intermediate molecular and cellular phenotypes



Ernst, Nature 2011

The deliverables

- Relevant cell type
- Target genes
- Causal variant
- Upstream regulator
- Relevant pathways
- Intermediate phenotypes

Review paper on GWAS dissection (Nature Biotech 2012)

Interpreting noncoding genetic variation in complex traits and human disease

Lucas D Ward[1,2] & Manolis Kellis[1,2]

Association studies provide genome-wide information about the genetic basis of complex disease, but medical research has focused primarily on protein-coding variants, owing to the difficulty of interpreting noncoding mutations. This picture has changed with advances in the systematic annotation of functional noncoding elements. Evolutionary conservation, functional genomics, chromatin state, sequence motifs and molecular quantitative trait loci all provide complementary information about the function of noncoding sequences. These functional maps can help with prioritizing variants on risk haplotypes, filtering mutations encountered in the clinic and performing systems-level analyses to reveal processes underlying disease associations. Advances in predictive modeling can enable data-set integration to reveal pathways shared across loci and alleles, and richer regulatory models can guide the search for epistatic interactions. Lastly, new massively parallel reporter experiments can systematically validate regulatory predictions. Ultimately, advances in regulatory and systems genomics can help unleash the value of whole-genome sequencing for personalized genomic risk assessment, diagnosis and treatment.

Table 1 The diversity of genetic architectures underlying human phenotypes

Types of variation

Architecture	Notes and examples	Role of computational and regulatory genomics
Classic monogenic traits	The earliest human genes characterized were those leading to inborn errors in metabolism, which were shown by Garrod in the early 1900s to follow Mendelian inheritance[130,131]. The modern study of human disease genes began with the cloning of loci responsible for high-penetrance monogenic disorders with Mendelian inheritance patterns, such as phenylketonuria and cystic fibrosis[130,132,133], that were most amenable to classical mapping approaches. Variants associated with monogenic traits were also the first to be identified through positional cloning in the 1980s, a classic success being the *CFTR* mutations responsible for most cases of cystic fibrosis[3,132,133].	As the underlying mutations tend to alter protein structure, the computational challenge in predicting their effect lies in molecular modeling and structural studies.
Monogenic traits with multiple disease alleles	Even monogenetic diseases differ greatly in the extent to which a single risk allele predominates among affected individuals (allelic heterogeneity). On one end of the spectrum, the F508del allele of *CFTR* is found in about 70% of patients with cystic fibrosis[134], even though thousands of alleles are known. In contrast, phenylketonuria is extremely heterogeneous, with different *PAH* alleles predominating among affected individuals in different populations[135]. A majority of mutations in this class are missense or nonsense coding mutations[3].	As noted above, for protein-coding mutations, the relevant problem is predicting the biochemical effect of the amino acid substitution. In cases of allele heterogeneity, the observed substitutions may be too numerous to characterize experimentally, necessitating computational models (**Fig. 3c**).
Multiple loci with independent contributions ('oligogenic')	Many variants increase or decrease the risk of a disease, with the final phenotype relying on the genotype at many loci (locus heterogeneity). One example well studied through linkage analysis is Hirschprung disease, a complex disorder with low sex-dependent penetrance for which at least ten underlying genes are involved, including the tyrosine kinase receptor *RET* and the gene *GDNF* which encodes its ligand[136]. Interestingly, the most common variant in the main susceptibility gene *RET* is noncoding, a SNP in an enhancer. Both coding and noncoding variants are involved typically in one or a few well-defined pathways.	Oligogenic traits, in which a handful of well-characterized loci contribute to the phenotype, may present the best opportunity to observe and quantify epistatic interactions. In cases where noncoding regions are implicated, these haplotypes can be functionally mapped to isolate the most likely causal variants (**Fig. 2**).
Large numbers of variants jointly contributing weakly to a complex trait	GWAS of complex traits are also discovering many weakly contributing loci. For example, a recent meta-analysis of several height studies found 180 loci reaching genome-wide significance[15,103,137], enriched near genes already known to underlie skeletal growth defects. In the height study and in a study of psychiatric disorders, it has been shown that polygenic association extends to thousands of common variants, extending far beyond genome-wide-significant loci[137,138].	In contrast to the variants underlying monogenic traits, the variants involved in complex traits are overwhelmingly not associated with missense or nonsense coding mutations, suggesting that their mechanisms are primarily regulatory[11]. Large sets of regulatory variants can be combined with reference annotations to elucidate relevant pathways and tissues (**Fig. 3b**, **Table 5**).
Variants regulating a 'molecular trait' with unknown effect on organismal phenotype or fitness	Variants are rapidly being discovered that directly affect molecular quantitative traits, such as gene expression or chromatin state, many of which may have no effect on organismal phenotype or fitness[38].	QTLs and allele-specific analyses are needed to characterize these variants (**Fig. 1b**,**c**). As the studies performed to date sample only a small fraction of the cell types in which a variant may have an effect, and variant-expression associations are highly tissue specific[139], it is possible that many such regulatory variants remain to be discovered.
Variants causing no known molecular phenotype and no effect on organismal phenotype or fitness	The idea that the majority of mutations are neutral from an adaptive perspective was controversial when first proposed, and now is widely accepted[140–142].	Although it is straightforward to calculate from the genetic code what fraction of protein-coding mutations will cause an amino acid change, an analogous estimate for other molecular phenotypes is far more challenging and requires comprehensive regulatory models at the nucleotide level.
Private and somatic variants	Somatic mutations within an organism are frequent driver mutations selected in cancer formation[143].	The interpretation of private and somatic variations (**Fig. 3d**) will also benefit tremendously from a systematic regulatory annotation, as they are likely to exploit existing regulatory pathways, even though they are subject to cellular, rather than organismal, selective pressures.

