18.700/18.701/6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution
Prof. Manolis Kellis

Lecture 17 – Genetics, GWAS, PRS, Mechanism, Disease Circuitry

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS

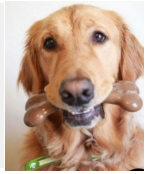
2. Polygenic scores (PGS)

3. From GWAS to biological insights

4. From Region to Mechanism / Circuitry

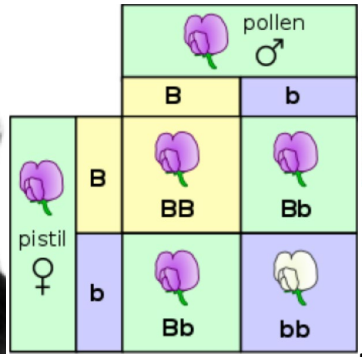
Genetics: Ancient Foreshadowings ➔ Mendelian traits ➔ Polygenicity



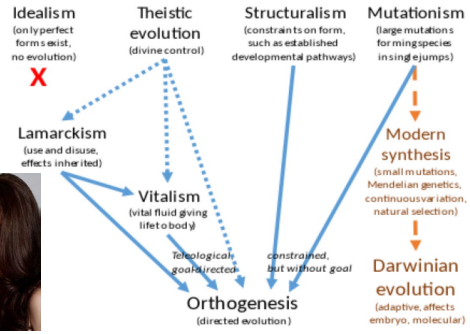
9000BC: Selective breeding of animals/plants

Inheritance: Eye/hair color long understood

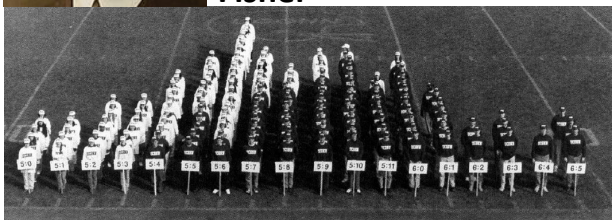
1866: Mendel: Discrete inheritance
No blending. Dominant/recessive alleles
Independent assortment



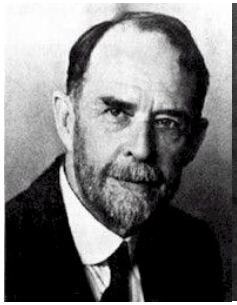


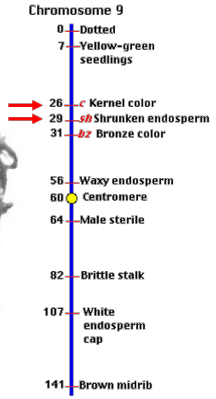
Biometrics: continuous phenotype variation.
Others: Saltationism, orthogenesis, vitalism, neo-Lamarckism, theistic evolution…

Fisher



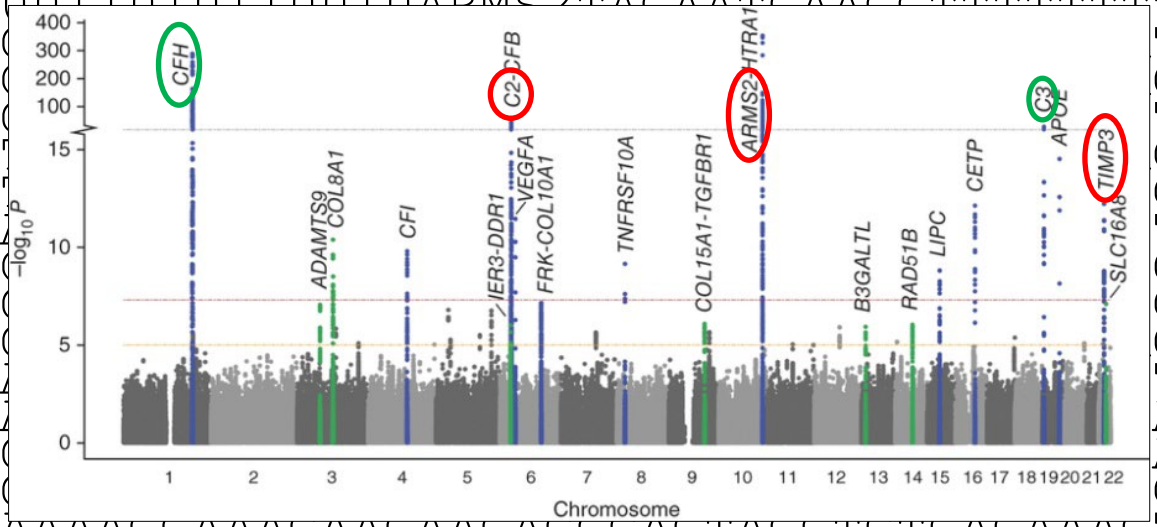
1918. Continuous phenotype variation explained by multiple Mendelian loci

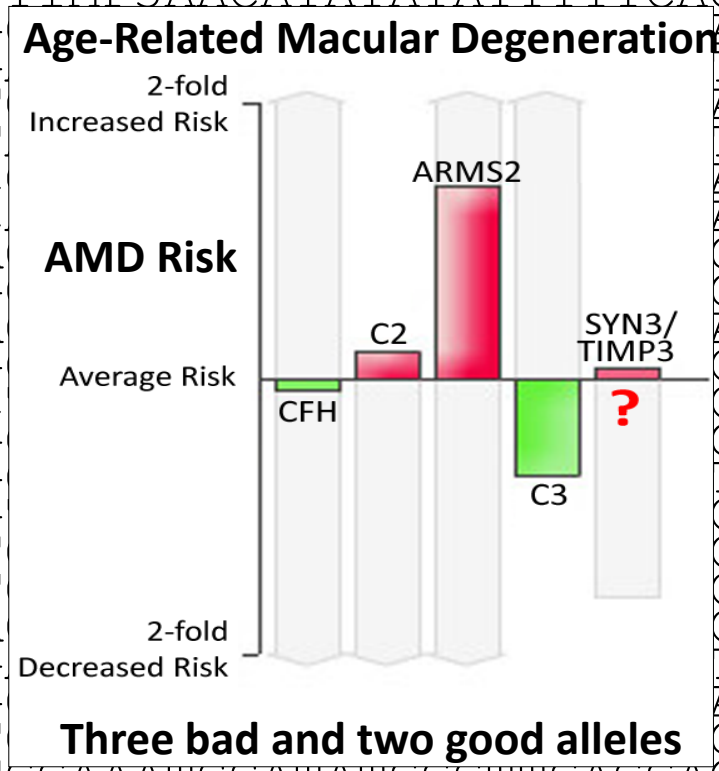




1913: Linkage/mapping, Morgan, Sturtevant
1980s: Mendelian Trait genes mapped

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



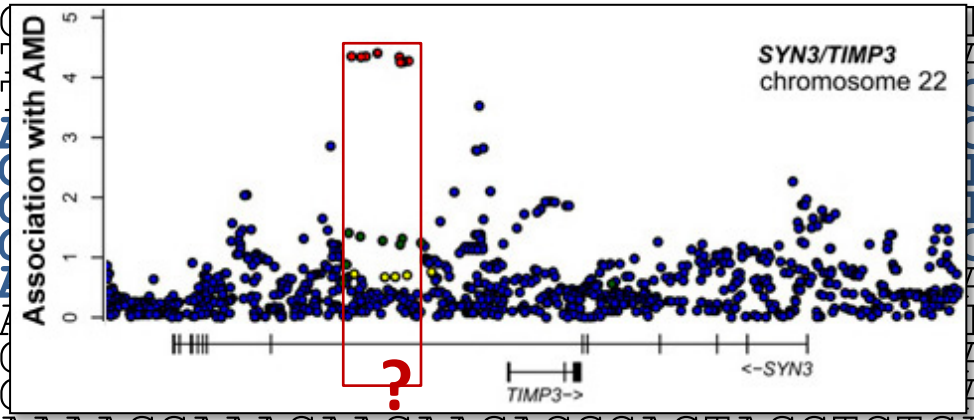


ARMS2

TIMP3/SYN3

C2

ARMS2gene: AAAGCTTCACAGATGATTTCAATGGATACTAGGGACCTCTGTTGCCTCCT
CTGGCAGAGCAGGACTGAGGGGTGGACCCTCCCTGAGACCACCCAACAATTCAGGGTGGA
TTTTTGTGTGACGGGAAAAGACAATGCTCC
AGGGGACAGCACCTTTGTCACCACATTATG
CAGGACCGATGGTAACTGAGGCGGAGGGGA
CCTCGGTGGTTCCTGTGTCCTTCATTTCCA
TTGGTGGAGAAGGAGCCAGTGACAAGCAGA
TCCCAGCTGCTAAAATCCACACTGAGCTCT
CCCAGAGGAGGTTCCAGCAGCCTCAGCACC
AGGCCTTACCCCAGACCTATTGAATCAGAA
TTAACCAGCCTTCAGGTGCTTCTGATGCAT
AGGCCTAGGCAATTCAGCCTTCCTCTGGTT
ATACTTCCAATGCACCTGCTACATGCCAGA
AAAAGGAAACAAGAAGACGCAGTAGGTCTGAGAAAGTGATGGGGTGAGCAGAAACCCAAA
GCTTATAGAAGGCCATCTGAGTGGCCCCTCAAGCCGGTGAATTGGCTTTAGGGTTTACTG

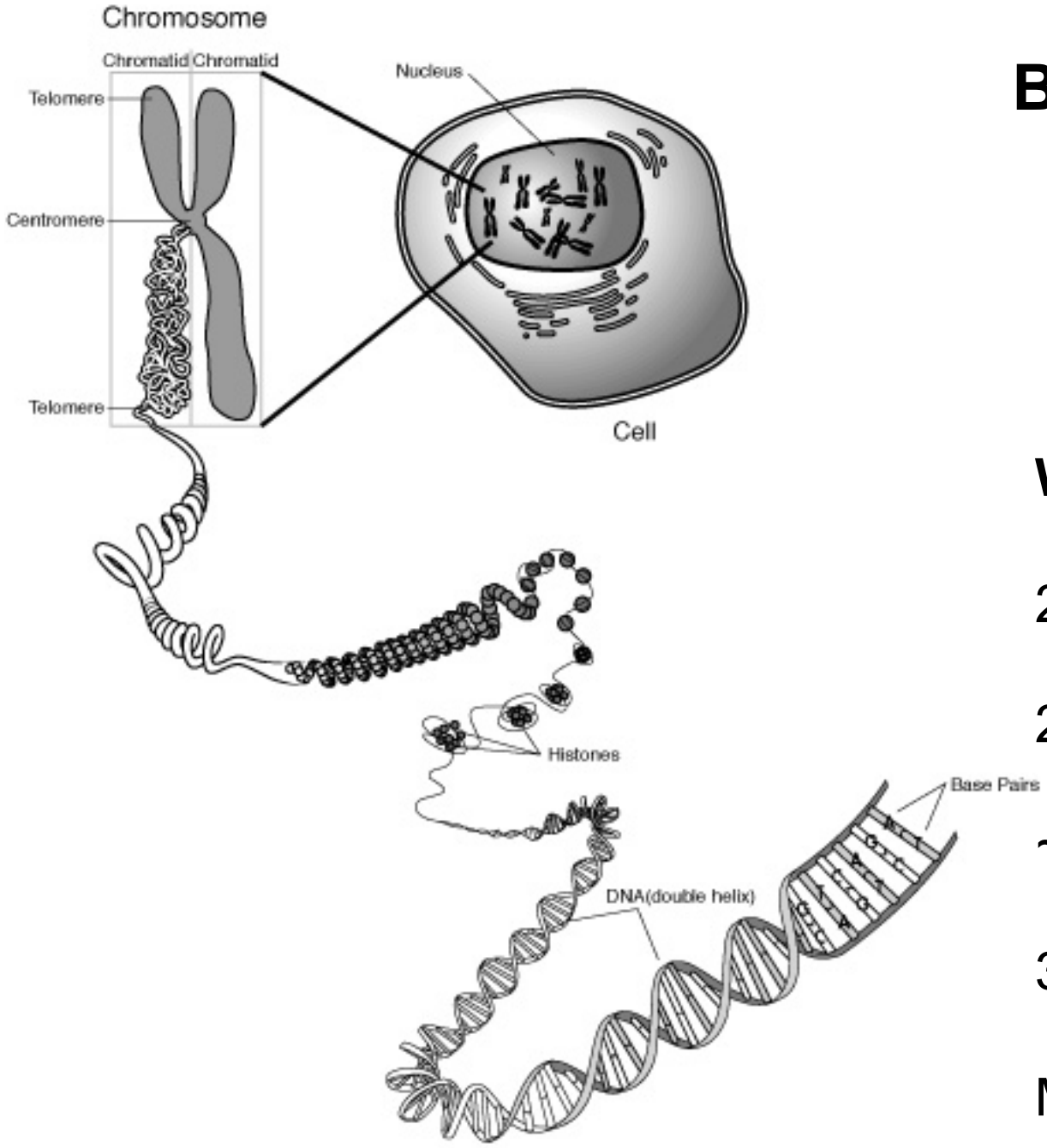


SYN3/TIMP3: ATATATTTTTCAGTGGCAGGAAGTCTTGCCCGAGGTGGGAATGTTACTG
GGTTAATATCTGGGGGAAAGAGAAATATTTTTCCCTTTGTTAGCTGGCTCTGGGCAGCCT
GAAAACTCTTGATCCTCTCTGTCTGCTGCTTGGGACATAATGACCTGCTTTCAATCCCTT
TCAATTACAGGATTTCTGATAGGAATTTGGAAAACAACCTAAATCCCAAAGCTTGGATGG
TAGCCCATGCTTCATTCCACGTCTCTGTACCCAGTTTTTCAAAGAGATTTTTTTTTTTCA

C2gene: TGTTTCCTTGACTGGCAGCTCAGCGGGGCCCTCCCGCTTGGATGTTCCGGGAA
AGTGATGTGGGTAGGACAGGCGGGGCGAGCCGCAGGTGCCAGAACACAGATTGTATAAAA
TGGGCTTCCAGGCAGCCCTGGAAGTCAAGAGAACACTCAGAAATGGGGAGGGAGAAGCAG

Questions for this module:

1. **How do we catalogue all genetic variants in the human genome (SNPs)**
2. **How do we systematically associate them with disease (GWAS)**
3. **How do we use GWAS to understand disease mechanism (Function)**
4. **How can we translate these insights into therapeutics (Manipulation)**



Building blocks of genetic variation

Within each cell:

2 copies of the genome

23 chromosomes

~20,000 genes

3.2B letters of DNA

Millions of polymorphic sites

Types of genetic variation

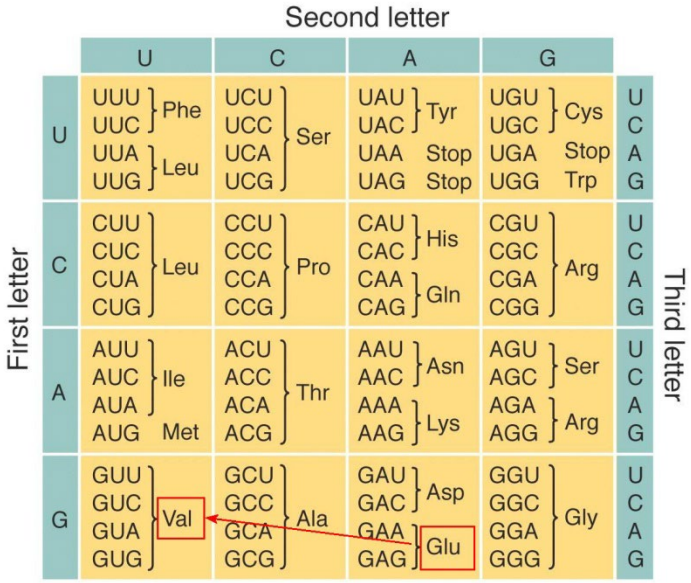
- **99% of DNA is shared** between two individuals
- Variation in the remainder explains all our **predisposition** differences
- **Remaining** phenotypic variation: environmental/stochastic differences

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

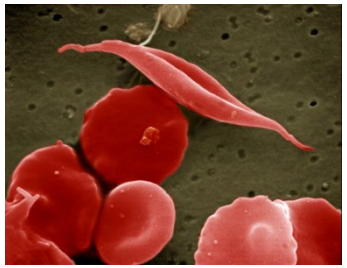
Single-nucleotide polymorphisms (SNPs)

CATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTG

CATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTG



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

Beyond SNPs: Tandem repeats and Indels

- Variable number tandem repeats

9 TCACAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT

10 TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT

12 TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTGCATTT

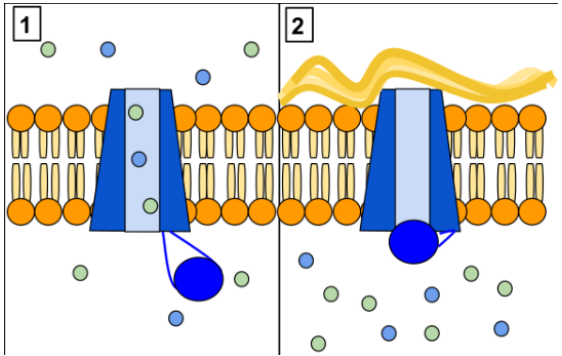
> 30 Huntington's Disease

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



CFTR Sequence:

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

Deleted in ΔF508

ΔF508 CFTR Sequence:

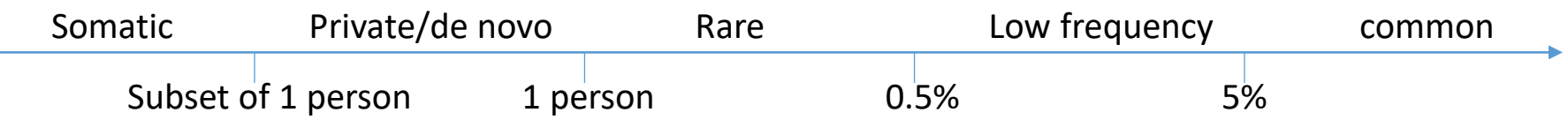
Nucleotide	ATC	ATT	GGT	GTT
Amino Acid	Ile 506	Ile	Gly	Val

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

Distinguishing the two alleles:

- Matching the human reference sequence (reference/alternate)
- Being more frequent in the population (major/minor)
- Matching the most recent common ancestor between human and chimpanzee (ancestral/derived)
- Based on their disease association (risk/non-risk)

Classifying variants by minor allele frequency:



Example: rs189107123

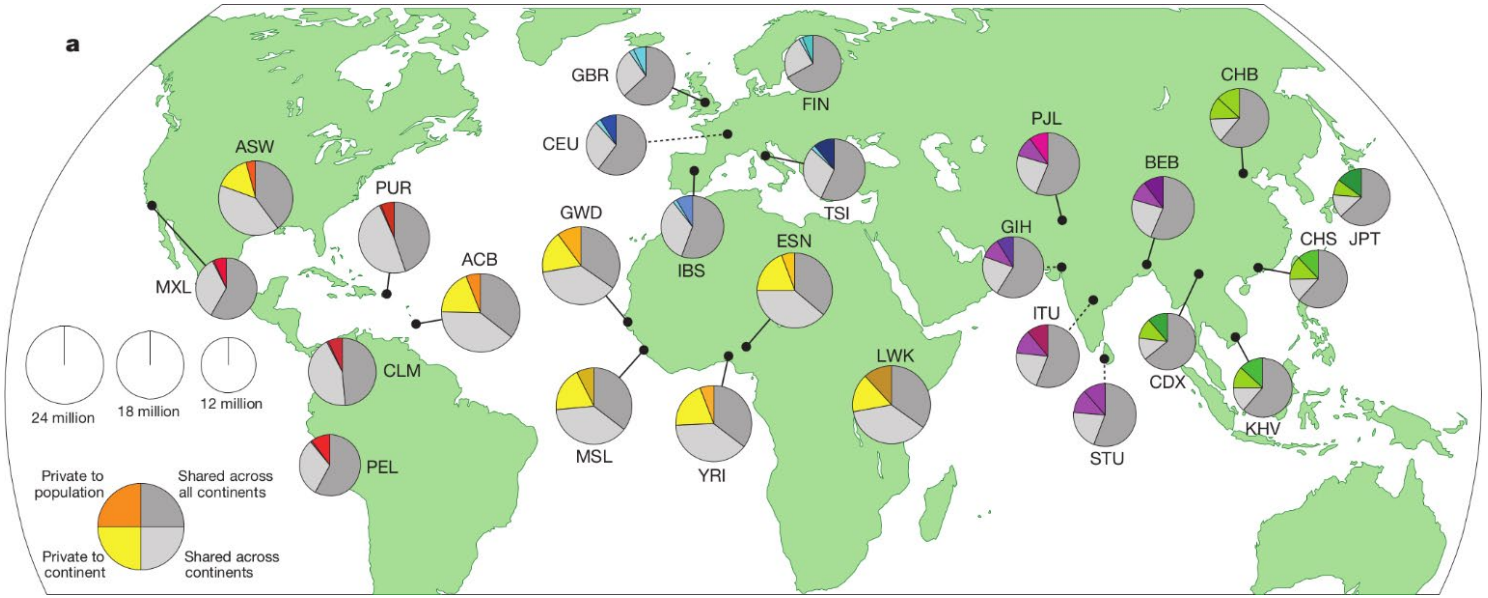
GAGGAGAACG[C/G]AACTCCGCCG

Reference allele: C

Minor allele: G (frequency 0.03 in Europeans)

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

Cataloguing genetic variants: Thousand Genomes Project



Esan	ESN	
Gambian	GWD	
Luhya	LWK	
Mende	MSL	
Yoruba	YRI	
Barbadian	ACB	
African-American SW	ASW	
Colombian	CLM	
Mexican-American	MXL	
Peruvian	PEL	
Puerto Rican	PUR	
Dai Chinese	CDX	
Han Chinese	CHB	
Southern Han Chinese	CHS	
Japanese	JPT	
Kinh Vietnamese	KHV	
CEPH	CEU	
British	GBR	
Finnish	FIN	
Spanish	IBS	
Tuscan	TSI	
Bengali	BEB	
Gujarati	GIH	
Telugu	ITU	
Punjabi	PJL	
Tamil	STU	

- 2,504 whole genome sequences at low depth (4x) across 26 subpopulations spanning the globe
- Develop sophisticated statistical tools (**phasing, imputation**) to account for noise, known patterns of variation (**linkage disequilibrium**; next section)

Thousand Genomes Consortium *Nature* 2016

Measuring known genetic variation: genotyping

- Key insight: Most **genetic variants** in an individual are recurrent in the population. Once they've been discovered/catalogued, build a **common array** for measuring them
- DNA microarrays were the key technological advance of the 1990s
- Idea: fragments of sample DNA containing SNPs will **hybridize** (reverse complement) to array **probes** (engineered DNA fragments)
- Tag fragments with fluorescent compound, use intensity to recover which probes were bound, which alleles were present in the sample
- Today: still the fundamental technology used in large-scale population genetic assays (GWAS, 23andMe)
- Next: study disease associations across populations, requiring new array designs due to differences in polymorphisms, LD across populations

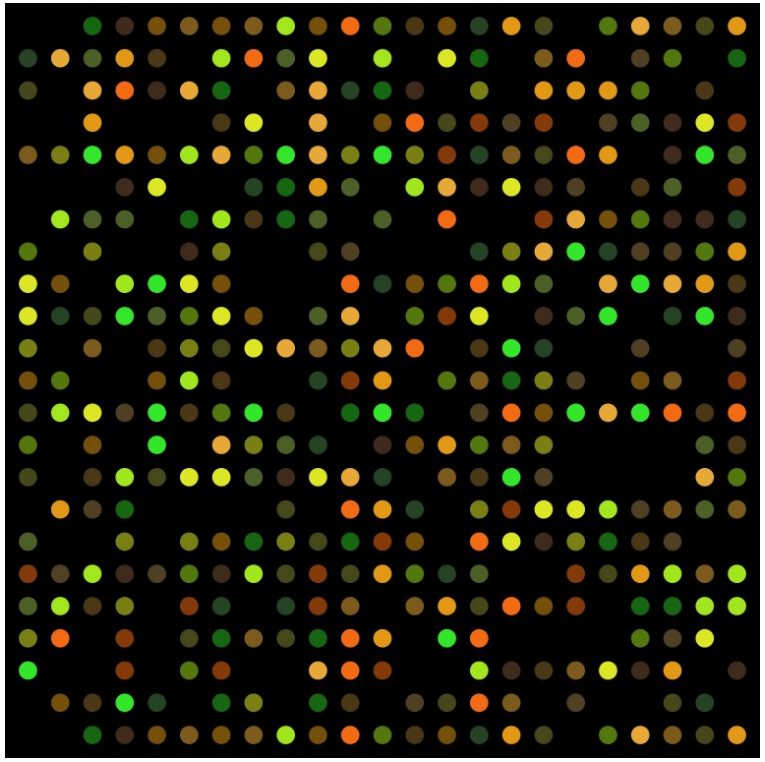
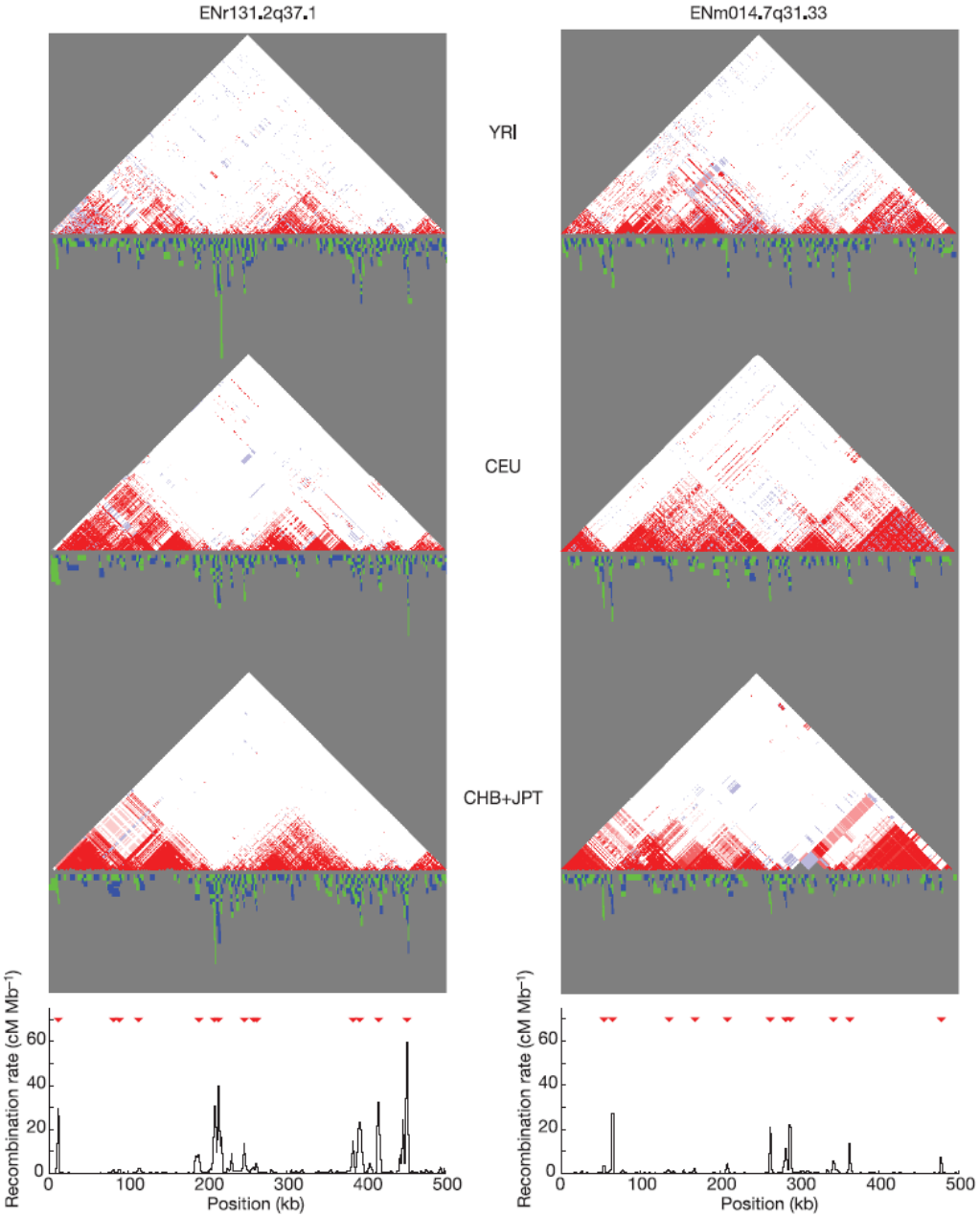