Types of common variant association: GWAS, eQTLs, ASEs, EWAS

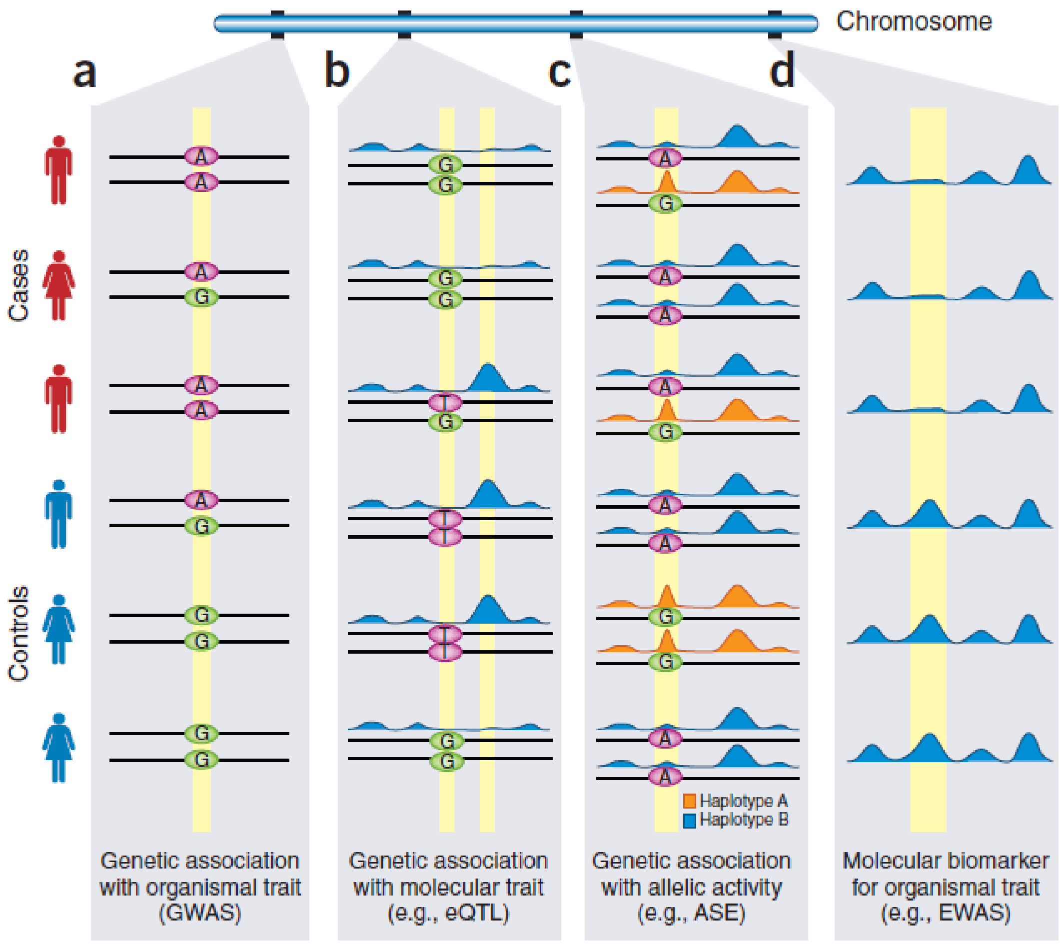


Table 3 Mechanisms through which noncoding variants influence human disease

Noncoding element disrupted	Molecular function and effect of mutations	Disease association
Splice junction and splicing enhancer	Splicing of mRNA is constitutive for some transcripts and highly tissue specific for others, relying on both canonical sequences at the exon-intron junction and weakly specified sequence motifs distributed throughout the transcript. Mutations affecting constitutive splice sites can have an effect similar to nonsense or missense mutations, resulting in aberrantly included introns or skipped exons, sometimes resulting in nonsense-mediated decay (NMD).	Splicing regulatory variants are implicated in several diseases[156,157]. A recent analysis suggests that the majority of disease-causing point mutations in OMIM may exert their effects by altering splicing[158]. Alternative splice site variants in the *WT1* gene are involved in Frasier syndrome (FS)[159]. Skipping of exon 7 of the *SMN* gene is involved in spinal muscular atrophy (SMA)[160].
Sequences regulating translation, stability, and localization	Sequences in the 5′ untranslated regions (UTRs) of mRNAs can influence translation regulation, such as upstream open reading frames (ORFs), premature AUG or AUC codons, and palindromic sequences that form inhibitory stem loops[161]. Sequence motifs in the 3′ UTR are recognized by microRNAs and RNA-binding proteins (RBPs).	Loss-of-function mutations in the 5′ UTR of *CDKN2A* predispose individuals to melanoma[162]. A rare mutation that creates a binding site for the miRNA hs-miR-189 in the transcript of the gene *SLITRK1* is associated with Tourette's syndrome[163].
Genes encoding *trans*-regulatory RNA	Noncoding RNAs participate in a panoply of regulatory functions; these RNAs range from the well-understood transfer and ribosomal RNA to the recently discovered long noncoding RNAs[164,165].	Both rare and common mutations in the gene *RMRP*, encoding an RNA component of the mitochondrial RNA processing RNase, have been associated with cartilage-hair hypoplasia[166]. Noncoding RNA mutations can cause many other diseases[167].
Promoter	Promoter regions are an essential component of transcription initiation and the assembly of RNA polymerase and associated regulators. Mutations can affect binding of activators or repressors, chromatin state, nucleosome positioning, and also looping contacts of promoters with distal regulatory elements. Genes with coding disease mutations can also harbor independently associated regulatory variants that correlate with expression, are bound by proteins in an allele-specific manner, and disrupt or create regulatory motifs[168].	Mutations in the promoter of the HIV-1 progression–associated gene *CCR5* are correlated with expression of the receptor it encodes and bind differentially to at least three transcription factors[169,170]. *APOE* promoter mutations are associated with Alzheimer's disease[171,172]. Heme oxygenase-1 (*HO1*) promoter mutations lead to expression changes and are associated with many diseases[173].
Enhancer	Enhancers are distal regulatory elements that often lie 10,000–100,000 nucleotides from the start of their target gene. Mutations within them can disrupt sequence motifs for sequence-specific transcription factors, chromatin regulators and nucleosome positioning signals. Structural variants including inversions and translocations can disrupt their regulatory activity by moving them away from their targets, disrupting local chromatin conformation, or creating interactions with insulators or repressors that can hinder their action. Although it is thought that looping interactions with promoter regions play a role, the rules of enhancer-gene targeting are still poorly understood.	The role of distal enhancers in disease was suggested even before the development of GWAS by the many Mendelian disorders for which some patients had translocations or other structural variants far from the promoter[174–176]. In one early study, point mutations were mapped in an unlinked locus in the intron of a neighboring gene, a million nucleotides away from the developmental gene *Shh*[177]; this distal locus acted as an enhancer of *Shh* and recapitulated the polydactyly phenotype in mouse. A number of GWAS hits have been validated as functional enhancers[178]: for example, common variants associated with cancer susceptibility map to a gene desert on chromosome 8, with one SNP demonstrated to disrupt a TCF7L2 binding site and to inhibit long-range activation of the oncogene *MYC*[179–181].
Synonymous mutations within protein-coding sequences	All of the aforementioned regulatory elements can also be encoded within the protein-coding exons themselves. Thus, synonymous mutations within protein-coding regions may be associated with noncoding functions, acting pre-transcriptionally at the DNA level or post-transcriptionally at the RNA level.	A synonymous variant in the dopamine receptor gene *DRD2* associated with schizophrenia and alcoholism has been shown to modulate receptor production through differences in mRNA folding and stability[182].

Types of evidence for GWAS interpretation

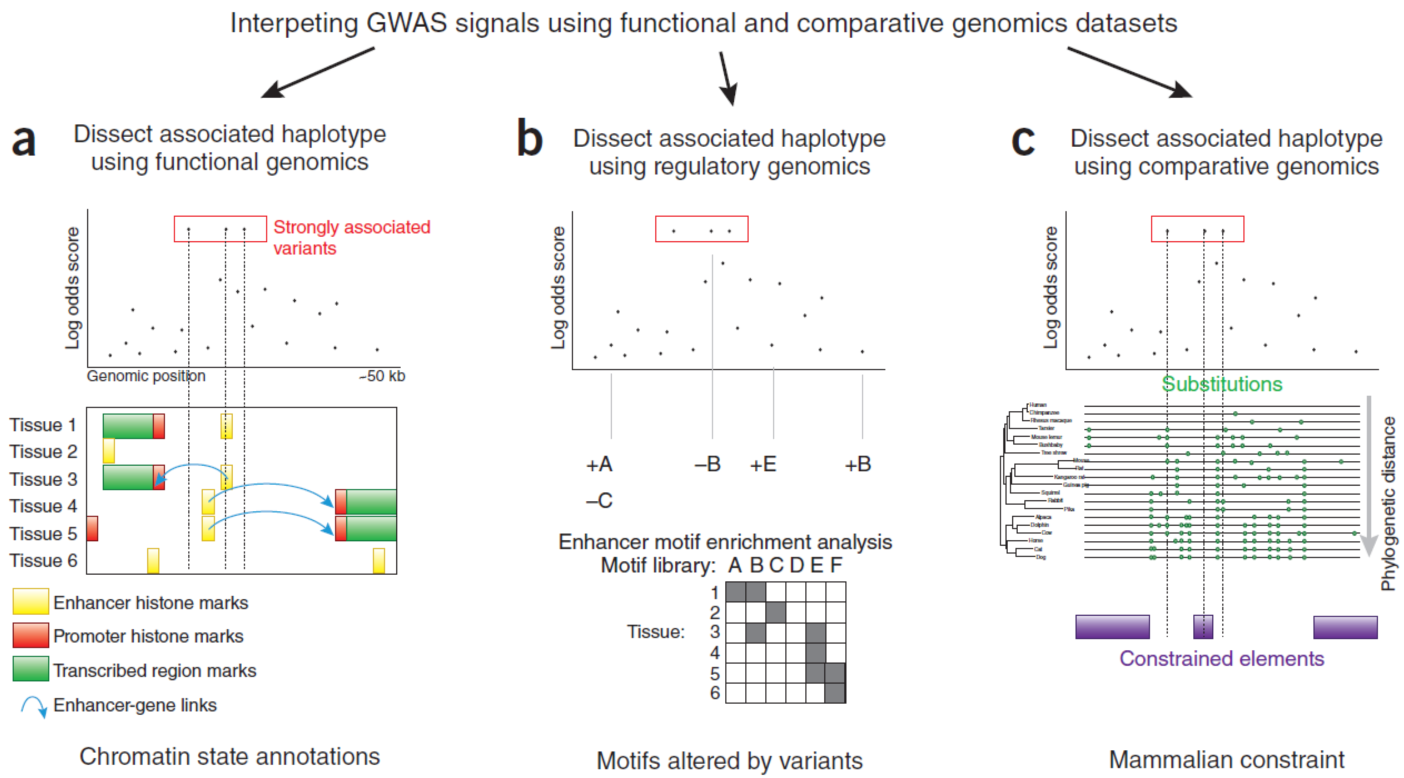


Figure 2 Dissecting haplotypes discovered through association tests. These three examples are ways to annotate loci containing several linked SNPs (in this case, three) to discover those most likely to be causal. (**a**) Functional genomics techniques are being developed to discover putative regulatory elements and link these elements to their target genes. Here, the middle SNP lies in an enhancer in tissue 1 and tissue 3, and regulates a gene to its left. (**b**) Regulatory genomics information leads to prediction of sequence motifs active in classes of enhancers, and this can be combined with the motif creation and/or disruption caused by variants. In this case, the middle SNP deletes a match to motif B, which is predicted to be active in enhancers found in both tissues 1 and 3. (**c**) Comparative genomics identifies regions of evolutionary constraint in noncoding sequence. Here, sequence surrounding only the middle SNP is constrained across mammals.

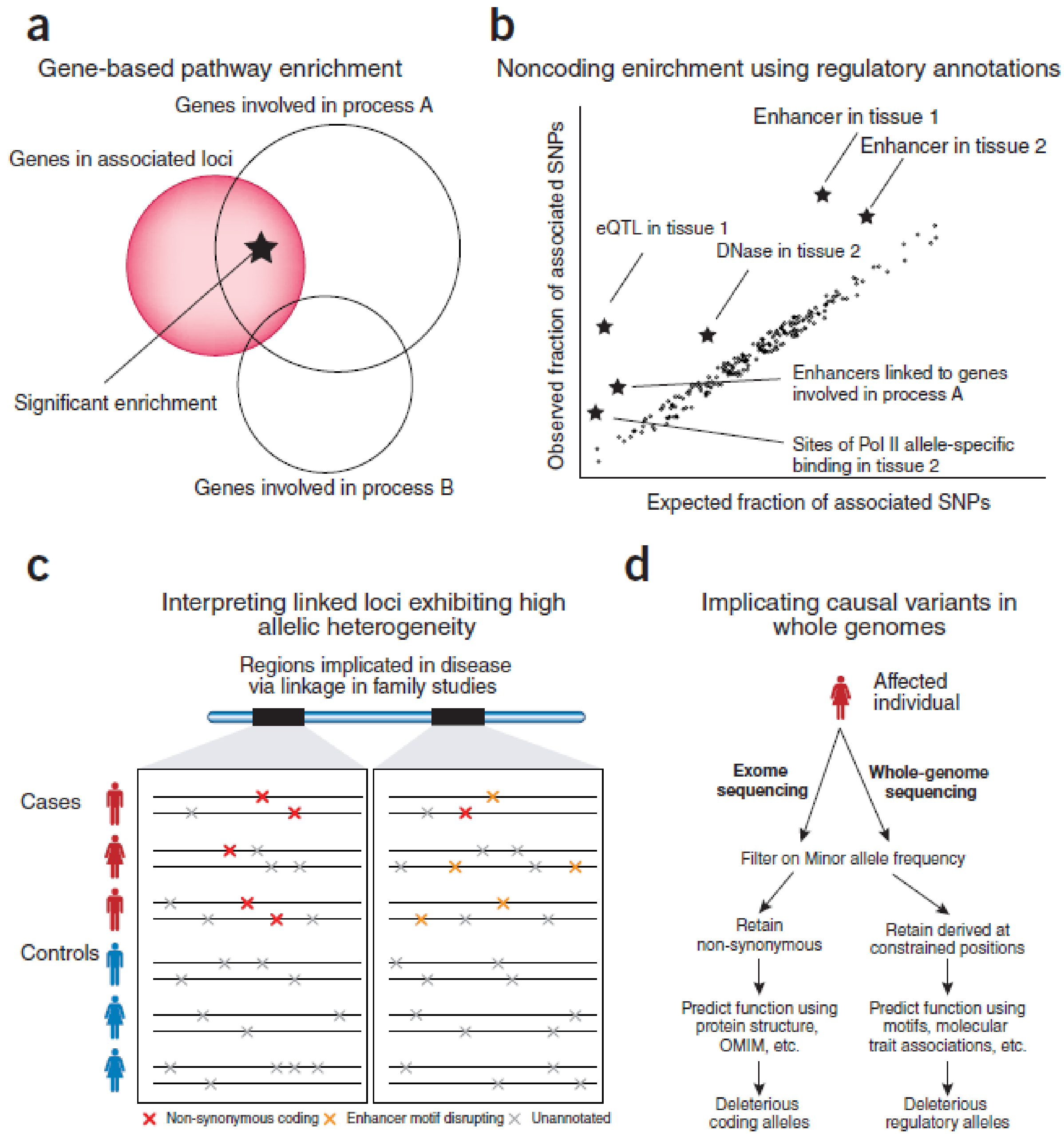
Tools for data integration

Table 4 Comparison of recent tools to systematically annotate variants[a]