Image credit: Wikimedia Commons

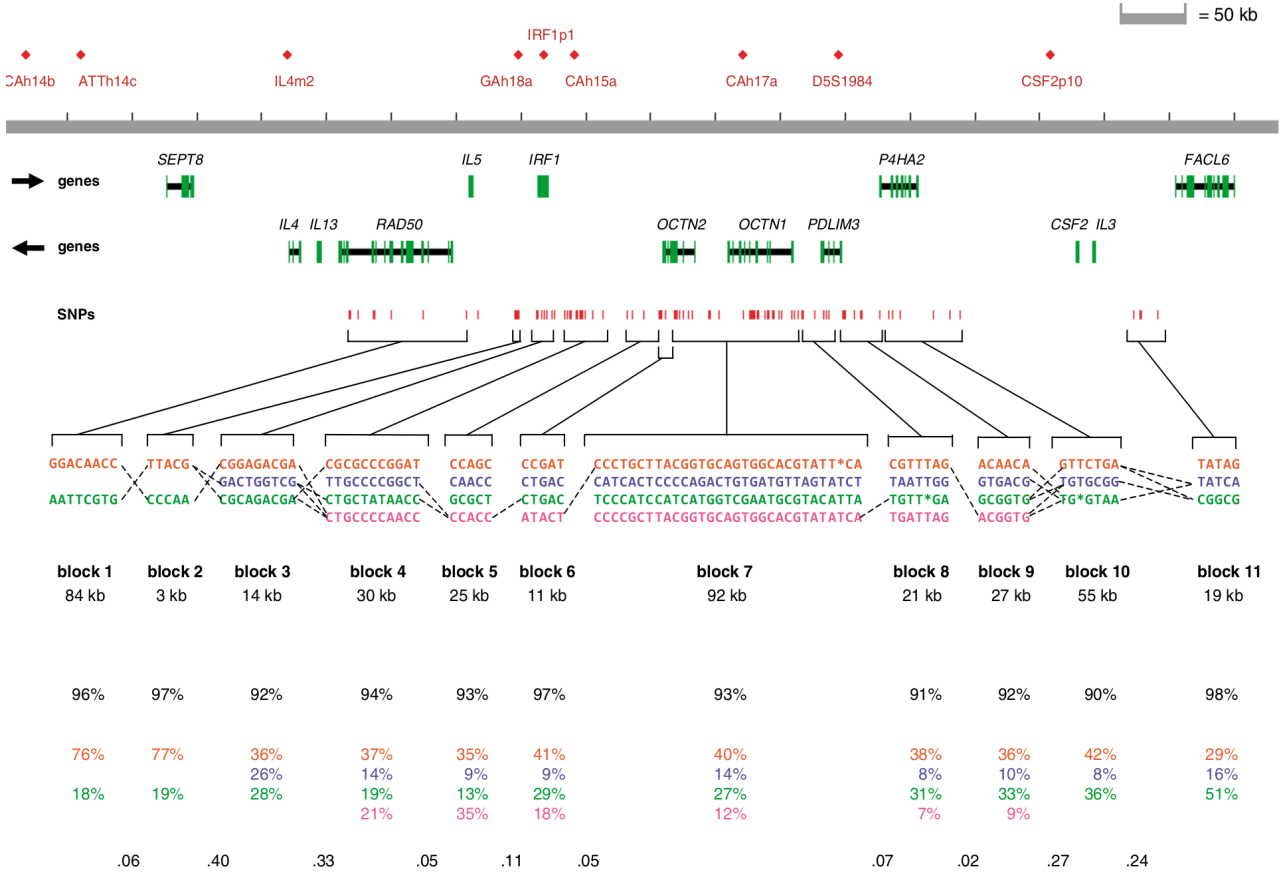
r^2 and recombination events across regions/populations



- 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)

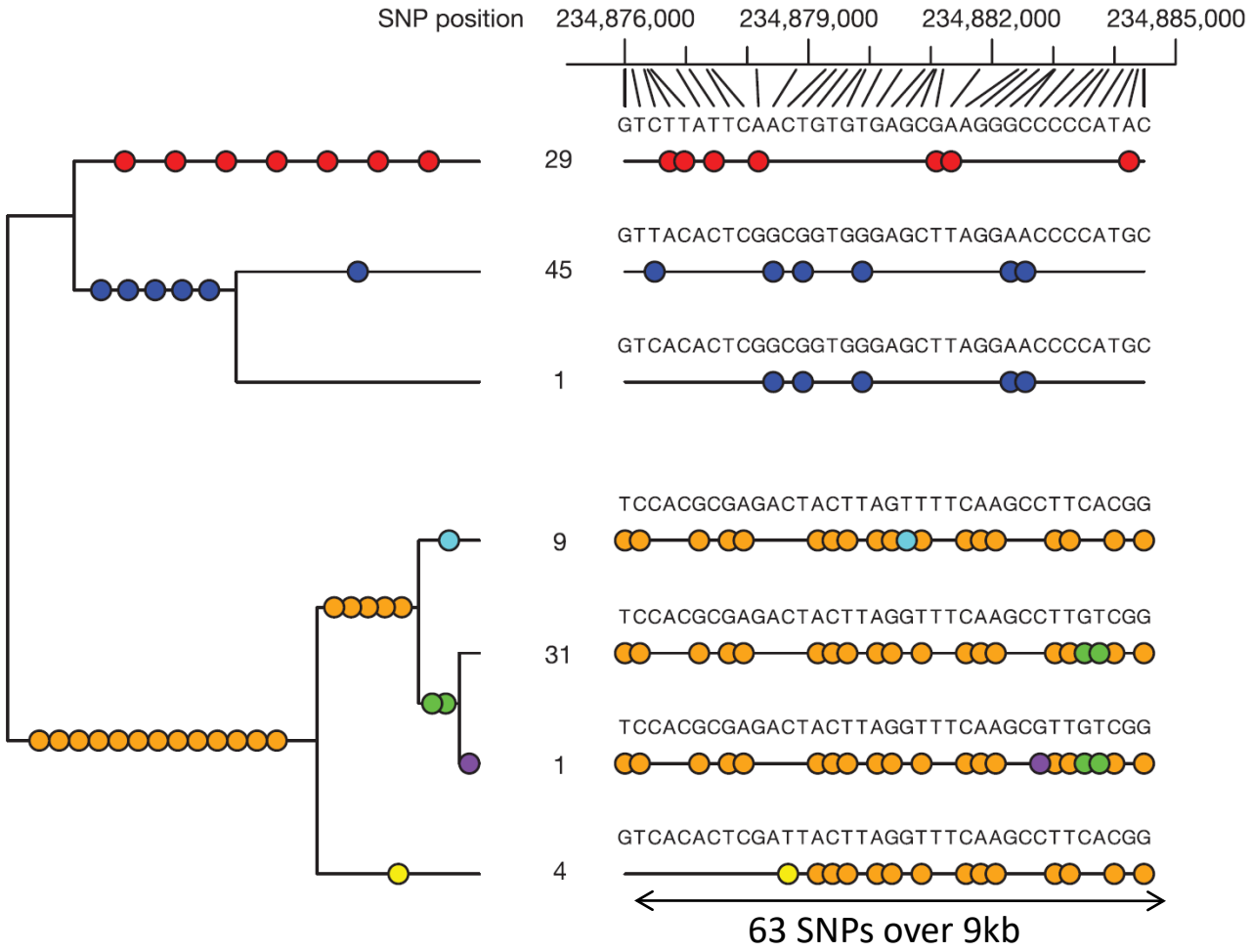
Long-range threading of haplotype blocks



- 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

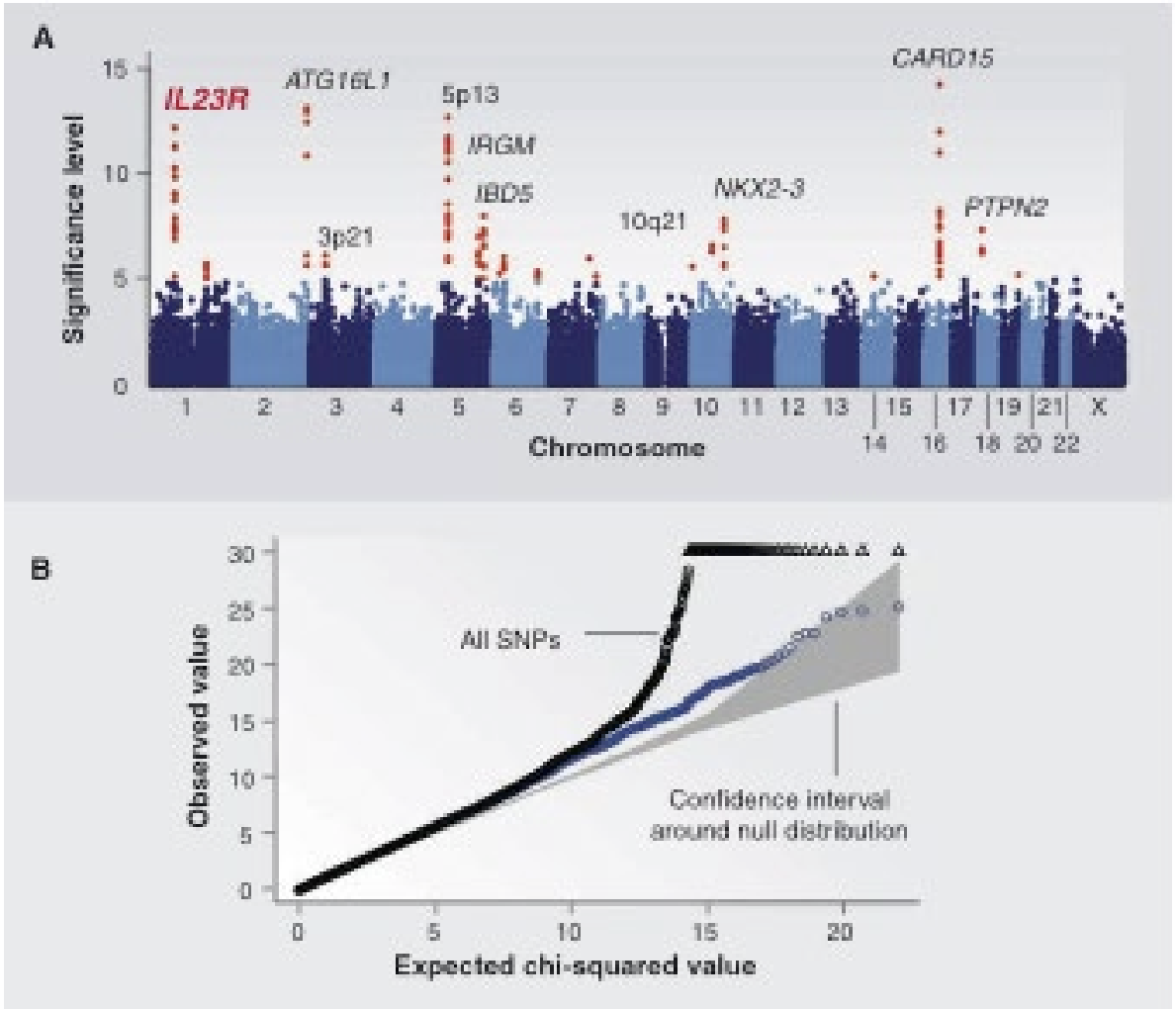
Daly et al *Nat Genet* 2001

Mutational history of multiple haplotypes



- 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)

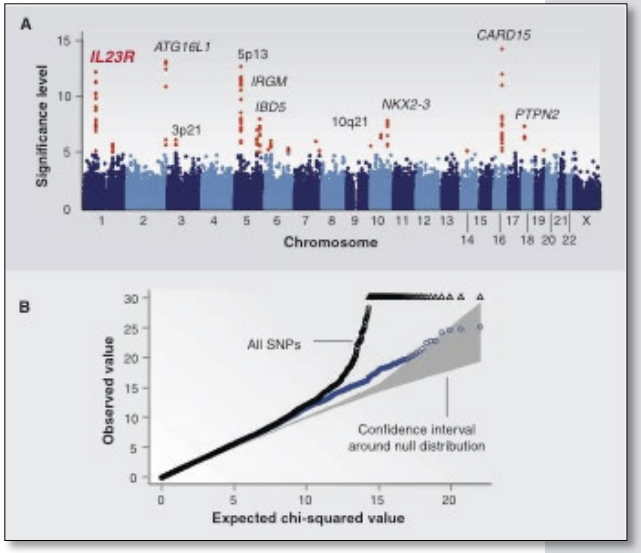
Genome-wide Association Studies (GWAS)

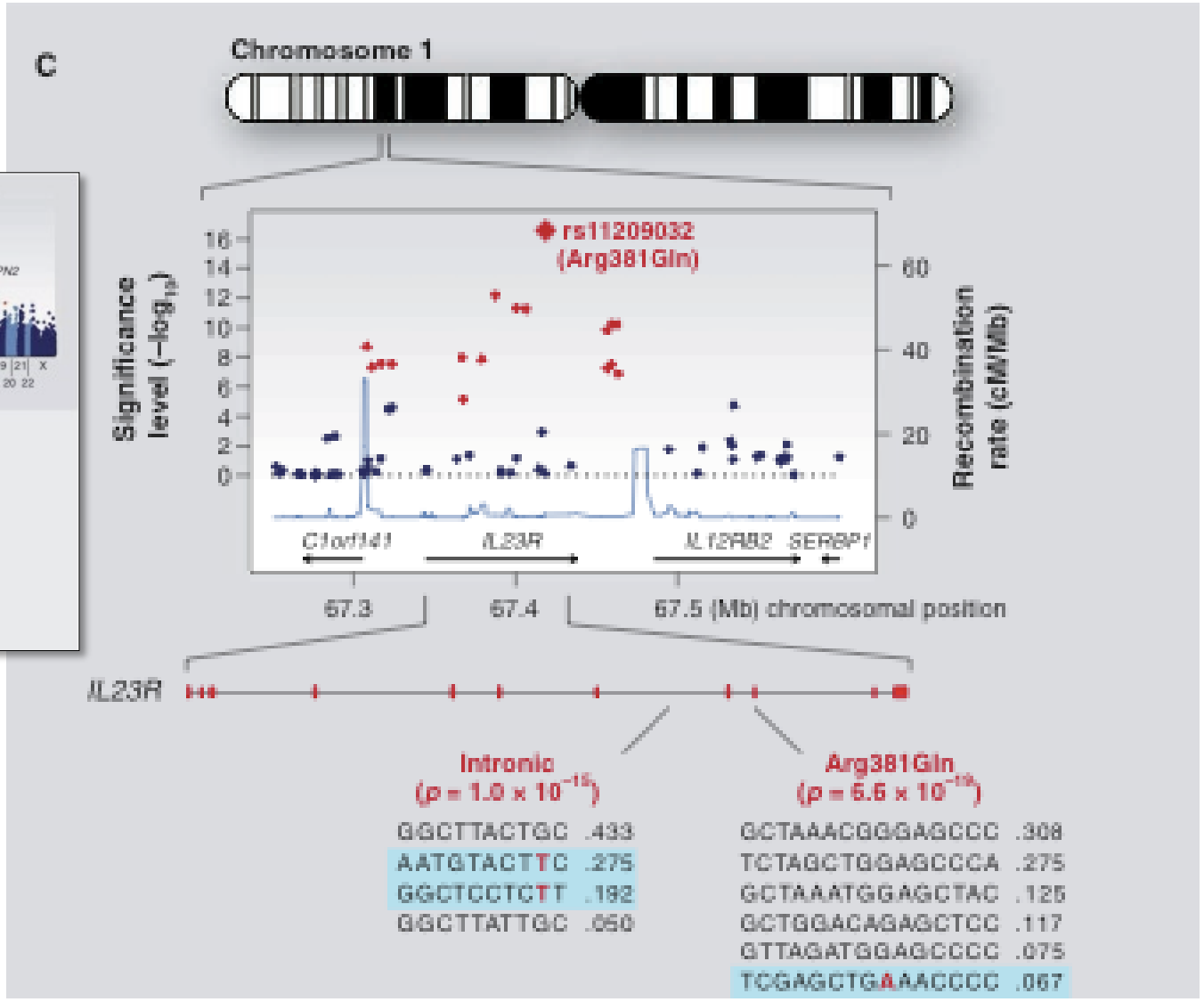


‘Manhattan’ plot

Q-Q plot

Fine Mapping: GWAS locus view





Testing for association

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

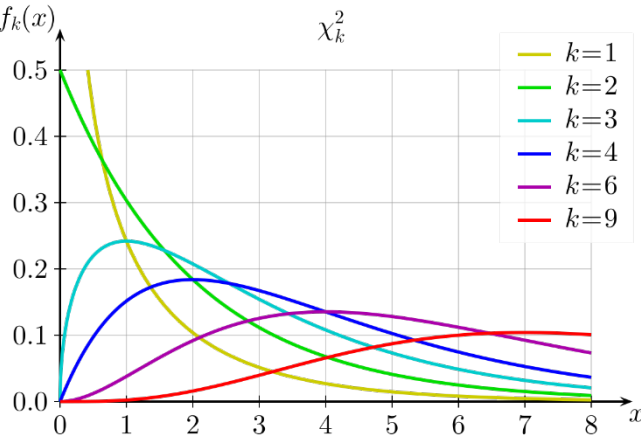
rs11209026	Allele A	Allele G
Cases	22	976
Controls	68	932

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

Expected	Allele A	Allele G
Cases	47	951
Controls	47	953

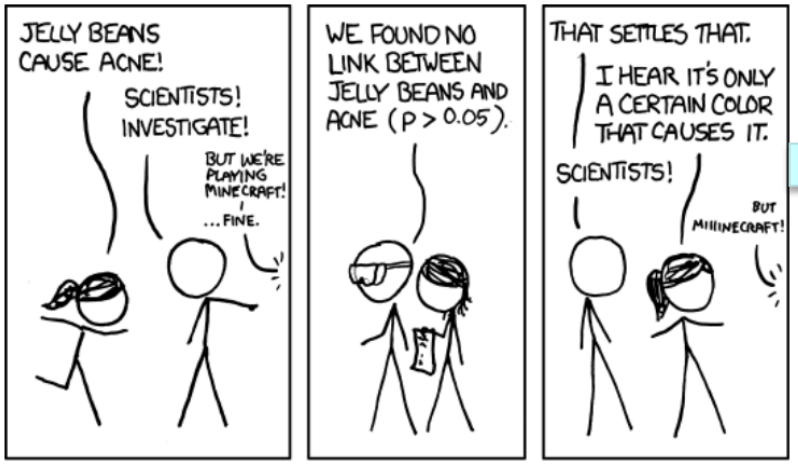
(O-E)^2/E	Allele A	Allele G
Cases	13.4	0.7
Controls	9.2	0.5

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

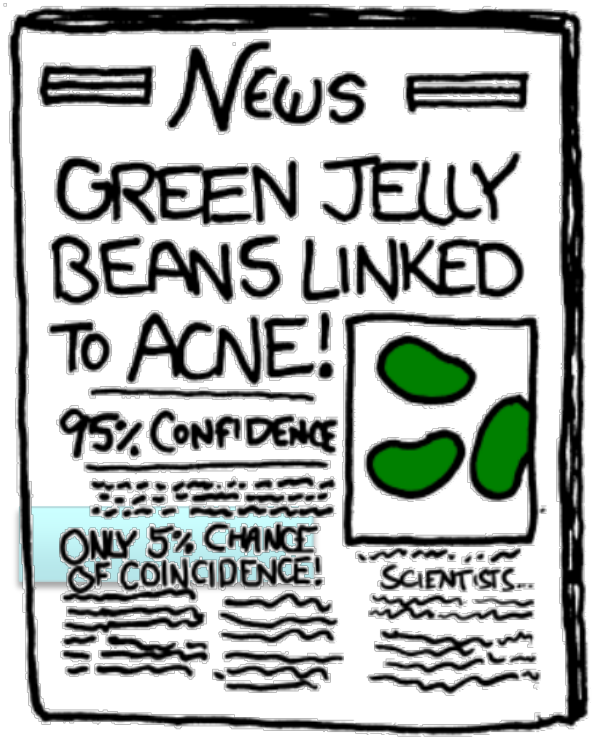


Simplest tests (single marker regression, χ^2) rule the day.
Association results requiring arcane statistics.
Complex multi-marker models are often less reliable

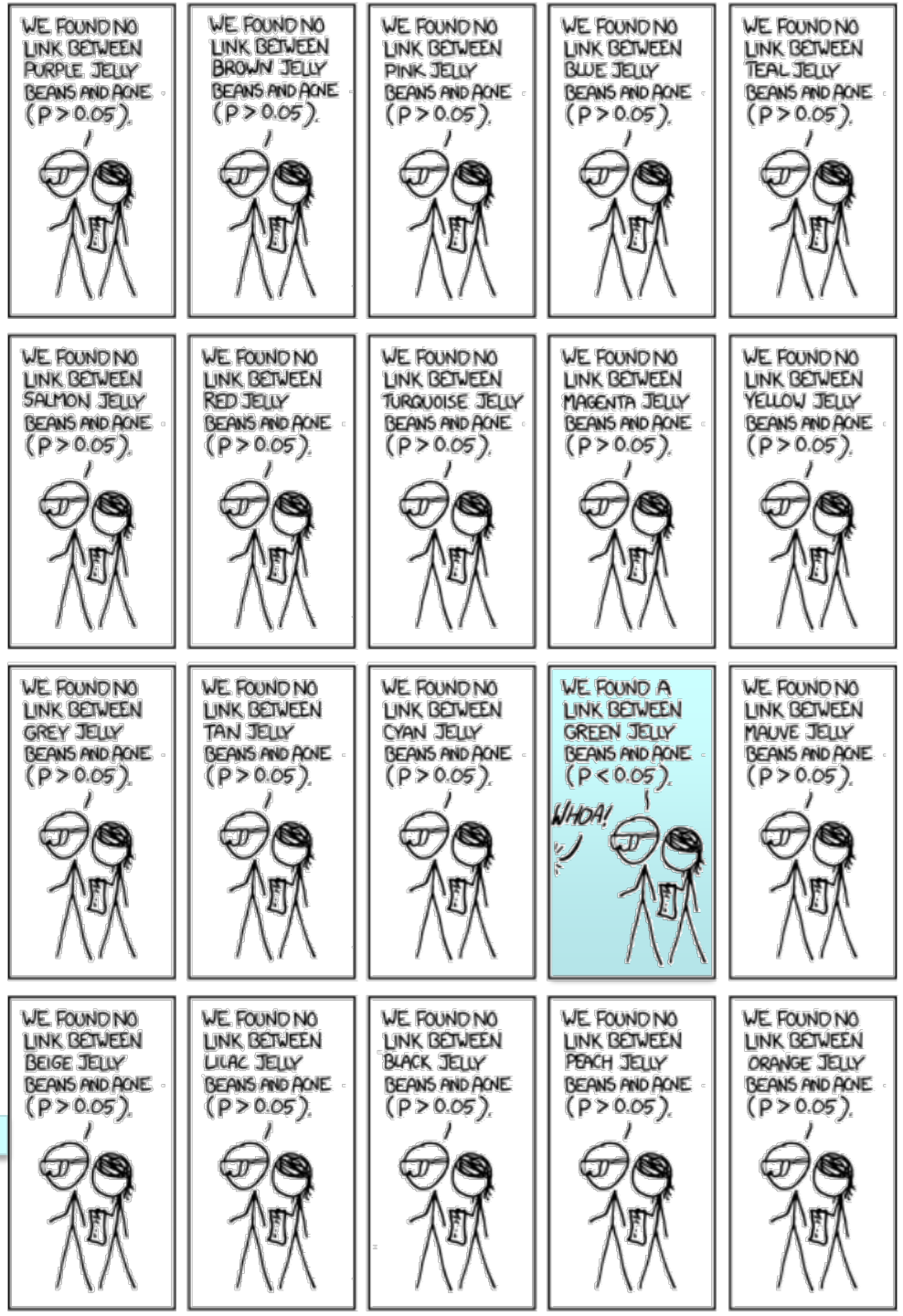
Multiple hypothesis testing



https://xkcd.com/882/
https://www.explainxkcd.com/wiki/index.php/882:_Significant



So, uh, we did the green study again and got no link. It was probably a-- "RESEARCH CONFLICTED ON GREEN JELLY BEAN/ACNE LINK; MORE STUDY RECOMMENDED!"



Correcting for 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)

Best practices are key

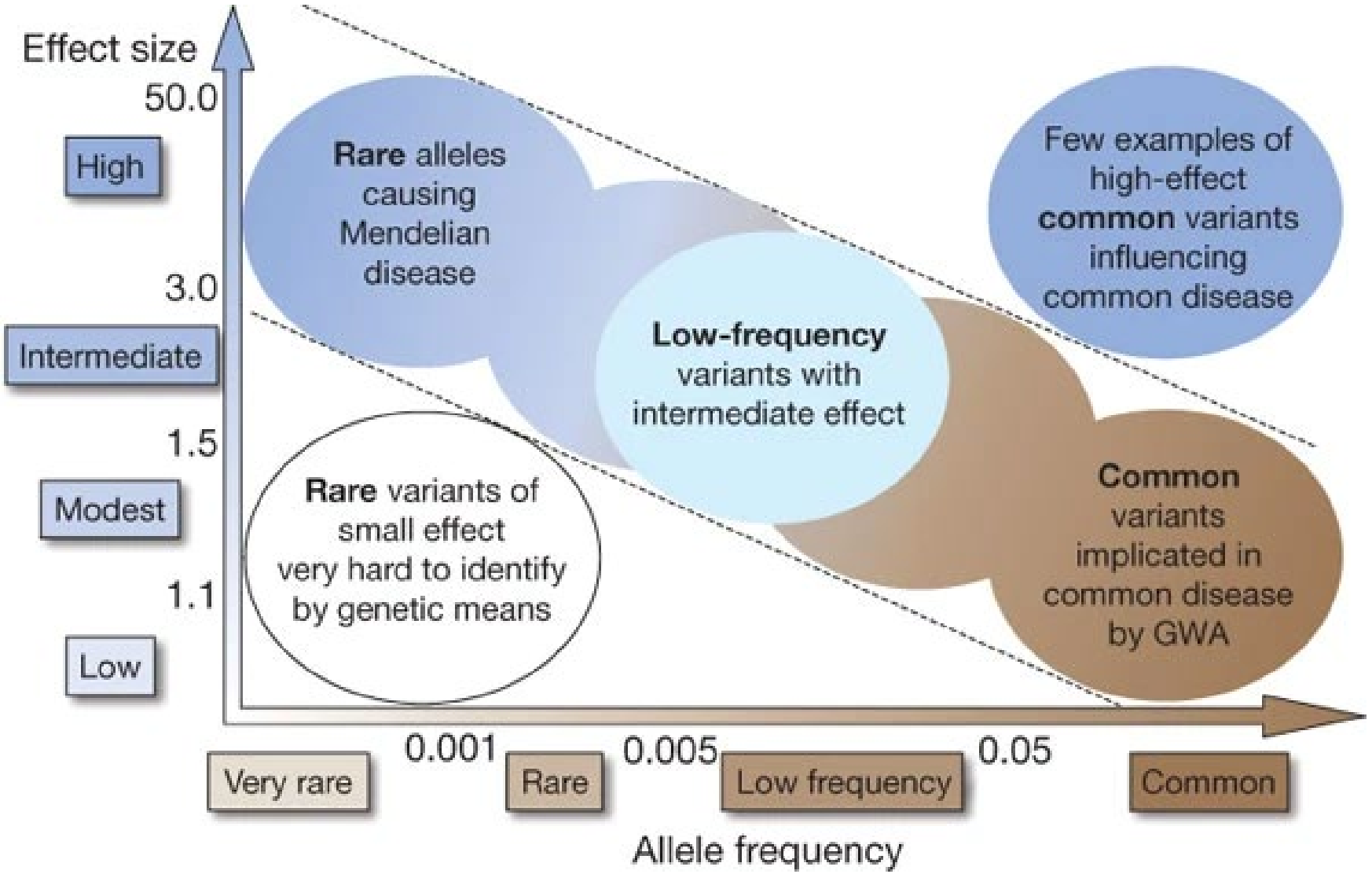
- Technical QC
  - Removal of failed SNPs, samples
- Genetic QC
  - Mendelian segregation and Hardy Weinberg Equilibrium
  - Estimating relatedness, gender
  - Population structure
- Analysis-based QC
  - Do initial runs of test statistics show inflation, biases towards missing data, specific allele frequencies

Reversing the curse: the story of GWAS

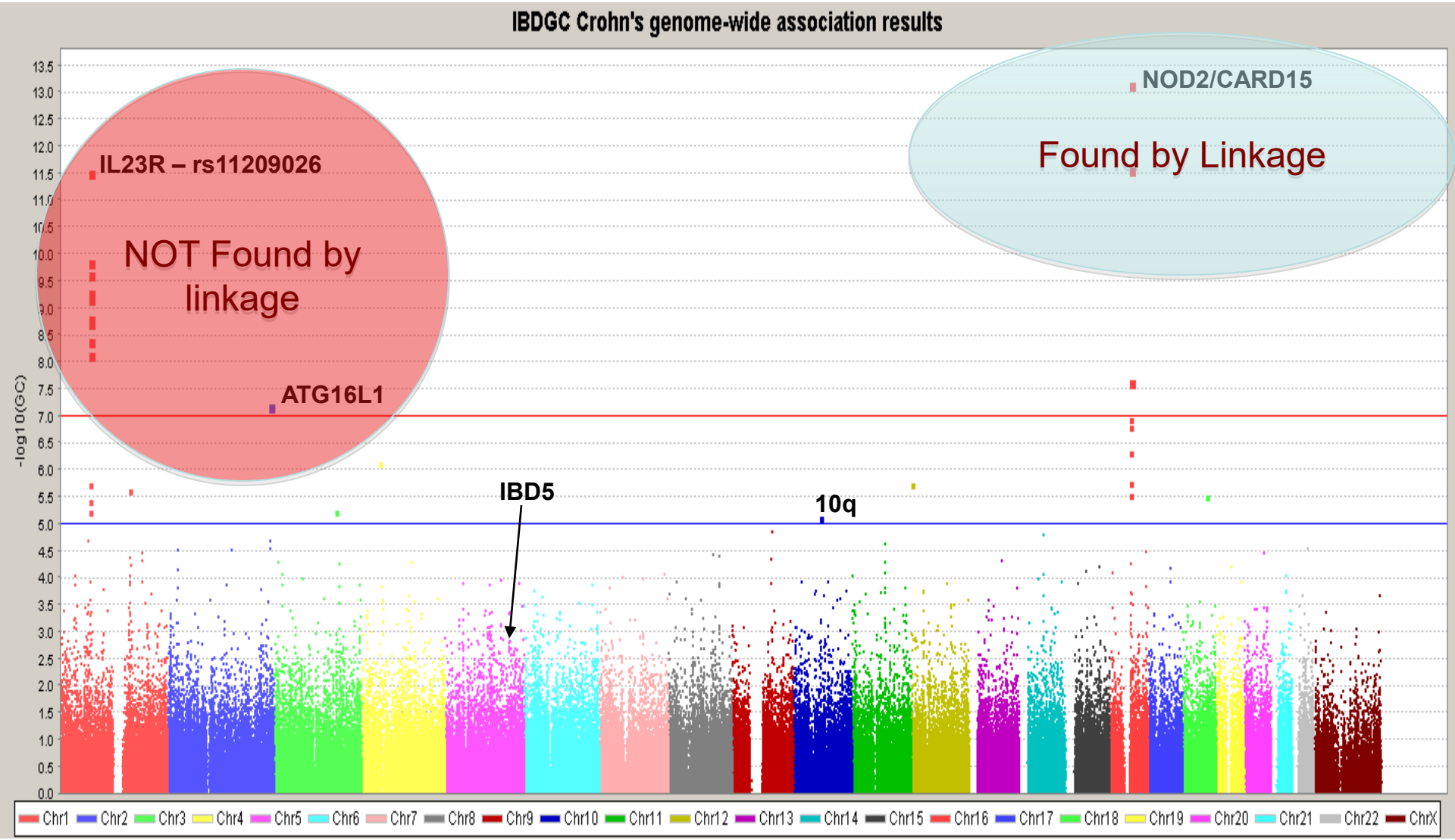
1996: Risch and Merikangas propose that a **p**-value of **5×10^{-8}** (equivalent to a **p**-value of 0.05 after a Bonferroni correction for 1 million independent tests) is a conservative threshold for declaring significant association in a genome-wide study.