Tool	Type	Input method	Protein annotation	Regulatory annotation	Other
SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation/)	Server	Variants	Deleteriousness scores	Conservation scores	dbSNP clinical association data
ANNOVAR[57]	Software	Variants, regions	User defined: user downloads desired variation, conservation, coding and noncoding functional annotations		
ENSEMBL VEP[56]	Server	Variants, regions	Deleteriousness scores	Regulatory motif alteration scores	OMIM, GWAS data
VAAST[58]	Software	Variants	Deleteriousness scores	Conservation scores	Aggregation to discover rare variants in case-control studies
HaploReg[54]	Server	Variants, studies	dbSNP consequence data	Chromatin state, protein binding, DNase, conservation, regulatory motif alteration scores	GWAS data, eQTL, LD calculation, enrichment analysis per study
RegulomeDB[55]	Server	Variants, regions	Not applicable	Histone modification, protein binding, DNase, conservation, regulatory motif alteration scores	eQTL, reporter assays, combined score analysis per variant

[a]Many such tools have been released as databases or software in the past decade; a sampling of the most recent are listed here.

Systems-level analyses



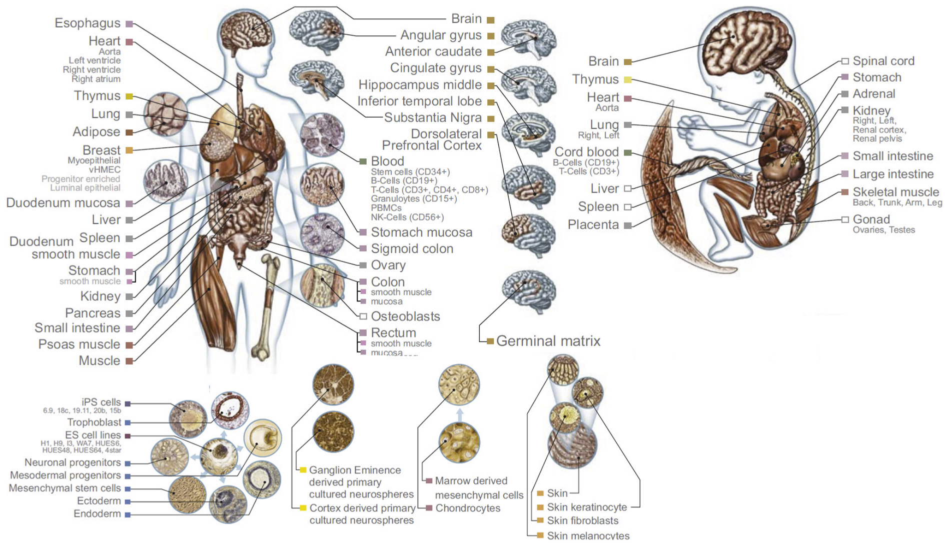
Enrichment analyses

Table 5 Examples of regulatory enrichment analyses of genetic associations

Class of test	Finding	Computational tools used
Gene-set enrichment near associated loci	Regulatory network of five proteins is implicated in Kawasaki disease[183]	Ingenuity Pathway Analysis (closed source)
	Genes differentially expressed in adipose overlap with genetic associations with obesity[184]	Microarray analysis of differential expression
	TGF-β pathway and Hedgehog signaling pathway are enriched among height GWAS loci[103]	GSEA using MAGENTA[185], network from text-mining using GRAIL[186], known disease genes from OMIM[4] and eQTL enrichment
Concordance with eQTL results	eQTL prioritization during replication facilitates validation of two Crohn's disease susceptibility loci[187]	eQTL enrichment
	GWAS involving immune system show enrichment for lymphoblastoid eQTL[64]	eQTL enrichment (RTC[64])
Chromatin state enrichment	Many GWAS show enrichment for enhancers in biologically relevant cell types[62]	ChromHMM to define discrete chromatin states (M.K. and colleagues[188]); and enrichment analysis
TF binding site and DNase hypersensitivity enrichment	Many GWAS show enrichment for ENCODE-annotated DNase and ChIP sites[189]	Enrichment analysis
	Many GWAS show enrichment for DNase in biologically relevant cell types[63]	Hot-spot algorithm to define discrete hypersensitive sites[190]; enrichment analysis
	FOXA1 and estrogen receptor binding sites are enriched among breast cancer GWAS loci[191]	Variant Set Enrichment (VSE[191])

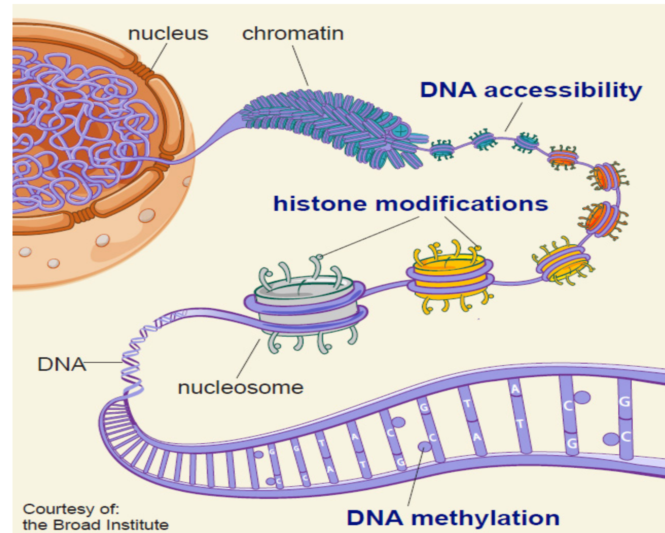
Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



x

Diverse epigenomic assays



Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood…)

Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood…)

ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

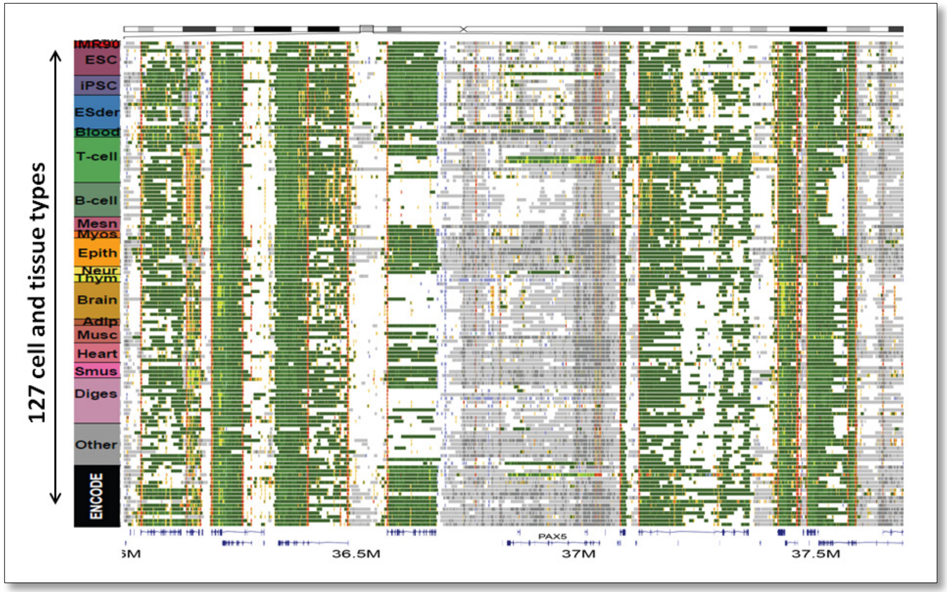
Histone modifications
- H3K4me3, H3K4me1, H3K36me3
- H3K27me3, H3K9me3, H3K27/9ac
- +20 more

Open chromatin:
- DNA accessibility

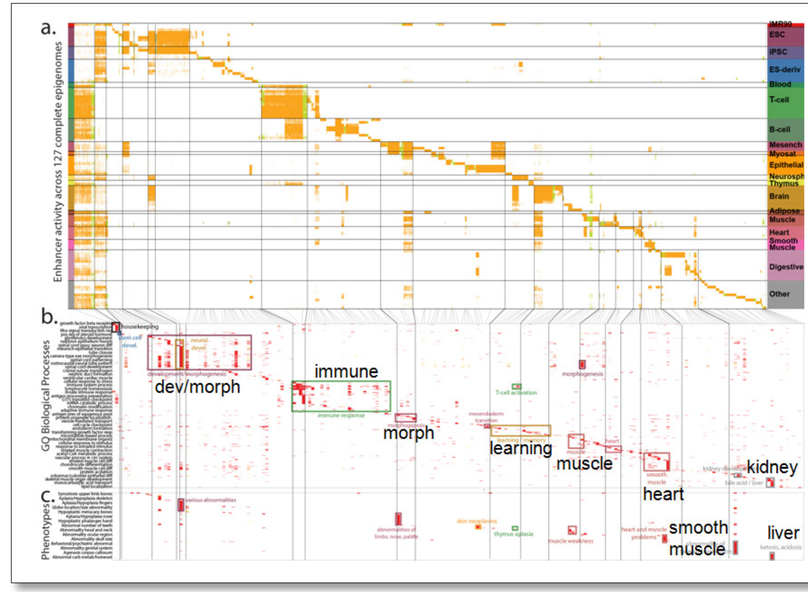
DNA methylation:
- WGBS, RRBS, MRE/MeDIP