2008: 3 groups publish empirically derived estimates based on dense genome-wide maps of common DNA and estimated appropriate dense-map numbers to be in the range of **2.5 to 7.2×10^{-8}**

Most common variants have small effects



Manolio et al 2009, PMID: 19812666

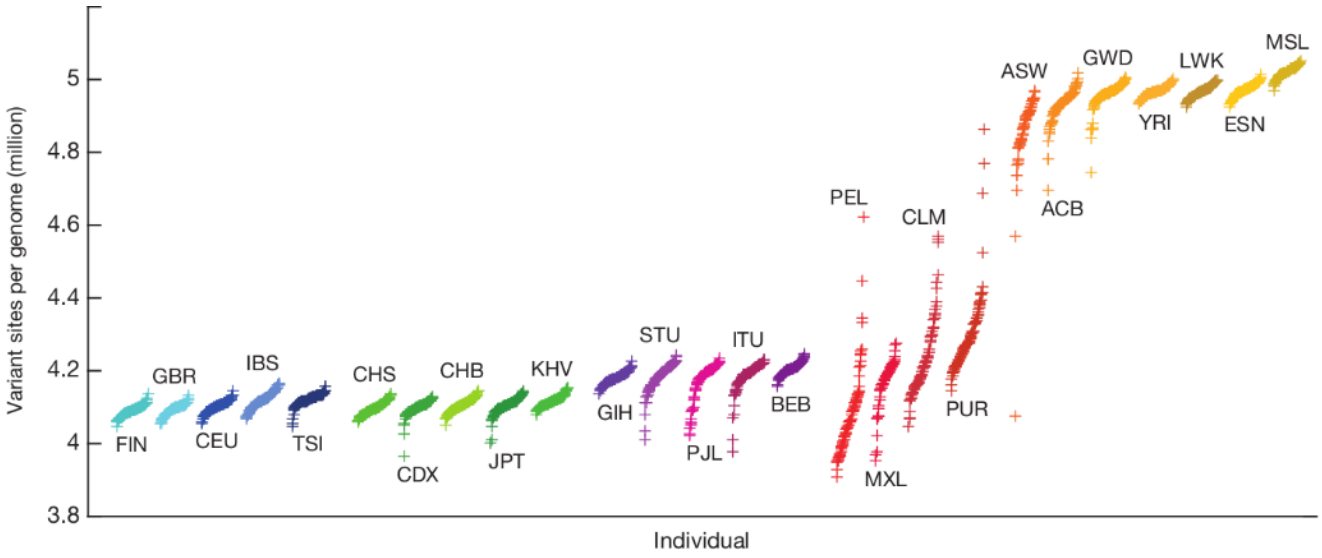

IBDGC Crohn's genome-wide association results
IL23R – rs11209026
NOT Found by linkage
ATG16L1
NOD2/CARD15
Found by Linkage
IBD5
10q
-log10(GC)
Chr1 Chr2 Chr3 Chr4 Chr5 Chr6 Chr7 Chr8 Chr9 Chr10 Chr11 Chr12 Chr13 Chr14 Chr15 Chr16 Chr17 Chr18 Chr19 Chr20 Chr21 Chr22 ChrX

Linkage vs. Association

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

Number of variants varies greatly by population



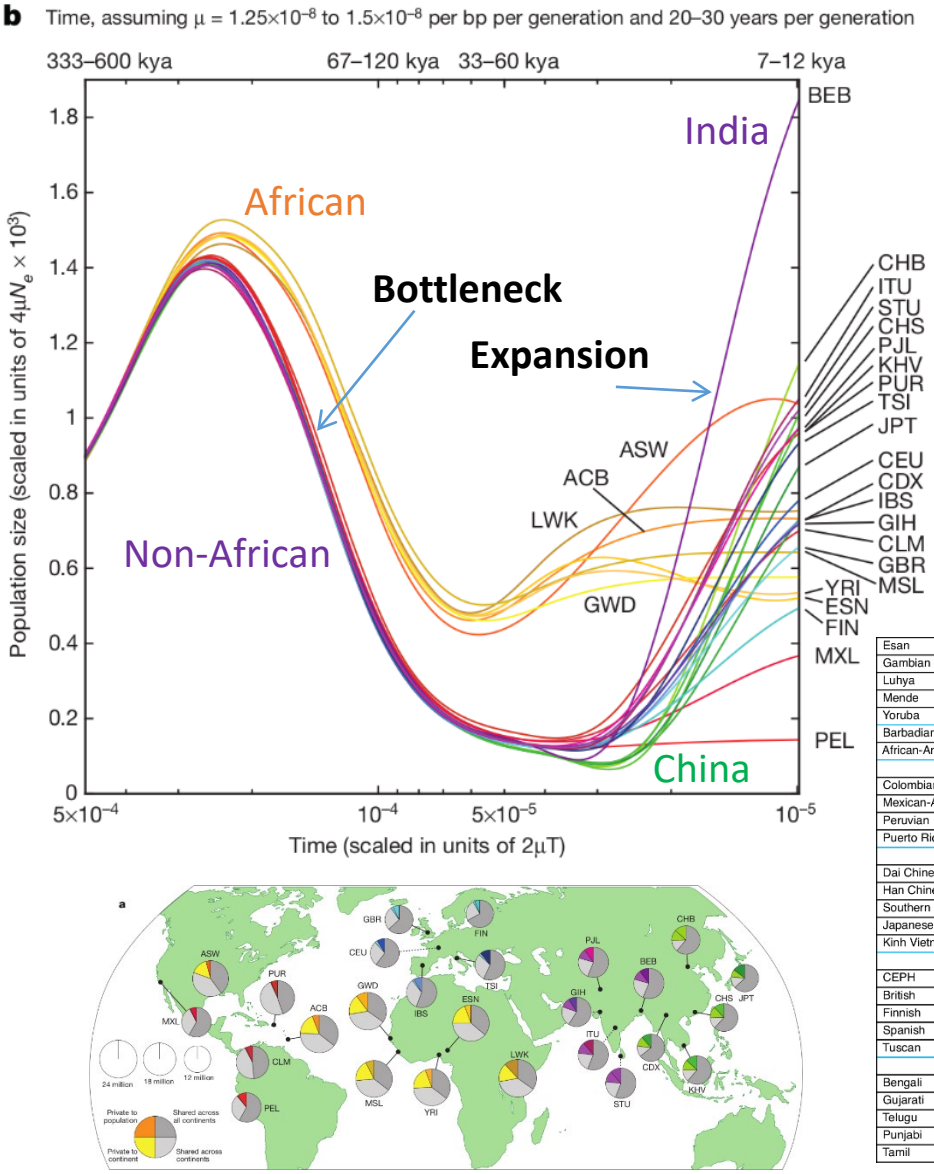
Africa	Esan	ESN
	Gambian	GWD
	Luhya	LWK
	Mende	MSL
	Yoruba	YRI
	Barbadian	ACB
	African-American SW	ASW
S.America	Colombian	CLM
	Mexican-American	MXL
	Peruvian	PEL
	Puerto Rican	PUR
E. Asia	Dai Chinese	CDX
	Han Chinese	CHB
	Southern Han Chinese	CHS
	Japanese	JPT
	Kinh Vietnamese	KHV
Europe	CEPH	CEU
	British	GBR
	Finnish	FIN
	Spanish	IBS
	Tuscan	TSI
India	Bengali	BEB
	Gujarati	GIH
	Telugu	ITU
	Punjabi	PJL
	Tamil	STU

- Over 100 million observed variants: 4-5M positions differ between each of us and the human reference
- Each of us carries 2-3K structural variants affecting 20mb of sequence
- Each of us carries hundreds of protein truncating variants, 10Ks of non-synonymous mutations
- African individuals have more variation in their genomes (**why?**)

Thousand Genomes Consortium *Nature* 2016

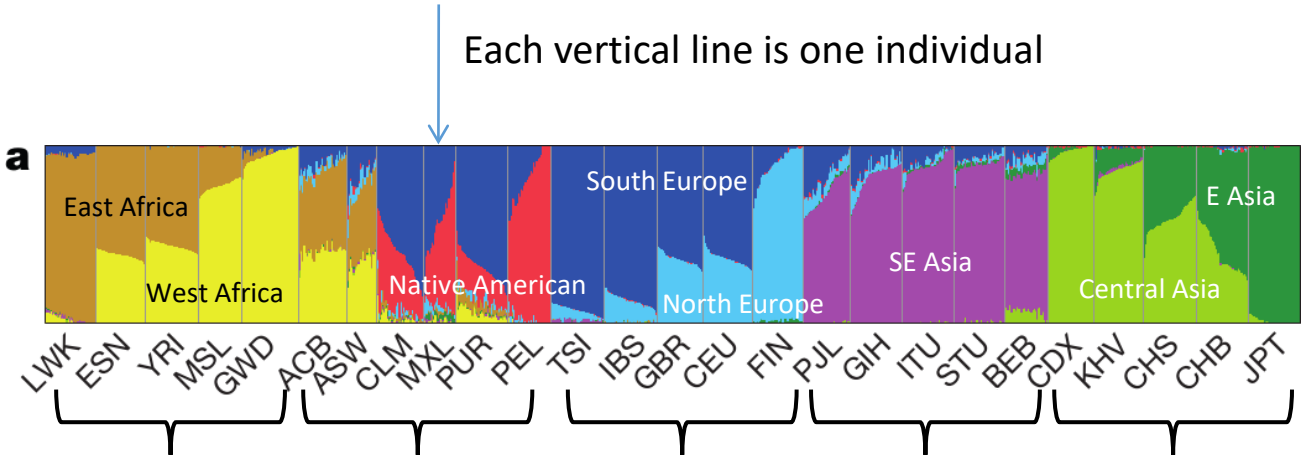
Population size, bottlenecks and expansion

- **Effective population size:** number of individuals needed in idealized model to recapitulate population properties
- Here, recapitulate the **coalescent time**: time to most recent common ancestor
- **Pairwise Markov sequential coalescent model** with population splits/growth enables comparison within vs. between populations
- 1KG suggests shared history beyond 150 kya
- Non-African population: Loss of heterozygosity, **bottleneck** 15-20 kya (migration out of Africa)
- After migration, rapid population expansion (with interesting exceptions: Finland, Peru, Mexico)
- Bottlenecks/founder effects: rare alleles suddenly rise in frequency due to small population size
- Selective sweeps: rare alleles suddenly rise in frequency due to positive selection
- Admixture between previously isolated populations



Thousand Genomes Consortium *Nature* 2016

Ancestry painting: population-level



Esan	ESN	
Gambian	GWD	
Luhya	LWK	
Mende	MSL	
Yoruba	YRI	
Barbadian	ACB	
African-American SW	ASW	
Colombian	CLM	
Mexican-American	MXL	
Peruvian	PEL	
Puerto Rican	PUR	
Dai Chinese	CDX	
Han Chinese	CHB	
Southern Han Chinese	CHS	
Japanese	JPT	
Kinh Vietnamese	KHV	
CEPH	CEU	
British	GBR	
Finnish	FIN	
Spanish	IBS	
Tuscan	TSI	
Bengali	BEB	
Gujarati	GIH	
Telugu	ITU	
Punjabi	PJL	
Tamil	STU	

African populations differentiate EW

American populations are an **admixture** of European, African, and native American populations

European populations differentiate NS

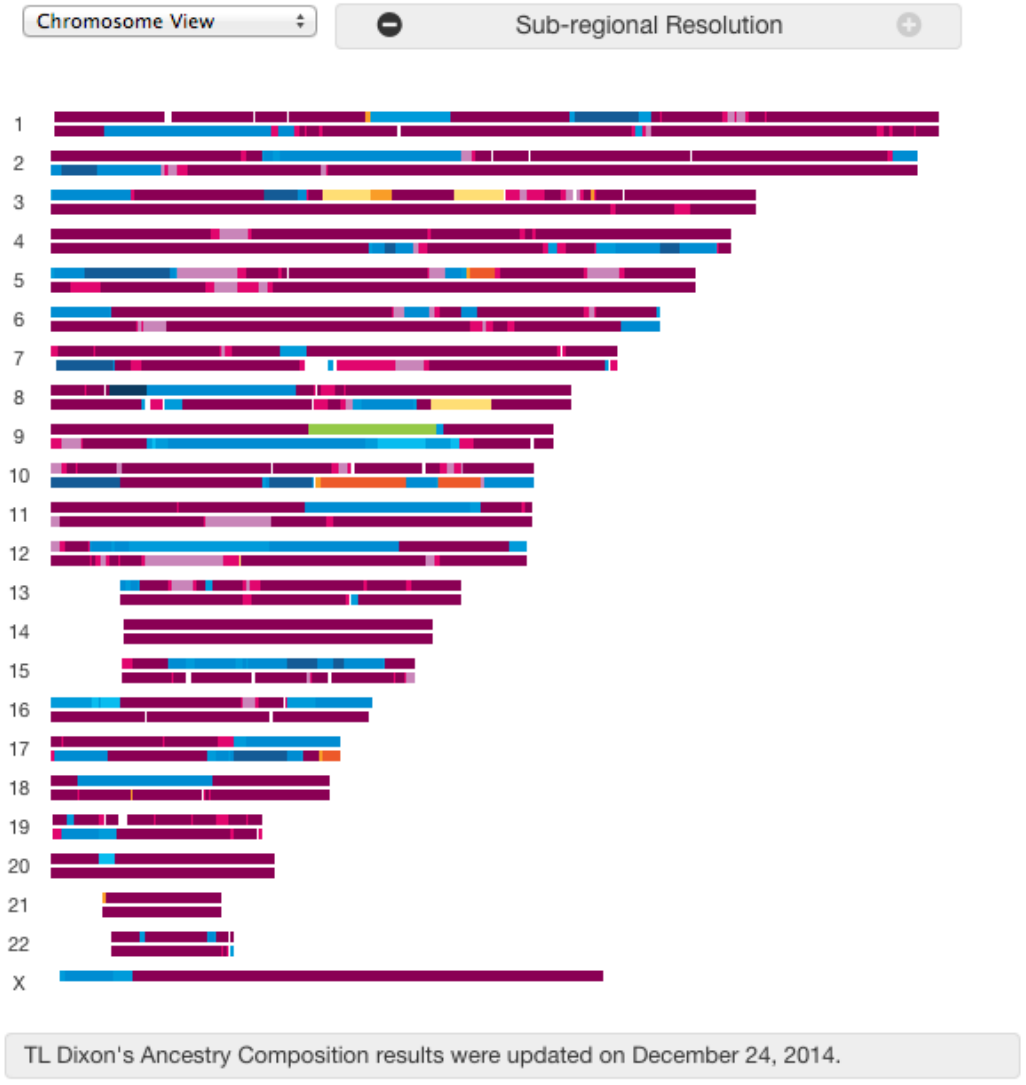
SE Asian populations differentiate NS

East Asian populations differentiate EW

- Goal: infer **ancestry** of segments of the genome, **population structure** (patterns of relatedness between ancestry groups)
- Sharing of genetic variants enables **ancestry painting** of individual genomes
- The history of migration, settlement, conquest is written on our genomes

Thousand Genomes Consortium *Nature* 2016

Ancestry painting (e.g. admixed individual)



Ancestry Composition tells you what percent of your DNA comes from each of 31 populations worldwide. This analysis includes DNA you received from all of your recent ancestors, on both sides of your family. The results reflect where your ancestors lived before the widespread migrations of the past few hundred years.

79.0%	Sub-Saharan African
72.3%	West African
2.9%	Central & South African
3.8%	Broadly Sub-Saharan African
18.4%	European
	Northern European
2.5%	British & Irish
0.2%	Scandinavian
11.4%	Broadly Northern European
0.6%	Ashkenazi
	Southern European
0.5%	Broadly Southern European
3.3%	Broadly European
1.9%	East Asian & Native American
0.8%	Native American
0.8%	Southeast Asian
0.2%	Broadly East Asian & Native American
0.7%	Unassigned
100%	**TL Dixon**

show all populations

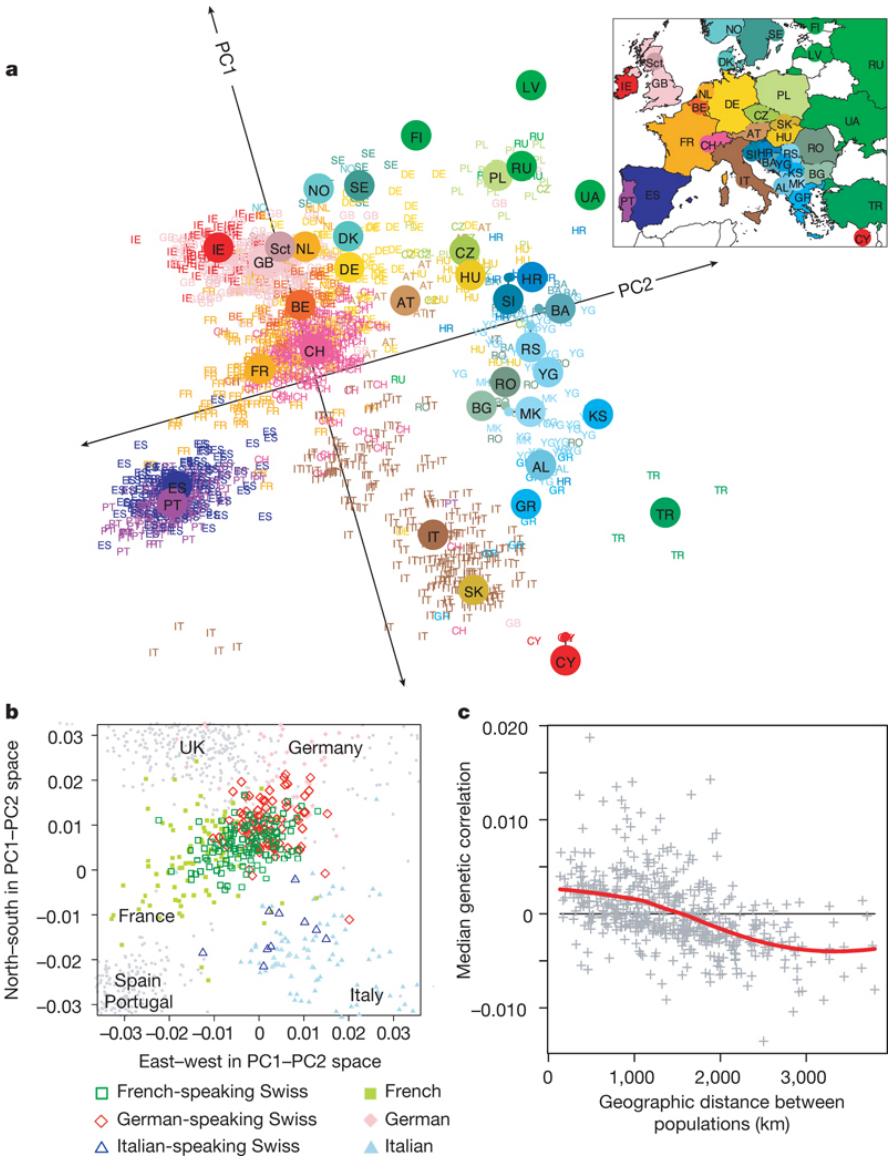
Which segments of a genome are shared with what populations

Genetic relatedness and geography

- Can we decompose genetic variation into the major forces shaping it?

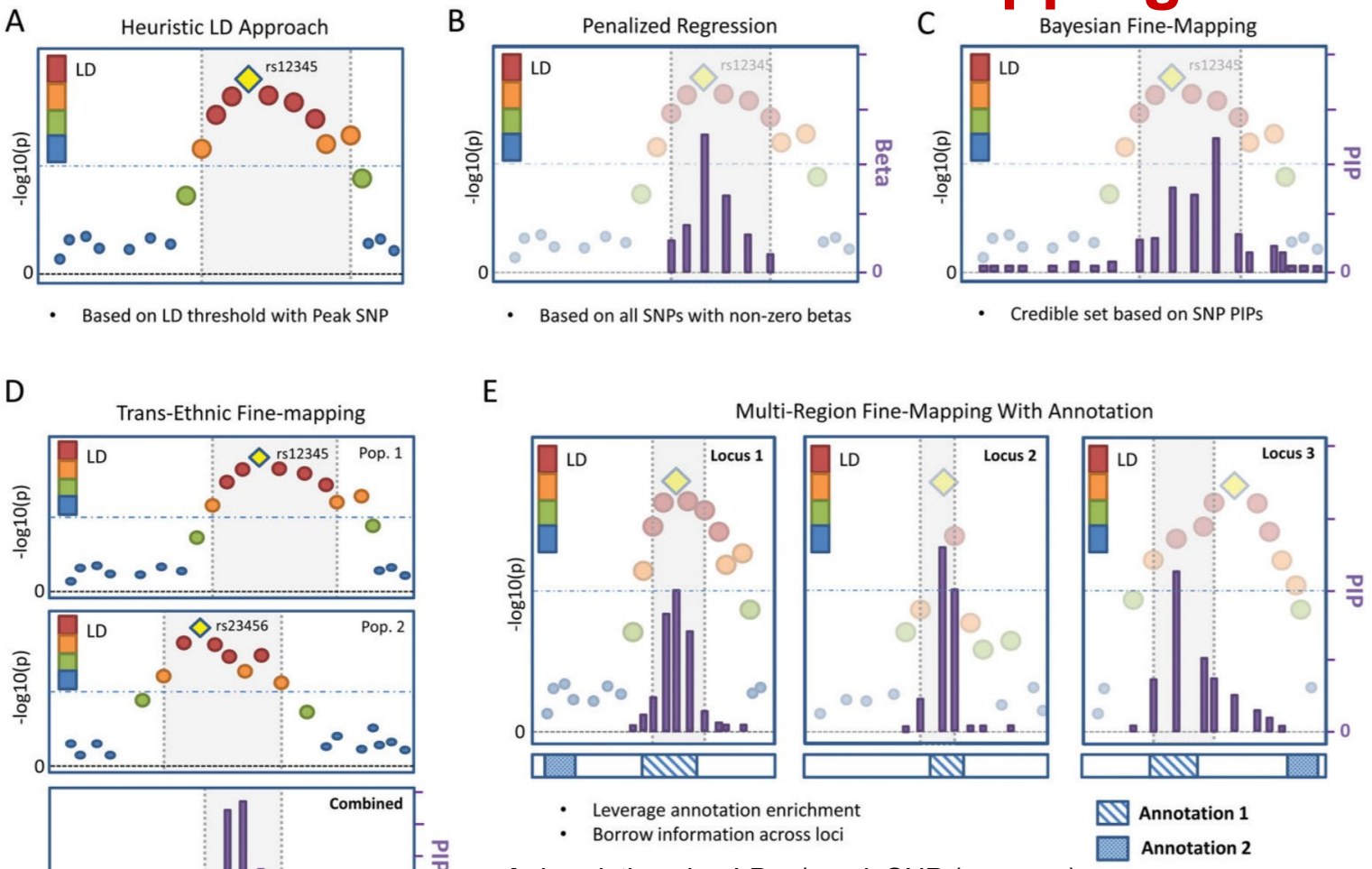
➔ PCA/SVD decomposition

- First components correspond to population structure.
- Population structure is shaped by geography! (people near each other are more likely to mate)
- In Europe, First two components correspond to N-S and E-W migration axes
- Country neighbors & borders visible at the genetic level



Novembre et al. *Nature* 2008

GWAS fine-mapping

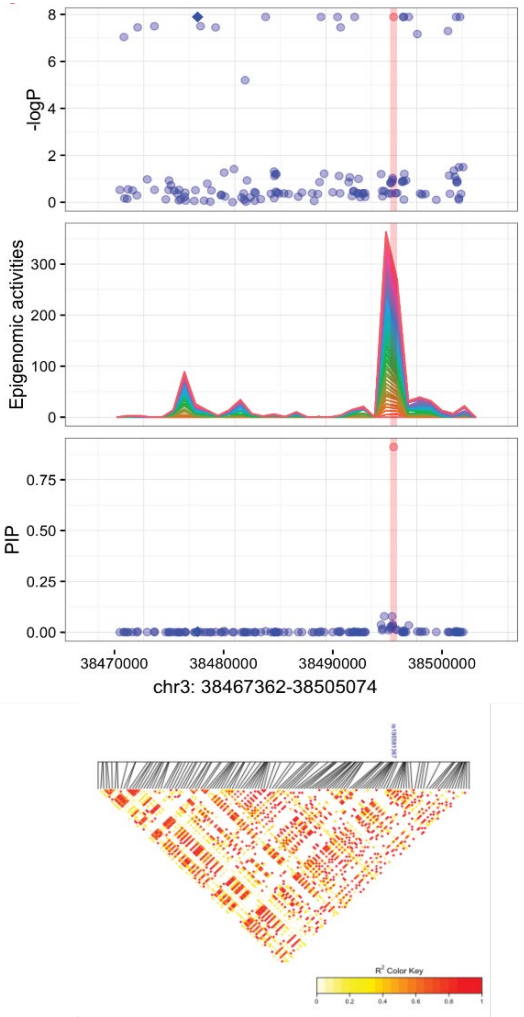


- LocusZoom of marginal SNP associations
- Y-axis: $-\log_{10}(p\text{-values})$
- X-axis: Variant positions
- Gold: peak SNP
- Other=degree LD w/peak SNP (red, orange, green, blue)
- Purple bars=additional variant-level statistics by fine-mapping
- (Penalized regression=Beta; Bayesian: posterior inclusion probabilities (PIPs))
- Light grey=regions selected by fine-mapping

- **A**=heuristic using LD w/ peak SNP (>orange)
- **B**=Penalized regression=Beta not shrunk to zero
- **C**=Bayesian PIPs summed to credible sets using $P_{coverage}$>95%
  (note: peak SNP not always highest PIP ← correlation structure of SNPs in region)
- **D**=2 pops w/ different local LD struct → meta-analysis narrow fine-mapping credible region
- **E**=Anno1 overlap in locus 1 & 2 → predict top-PIP SNP in locus 3 (overlaps anno1)

Fine-mapping disease associations: (1) Epigenomics / functional data (next lecture)

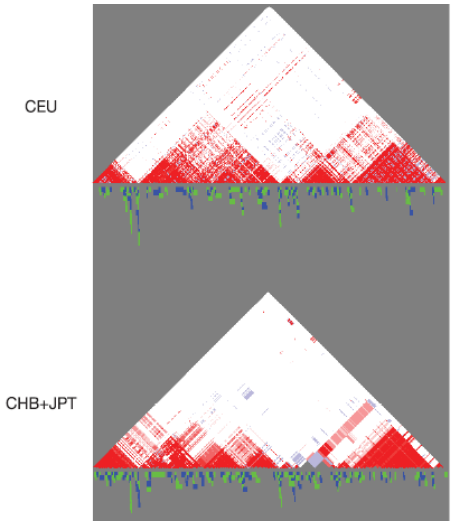
- **Association mapping** refers to identifying variants/gene associated with disease
- This is confounded by LD
- Many variants are strongly correlated to the true causal variant, and will show nearly as strong associations
- Use estimated correlations to explain correlated associations and recover the true underlying effects



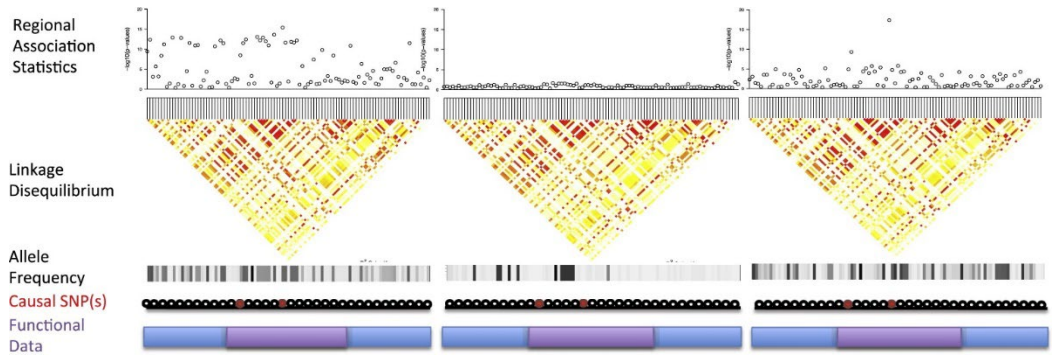
Li and Kellis *BiorXiv* 2016

Fine-mapping disease associations (2) Multi-ethnic analysis

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

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS
2. Polygenic scores (PGS)
3. From GWAS to biological insights
4. From Region to Mechanism / Circuitry

Potential of PRS in clinical practice

AHA SCIENTIFIC STATEMENT

Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association

Jack W. O'Sullivan, MBBS, DPhil, Chair; Sridharan Raghavan, MD, PhD; Carla Marquez-Luna, PhD; Jasmine A. Luzum, PharmD, PhD; Scott M. Damrauer, MD, FAHA; Euan A. Ashley, MBChB, DPhil, FAHA; Christopher J. O'Donnell, MD, MPH; Cristen J. Willer, DPhil; Pradeep Natarajan, MD, MMSc, Vice Chair; on behalf of the American Heart Association Council on Genomic and Precision Medicine; Council on Clinical Cardiology; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Radiology and Intervention; Council on Lifestyle and Cardiometabolic Health; and Council on Peripheral Vascular Disease