Gene expression
- RNA-seq, Exon Arrays

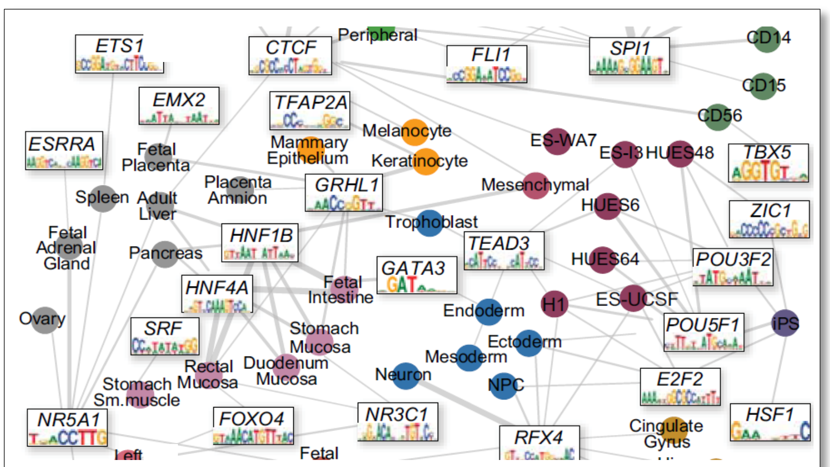
Enhancer modules, regulators, and target genes



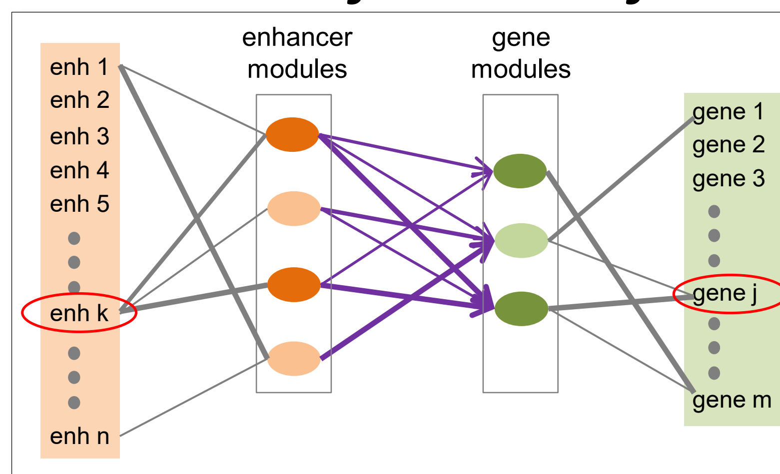
1. Map chromatin states across 127 tissue/cells



2. Group enhancers into modules of common function

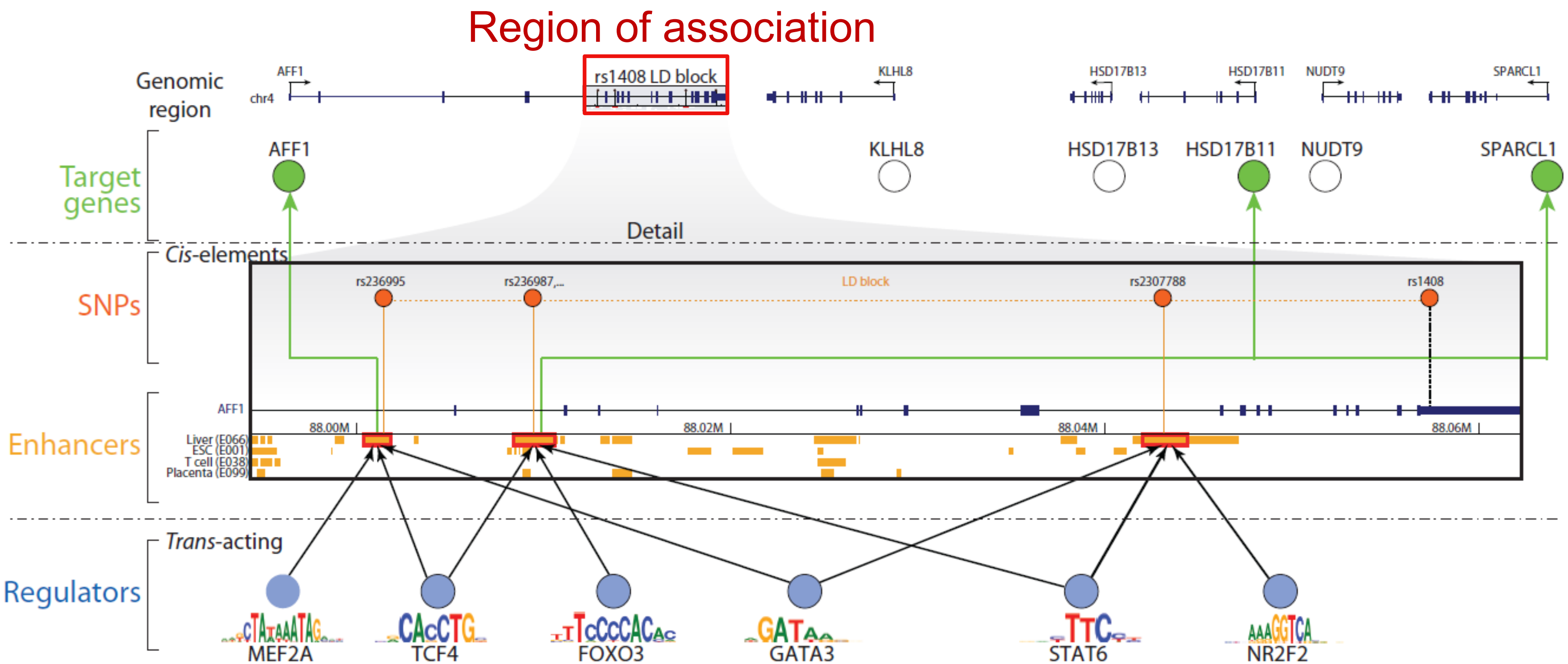


3. Predict module regulators using motif enrichment



4. Predict target genes using common activity

Use resulting annotations and networks for GWAS



- 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

HaploReg: systematic mining of GWAS variants

Query SNP: rs4684847 and variants with r^2 >= 0.8

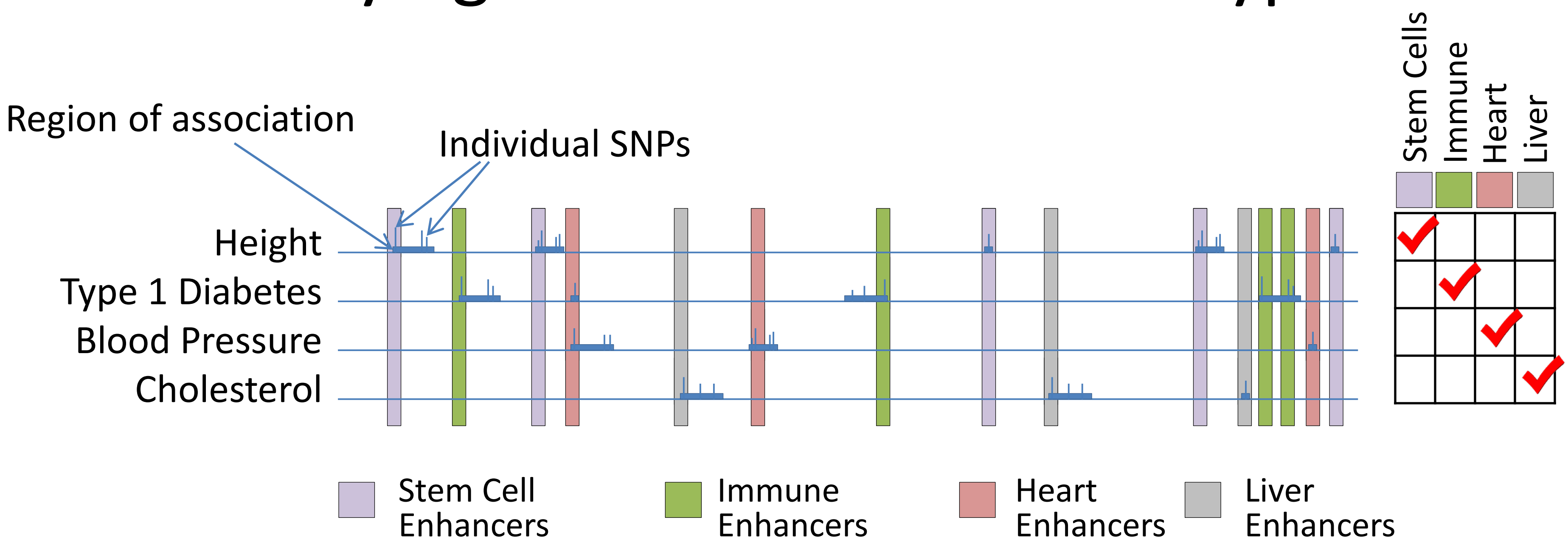
pos (hg19)	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	Drivers disrupted	GENCODE genes	dbSNP func annot
chr3:12329783	chr3:12288284	0.95	0.97	rs17036160	C	T	0.01	0.06	0.04	0.12		24 organs	7 organs	4 organs			4 altered motifs		PPARG	intronic
chr3:12336507	chr3:12295008	0.95	0.97	rs11709077	G	A	0.01	0.07	0.04	0.12		LNG	9 organs	15 organs			4 altered motifs		PPARG	intronic
chr3:12344730	chr3:12303231	0.94	0.97	rs11712037	C	G	0.01	0.08	0.04	0.12			6 organs	BLD			AP-1,TCF11::MafG		PPARG	intronic
chr3:12351521	chr3:12310022	0.95	0.97	rs35000407	T	G	0.01	0.07	0.04	0.12		LNG	5 organs				Smad		PPARG	intronic
chr3:12360884	chr3:12319385	0.95	0.97	rs150732434	TG	T	0.01	0.07	0.04	0.12		FAT	7 organs	MUS,VAS	CFOS		Hdx,Sox,TATA		PPARG	intronic
chr3:12365308	chr3:12323809	0.95	0.97	rs13083375	G	T	0.01	0.07	0.04	0.12		BLD	BLD, FAT				Homez,Sox,YY1		PPARG	intronic
chr3:12369401	chr3:12327902	0.95	0.97	rs13064760	C	T	0.01	0.07	0.04	0.12			7 organs				9 altered motifs		PPARG	intronic
chr3:12375956	chr3:12334457	0.95	0.97	rs2012444	C	T	0.01	0.07	0.04	0.12			SKIN, FAT, BLD				7 altered motifs		PPARG	intronic
chr3:12383265	chr3:12341766	0.96	0.99	rs13085211	G	A	0.18	0.10	0.04	0.12			FAT, SKIN				NRSF		PPARG	intronic
chr3:12383714	chr3:12342215	0.96	0.99	rs7638903	G	A	0.18	0.10	0.04	0.12			6 organs	CRVX					PPARG	intronic
chr3:12385828	chr3:12344329	0.95	1	rs11128603	A	G	0.18	0.10	0.04	0.12			CRVX				RXRA		PPARG	intronic
chr3:12386337	chr3:12344838	1	1	rs4684847	C	T	0.01	0.07	0.04	0.12			6 organs						PPARG	intronic
chr3:12388409	chr3:12346910	0.99	1	rs7610055	G	A	0.17	0.09	0.04	0.12			BLD				4 altered motifs		PPARG	intronic
chr3:12389313	chr3:12347814	0.99	1	rs17036326	A	G	0.17	0.09	0.04	0.12			FAT, BL						PPARG	intronic
chr3:12390484	chr3:12348985	0.99	1	rs17036328	T	C	0.17	0.09	0.04	0.12			FAT, CR						PPARG	intronic
chr3:12391207	chr3:12349708	0.99	1	rs6802898	C	T	0.61	0.15	0.04	0.12			FAT, BL						PPARG	intronic
chr3:12391583	chr3:12350084	0.99	1	rs2197423	G	A	0.17	0.09	0.04	0.12		FAT, LIV	8 organ						PPARG	intronic
chr3:12391813	chr3:12350314	0.99	1	rs7647481	G	A	0.17	0.09	0.04	0.12		4 organs	9 organ						PPARG	intronic
chr3:12392272	chr3:12350773	0.99	1	rs7649970	C	T	0.17	0.09	0.04	0.12		5 organs	9 organ						PPARG	intronic
chr3:12393125	chr3:12351626	1	1	rs1801282	C	G	0.01	0.07	0.04	0.12		FAT, LIV	9 organ						PPARG	missense
chr3:12393682	chr3:12352183	0.99	1	rs17036342	A	G	0.17	0.09	0.04	0.12		FAT	9 organ						PPARG	intronic
chr3:12394840	chr3:12353341	0.99	1	rs1899951	C	T	0.61	0.15	0.04	0.12		FAT	9 organs				Mef2		PPARG	intronic
chr3:12395645	chr3:12354146	0.99	1	rs4684848	G	A	0.61	0.15	0.04	0.12		FAT, BLD	9 organs	ADRL,GI,CRVX	5 bound proteins				PPARG	intronic
chr3:12396845	chr3:12355346	0.93	1	rs4135250	A	G	0.17	0.09	0.04	0.13			4 organs	PLCNT					PPARG	intronic
chr3:12396913	chr3:12355414	0.98	1	rs71304101	G	A	0.01	0.07	0.04	0.12			4 organs	PLCNT			Crx,NF-E2		PPARG	intronic
chr3:12396955	chr3:12355456	0.96	1	rs2881654	G	A	0.61	0.15	0.04	0.12			4 organs				7 altered motifs		PPARG	intronic