"*These observations point to the **possibility of using genetic profiling to inform clinical practice** in significantly larger groups of individuals than for whom monogenic cardiovascular variants are considered. As a result of exponential increases in the proportion of individuals with broad genetic profiling, **cardiovascular PRSs are beginning to enter clinical practice**. Such PRSs may be appropriately considered in select scenarios, given the current evidence base.*"

O'Sullivan, et al. *Circulation* (2022);

Potential relevance of PRS in clinical practice

Example: coronary artery disease

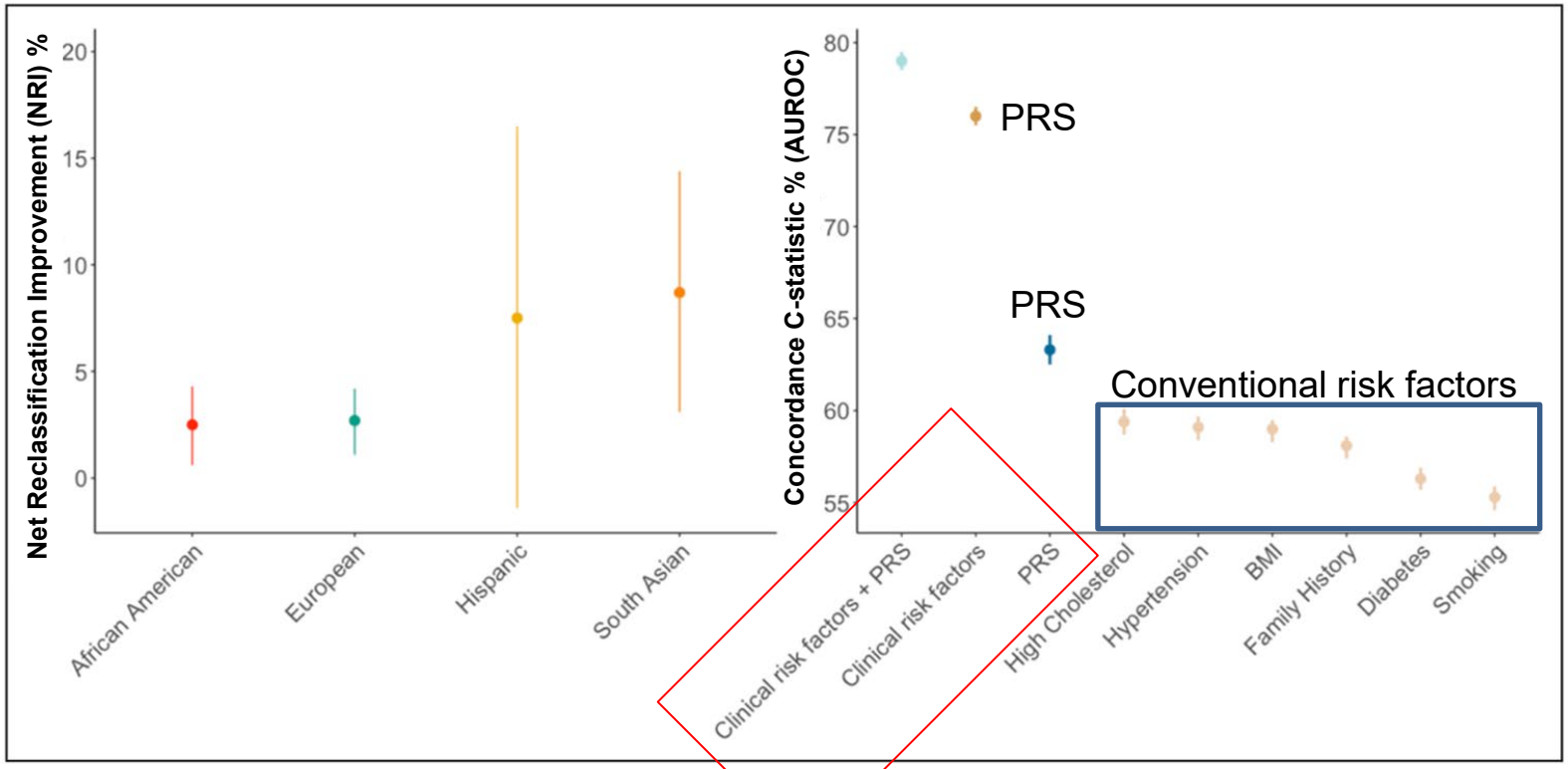


Figure 3. **Predictive ability of polygenic risk scores for coronary artery disease.**

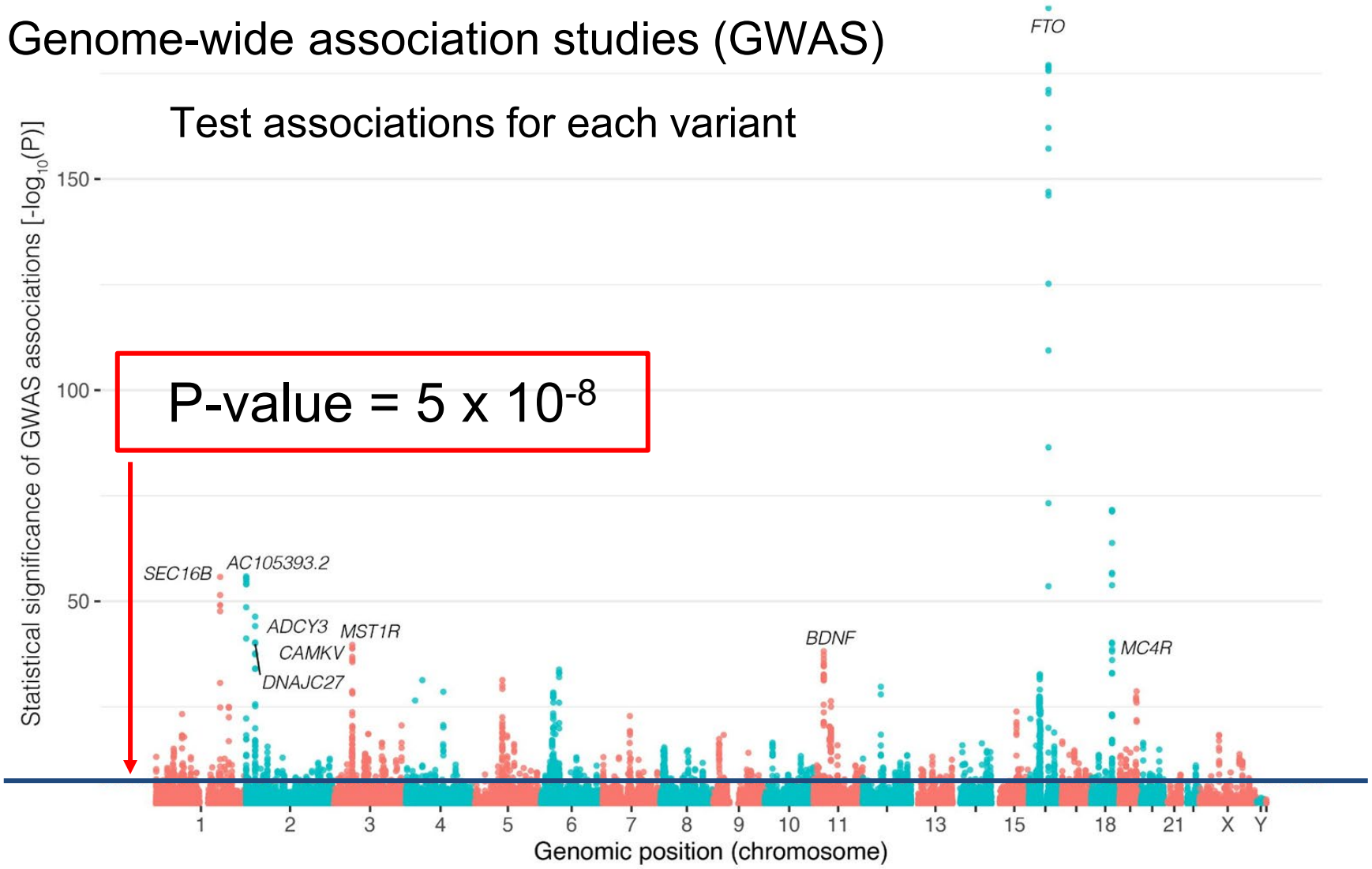
- PRS has higher risk stratification ability than conventional risk factors
- PRS & conventional risk factors leads to improvement

O'Sullivan, et al. *Circulation* (2022);

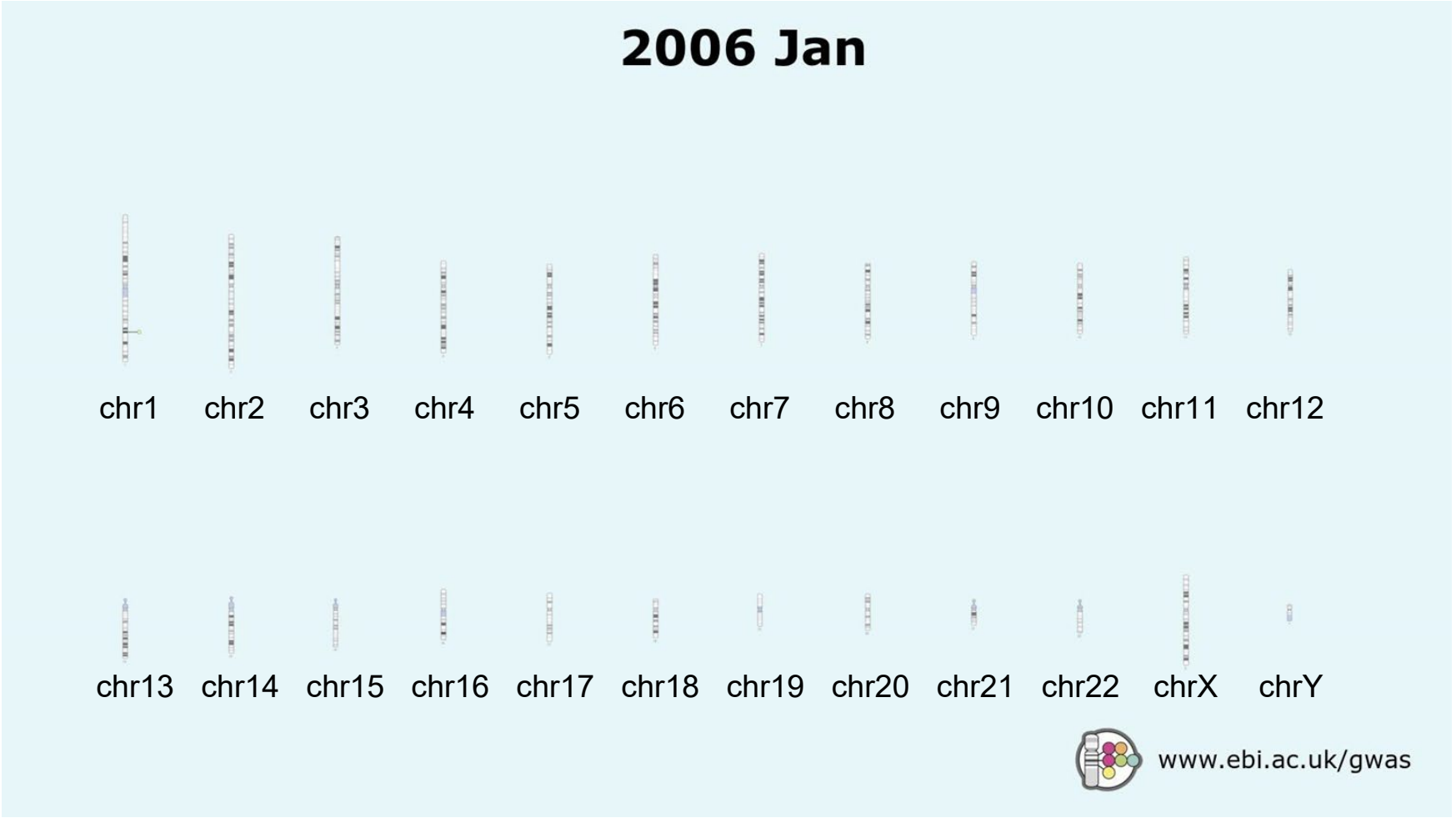
GWAS reveals complex traits are polygenic

Genome-wide association studies (GWAS)