Adipose_Derived_Mesenchymal_Stem_Cell_Cultured_Cells, CD4+_CD25-_IL17+_PMA-Ionomcyin_stimulated_Th17_Primary_Cells, Muscle_Satellite_Cultured_Cells, Penis_Foreskin_Fibroblast_Primary_Cells_skin01, Penis_Foreskin_Fibroblast_Primary_Cells_skin02, Penis_Foreskin_Keratinocyte_Primary_Cells_skin02, Penis_Foreskin_Keratinocyte_Primary_Cells_skin03, A549_EtOH_0.02pct_Lung_Carcinoma, HeLa-S3_Cervical_Carcinoma, NHEK-Epidermal_Keratinocytes

- **Start with any list of SNPs or select a GWA study**
  - Mine ENCODE and Roadmap epigenomics data for hits
  - Hundreds of assays, dozens of cells, conservation, motifs
  - Report significant overlaps and link to info/browser
- **Try it out: http://compbio.mit.edu/HaploReg**

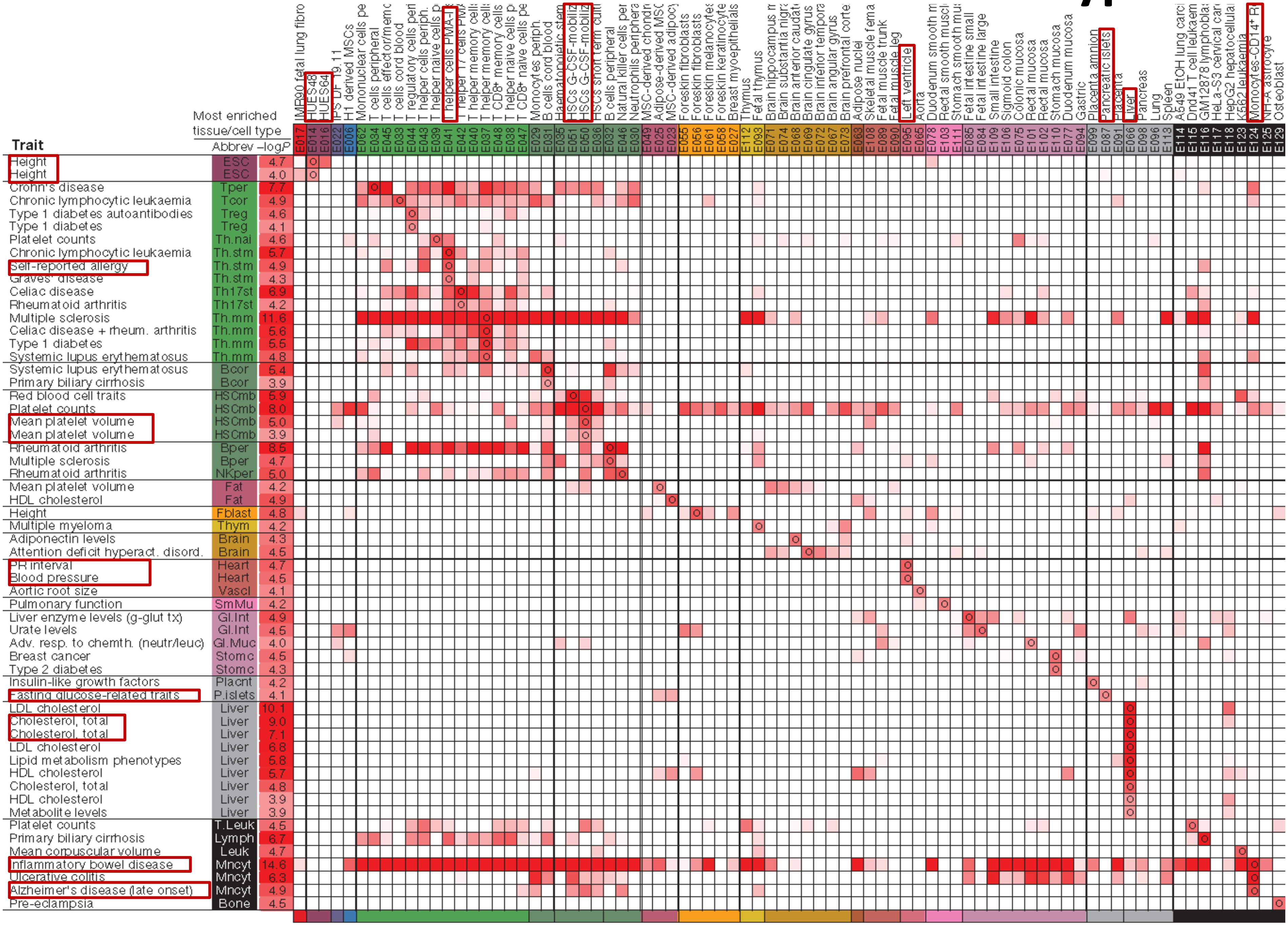
Ward, Kellis NAR 2011

Identifying disease-relevant cell types

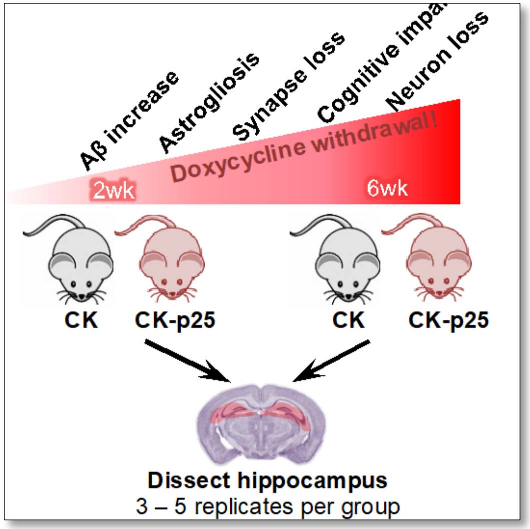


- For every trait in the GWAS catalog:
  - Identify all associated regions at P-value threshold
  - Consider all SNPs in credible interval ($R^2 \geq .8$)
  - Evaluate overlap with tissue-specific enhancers
  - Keep tissues showing significant enrichment ($P<0.001$)
- Repeat for all traits (rows) and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

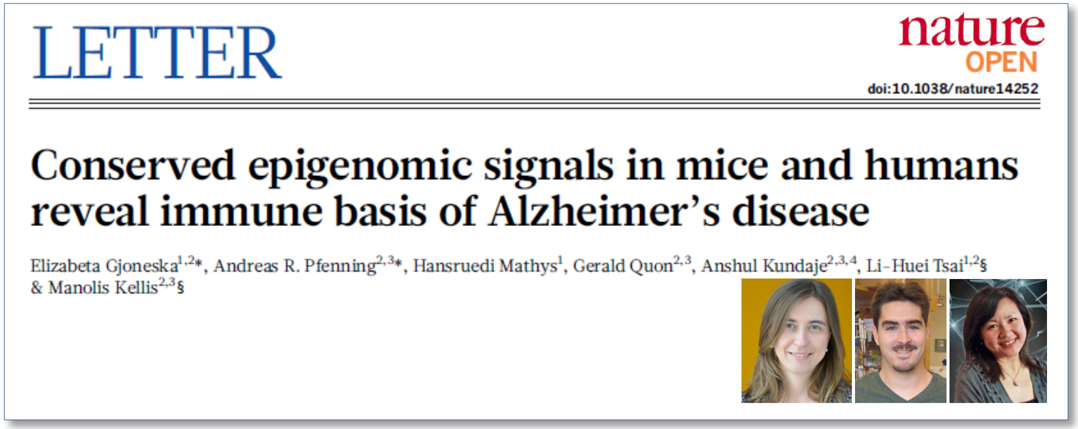
GWAS hits in enhancers of relevant cell types

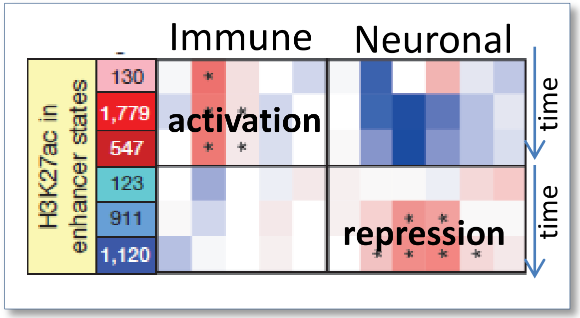


Immune activation + neural repression in human + mouse

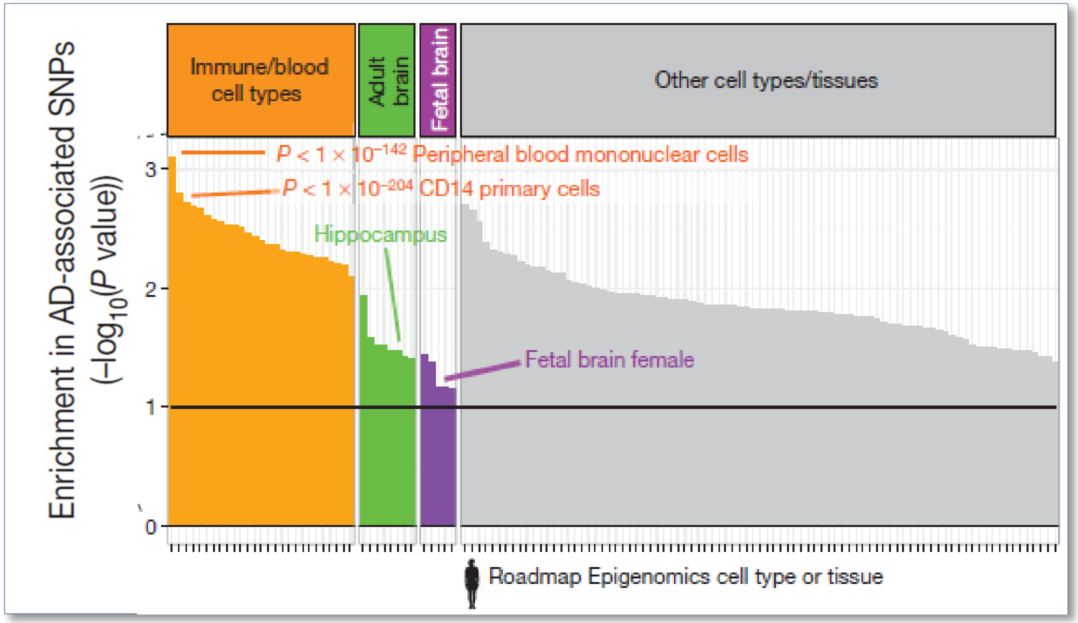


Epigenomics of AD progression





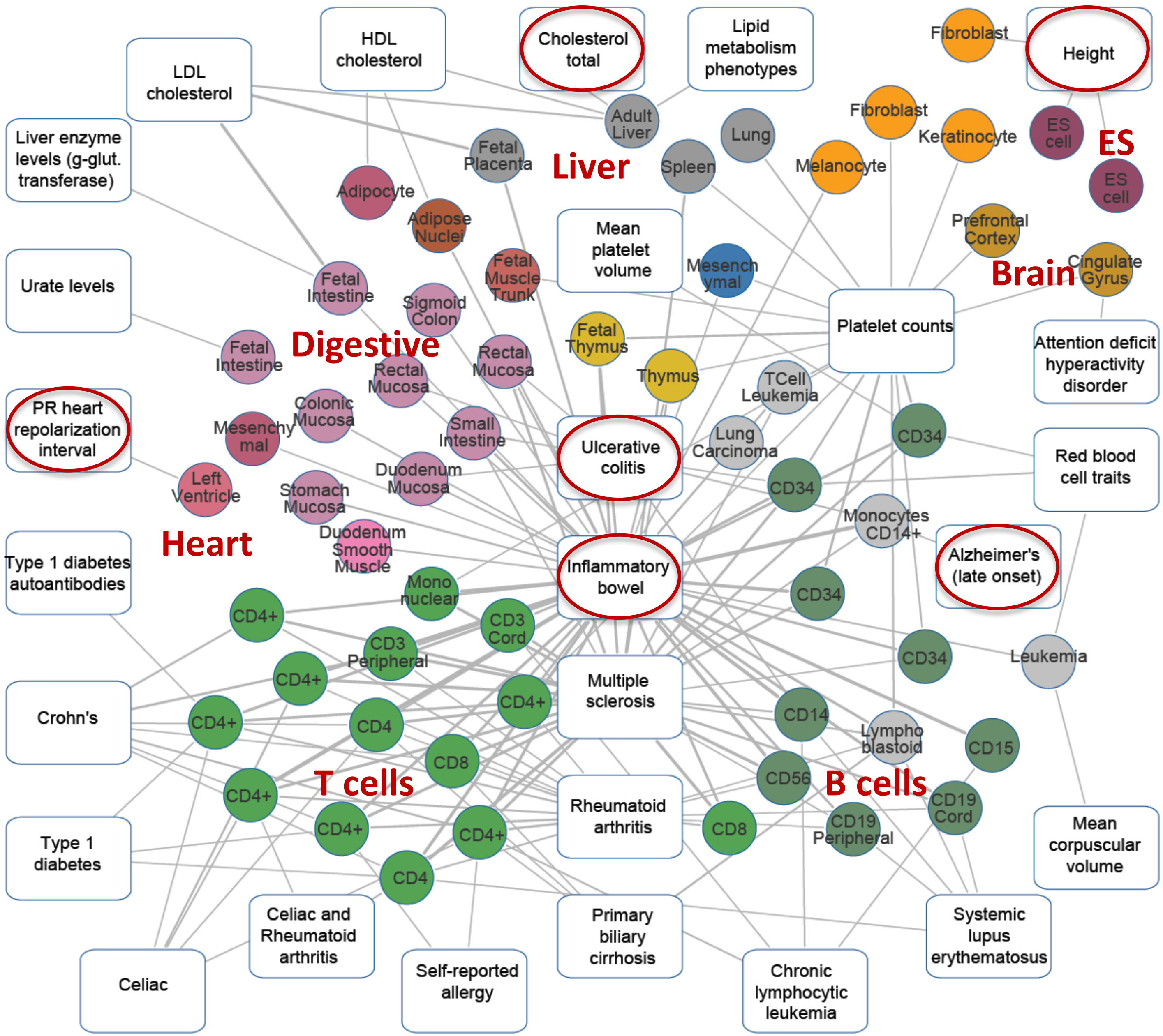
Immune activation precedes neuronal repression



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

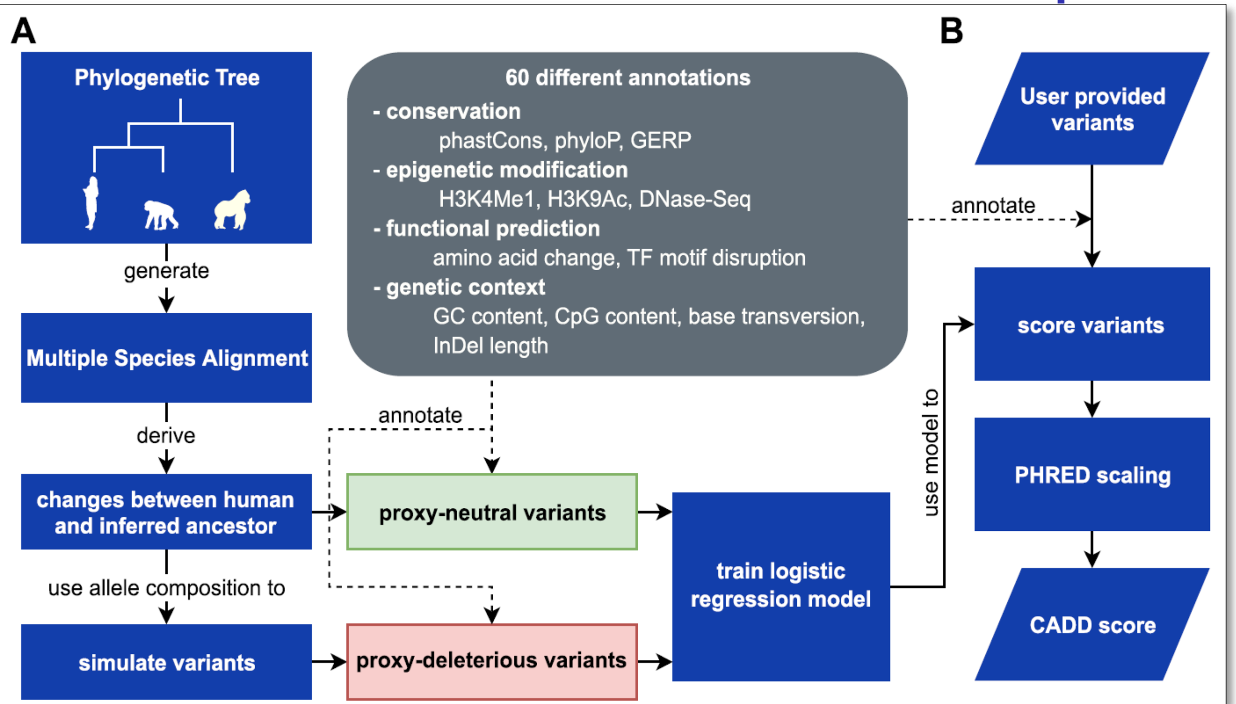


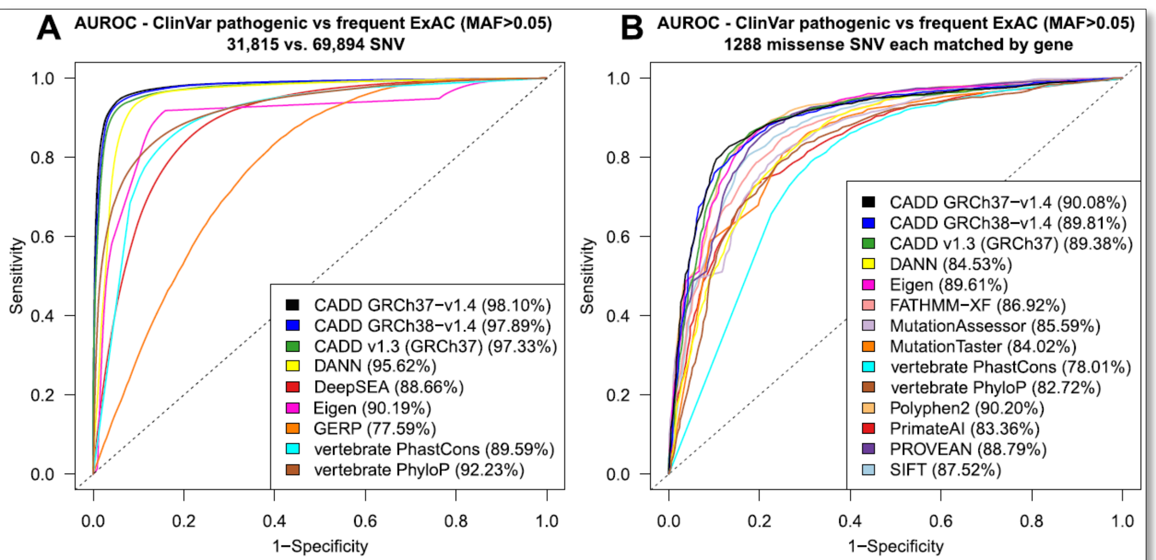
Today: Deep Learning for Human Genetics and Disease

1. Human Genetics: Inheritance, Mendel, Fisher, SNPs, STRs, alleles
2. 'Disease gene' hunting (locus, really): Common/rare alleles, Linkage vs. GWAS
3. Evolution/scaling of GWAS power: Sharing, inflection points
4. Challenge of fine-mapping: Co-inheritance, LD, Haplotypes, Recomb.
5. From locus to mechanism - Case study: FTO and Obesity
6. Challenges in disease mechanism inference
7. Machine learning tools for variant interpretation
    - Deep variant
    - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
    - DeepSEA

7. 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,*]

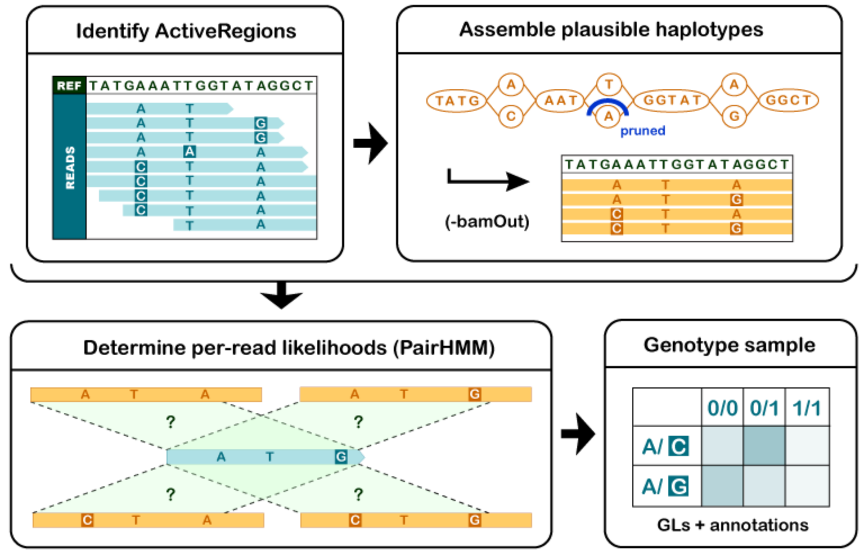
Large number of methods for variant prioritization

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

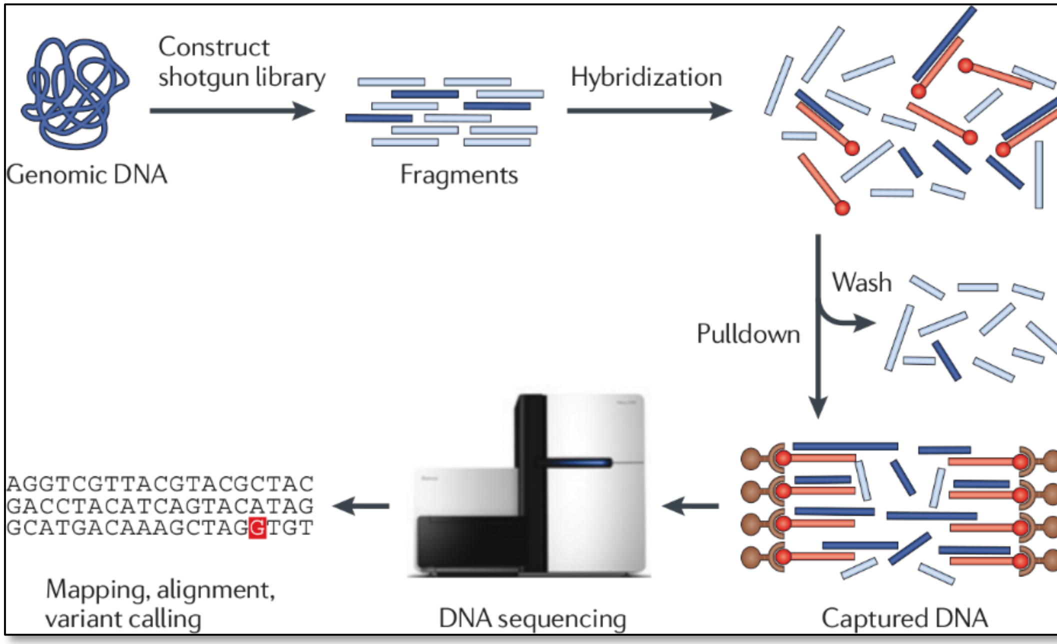


<u>Tour de Force, combining many methods:</u>

- **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

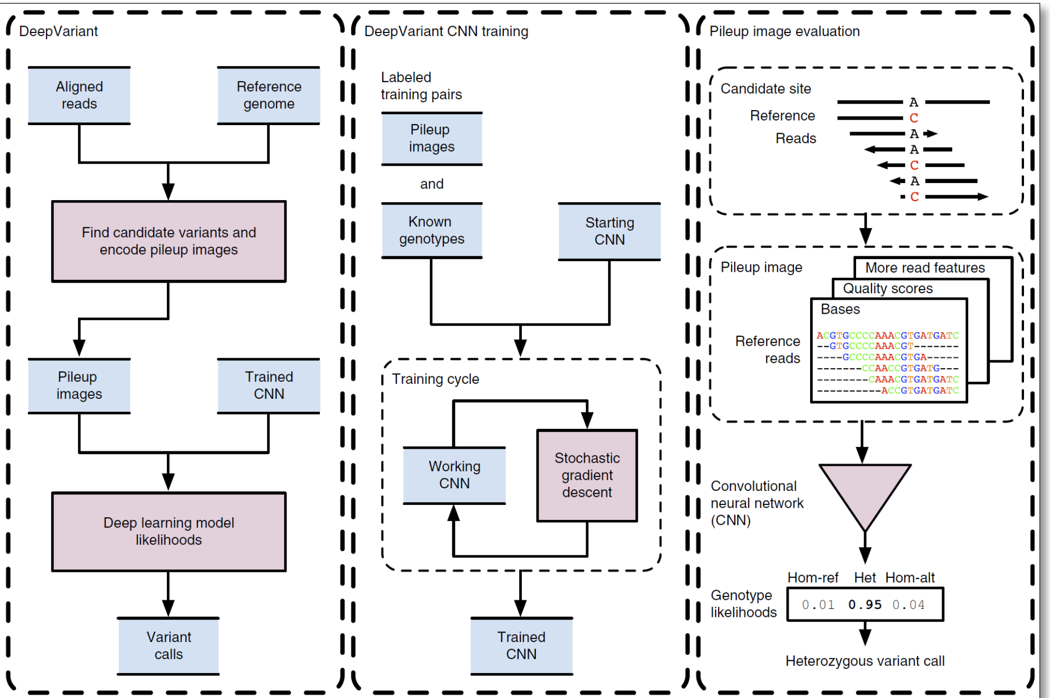
Exome variant calling: atlas2



- 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

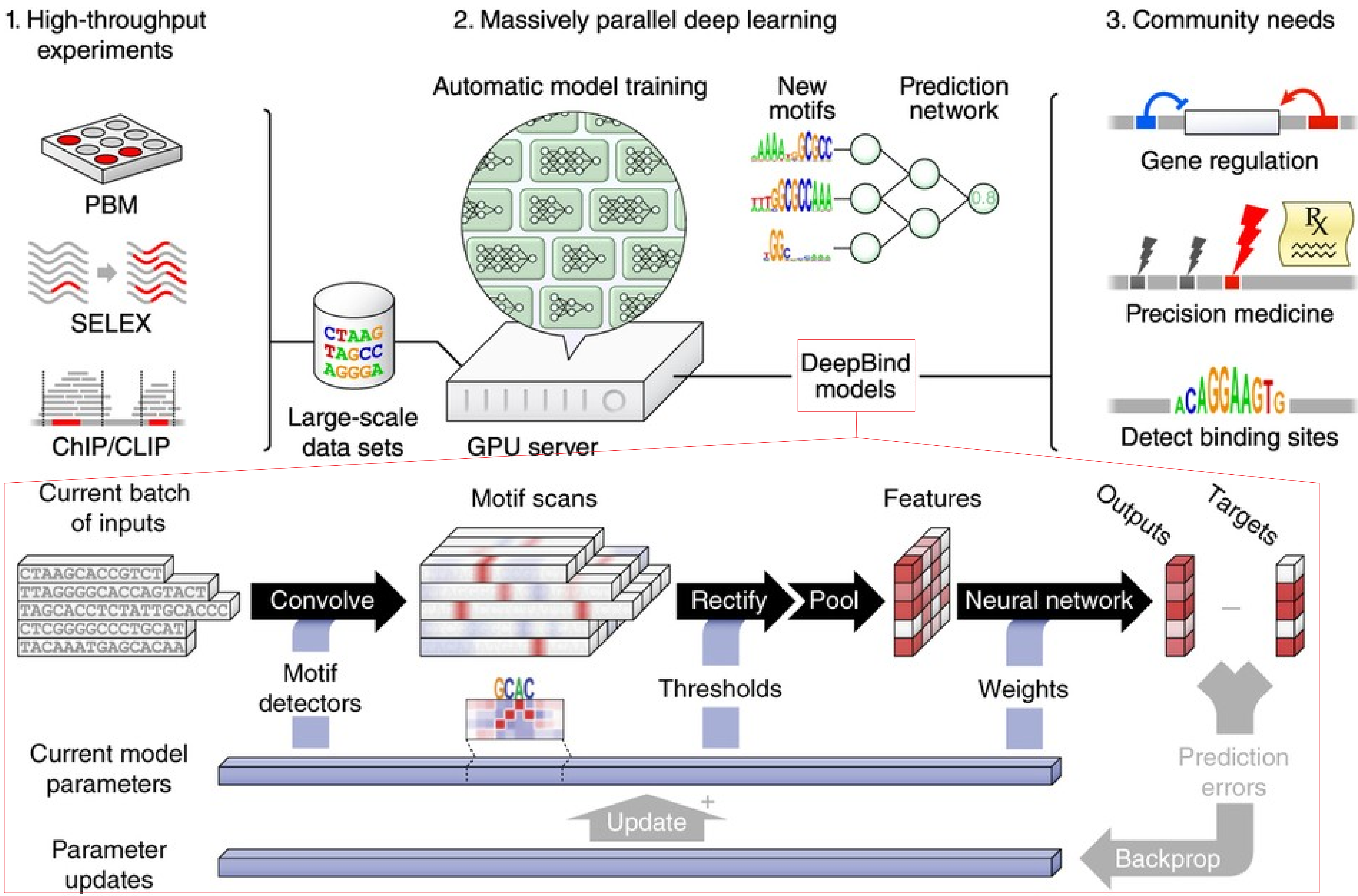


A universal SNP and small-indel variant caller using deep neural networks

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]

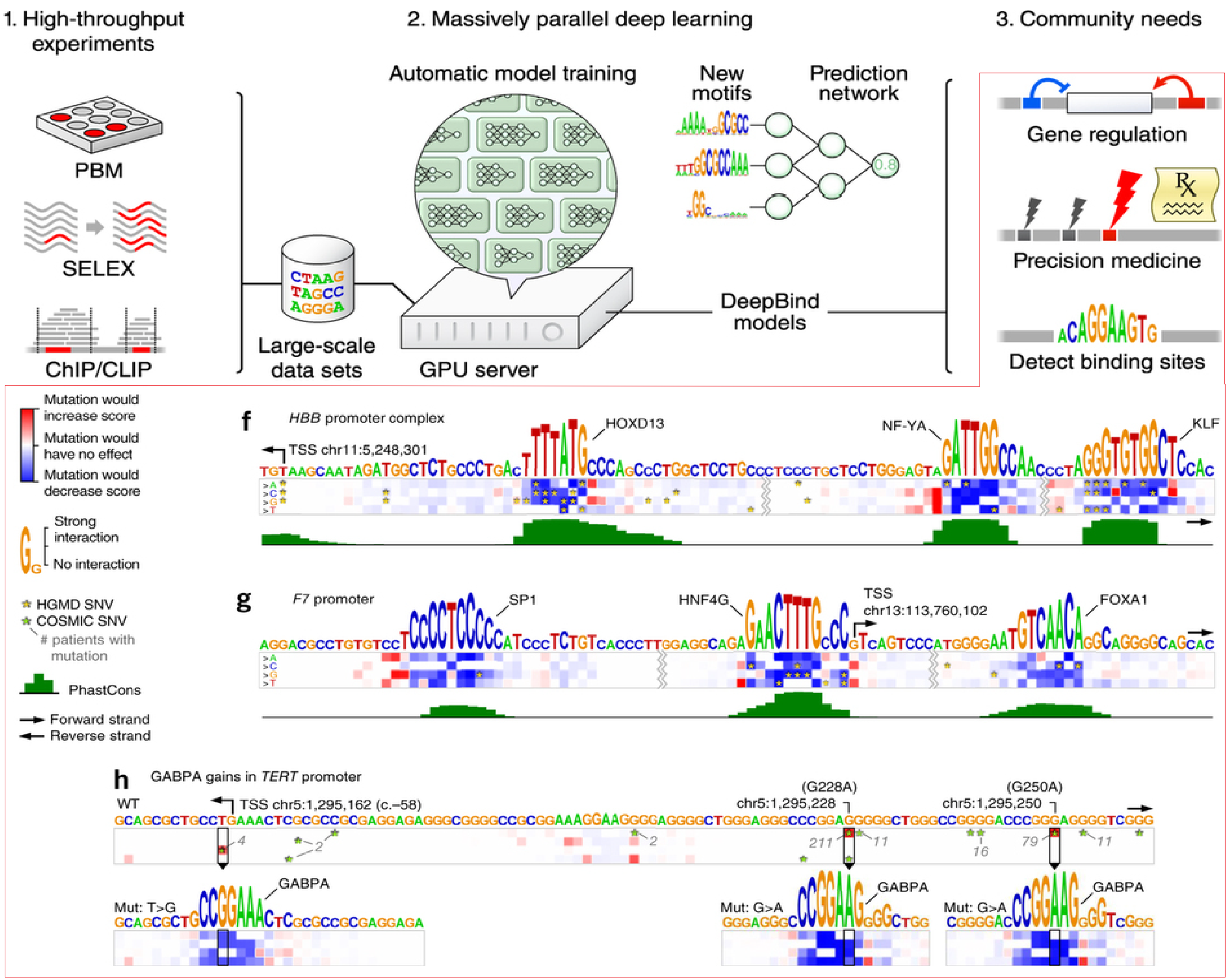
Method	Type	F1	Recall	Precision	TP	FN	FP	FP.gt	FP.al	Version
DeepVariant (live GitHub)	Indel	0.99507	0.99347	0.99666	357,641	2350	1,198	217	840	Latest GitHub v0.4.1-b4e8d37d
GATK (raw)	Indel	0.99366	0.99219	0.99512	357,181	2810	1,752	377	995	3.8-0-ge9d806836
Strelka	Indel	0.99227	0.98829	0.99628	355,777	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

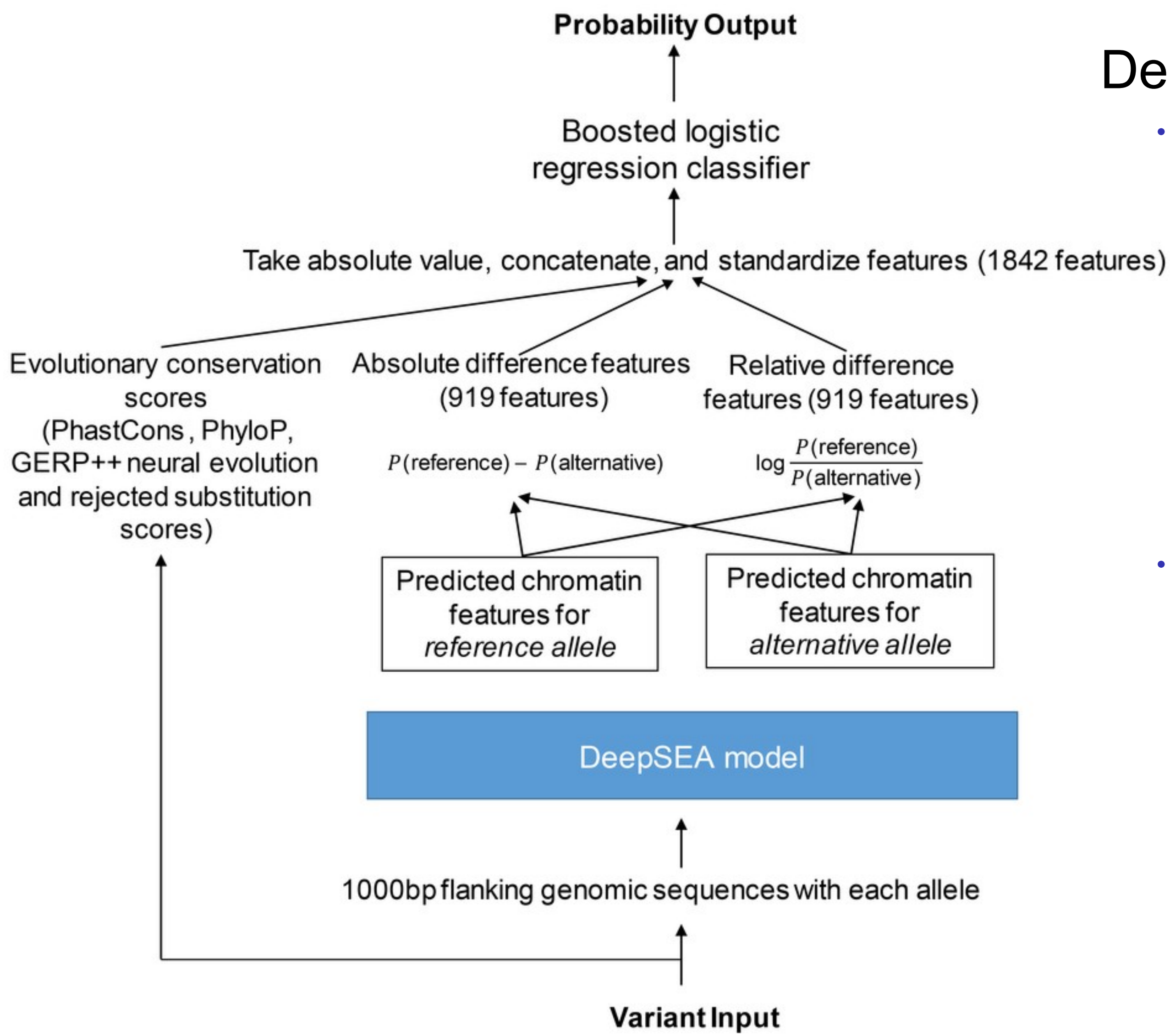
The key deep learning techniques:

- 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 (locus, really): Common/rare alleles, Linkage vs. GWAS
3. Evolution/scaling of GWAS power: Sharing, inflection points
4. Challenge of fine-mapping: Co-inheritance, LD, Haplotypes, Recomb.
5. From locus to mechanism - Case study: FTO and Obesity
6. Challenges in disease mechanism inference
7. Machine learning tools for variant interpretation
   - Deep variant
   - Eigen, FunSeq2, LINSIGHT, CADD, FATHMM, ReMM, Orion, CDTS
   - DeepSEA