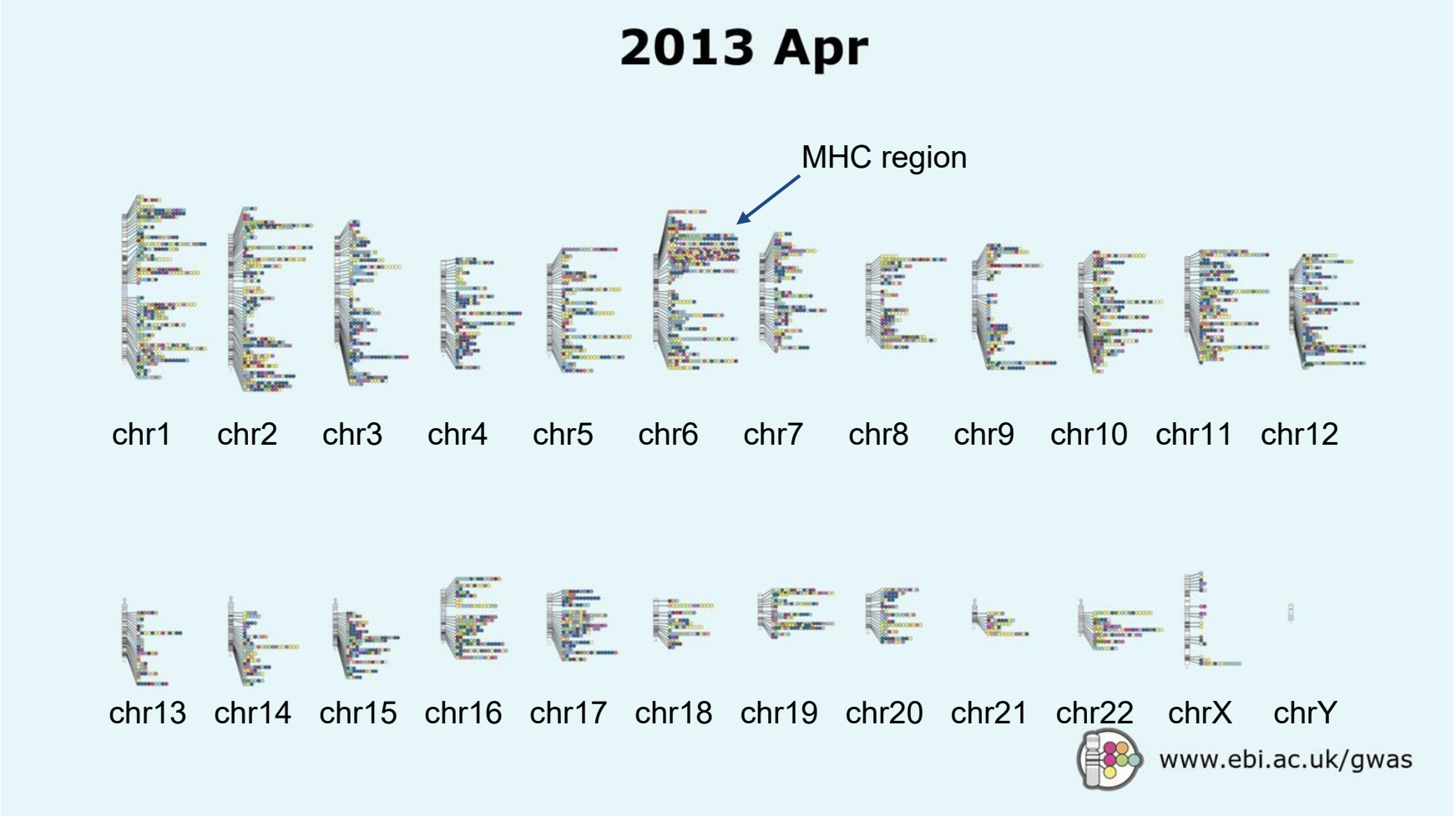
Test associations for each variant



Mapping disease-associated variants with GWAS

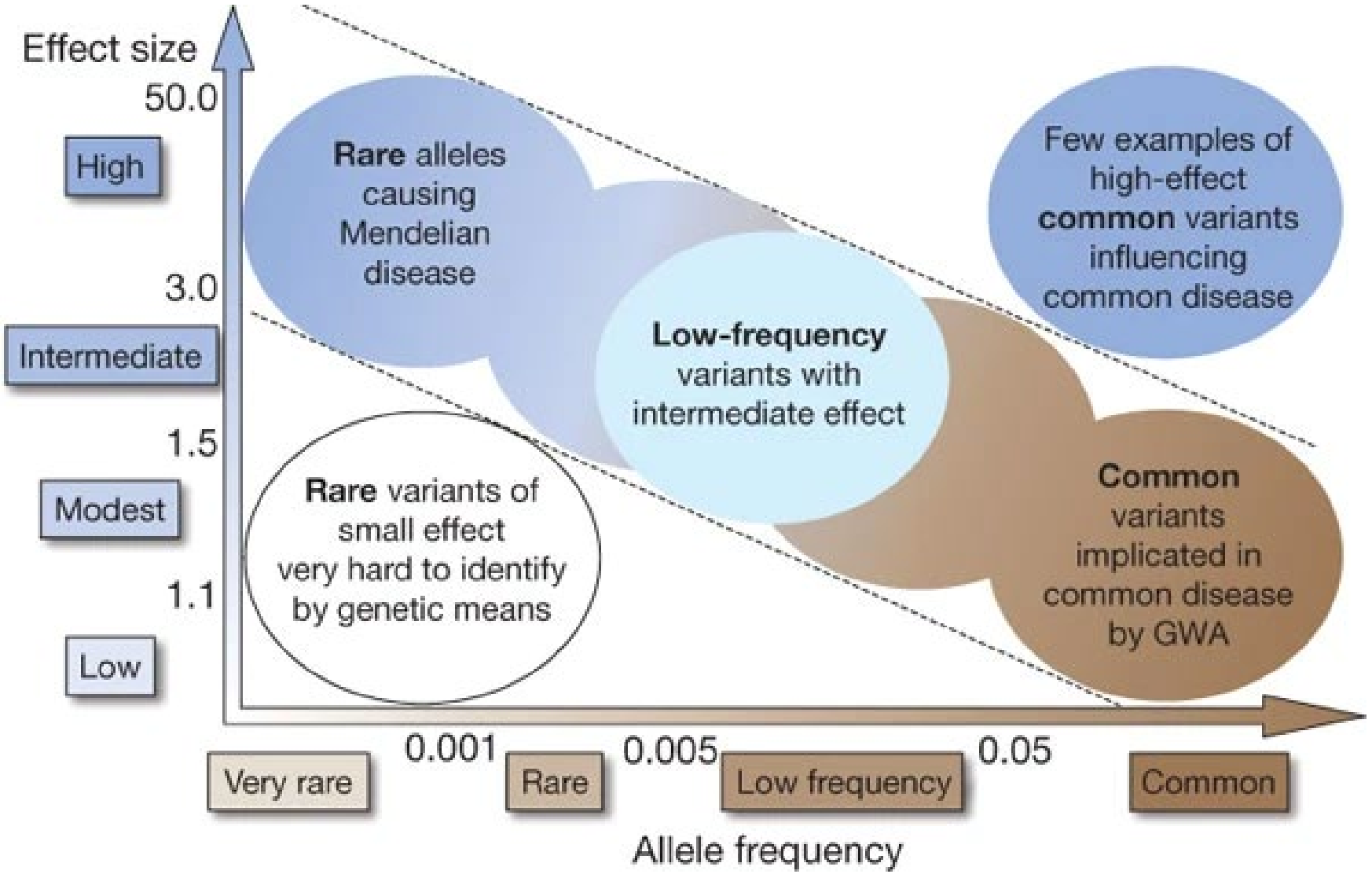


Mapping disease-associated variants with GWAS



Mapping disease-associated variants with GWAS

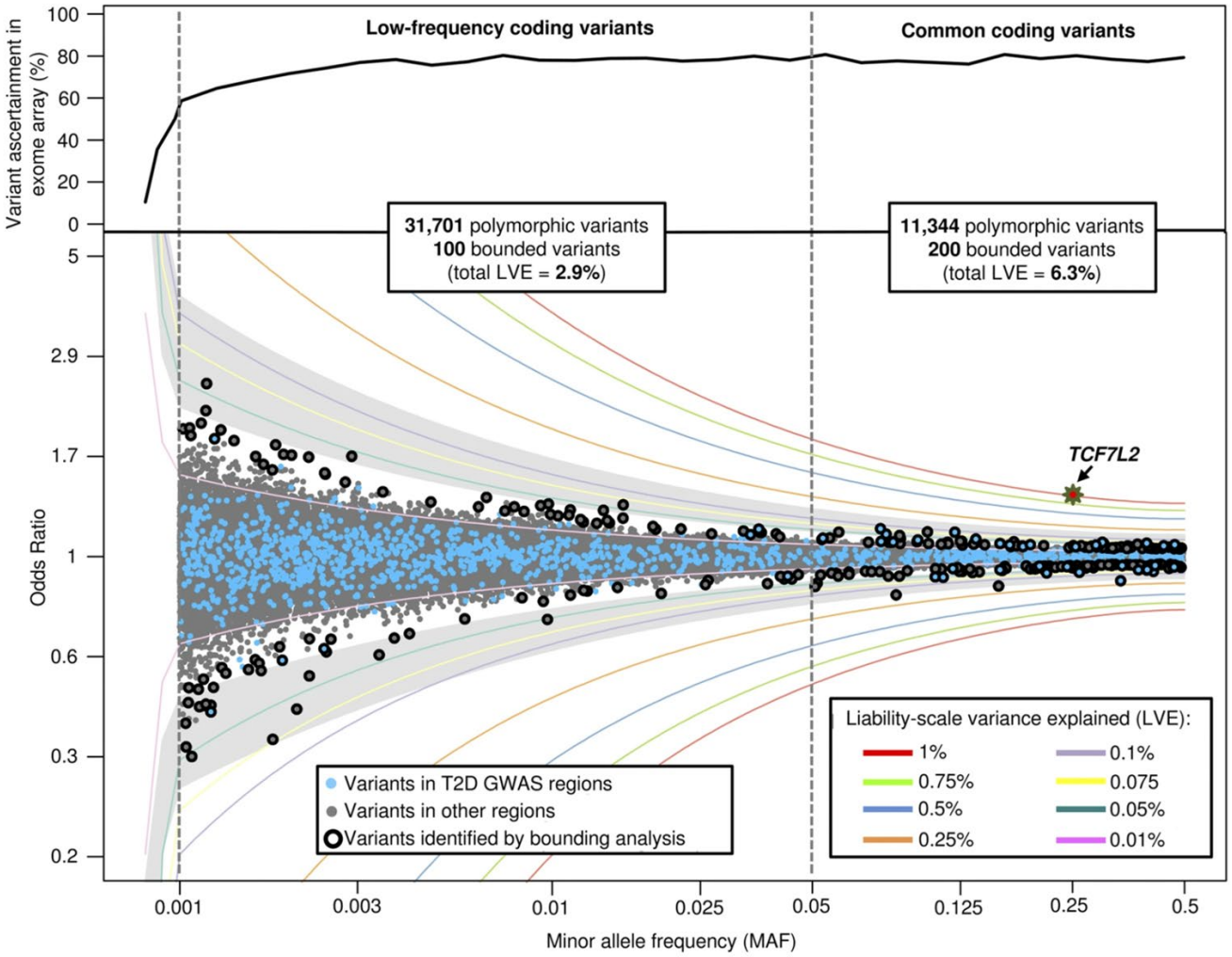
Most common variants have small effects



Manolio et al 2009, PMID: 19812666

Most common variants have small effects

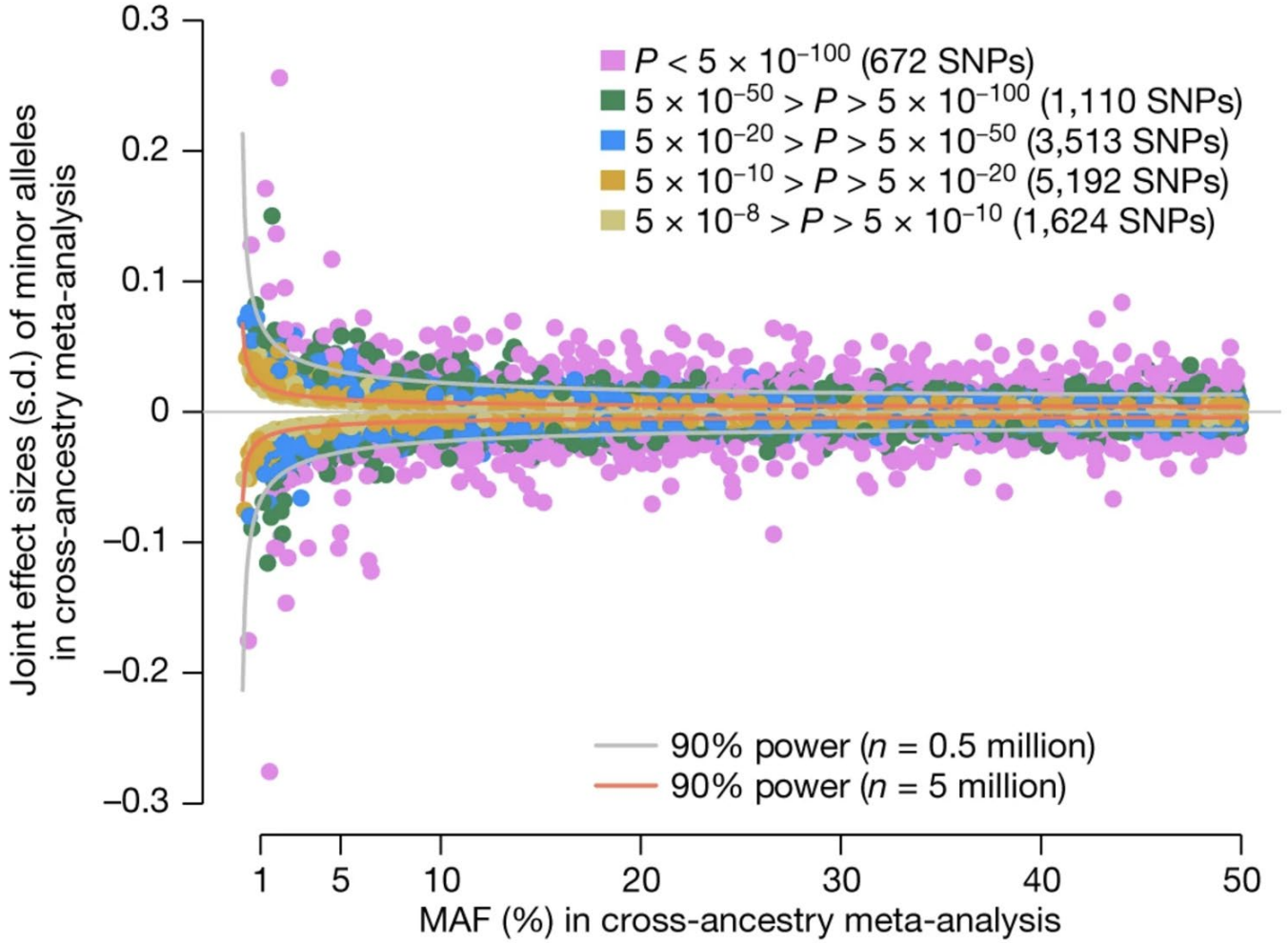
Type 2 diabetes



Fuchsberger, et al. 2016

Most common variants have small effects

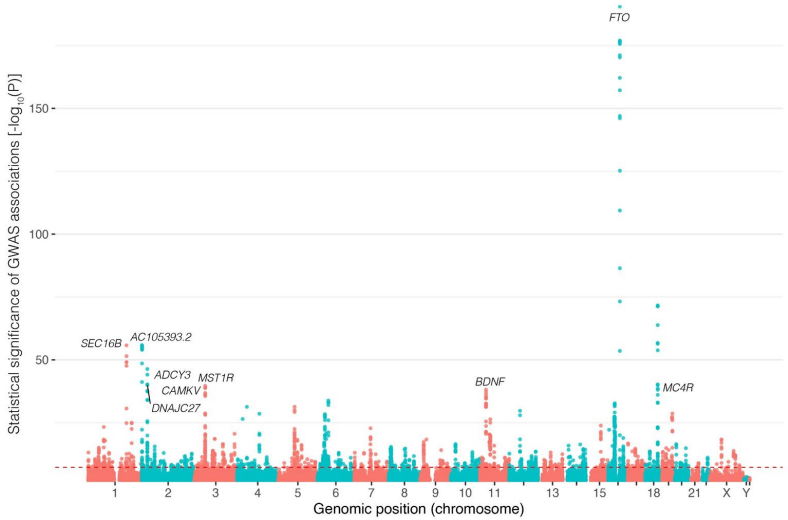
Standing height (n = 5 million, 2022)



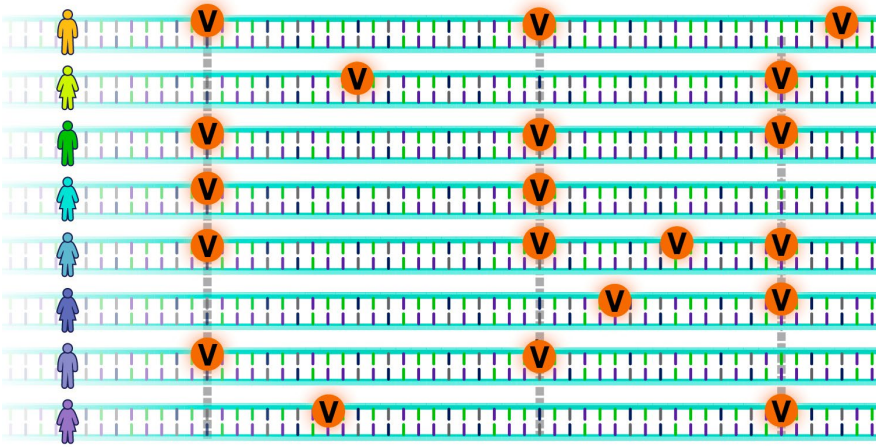
Yengo*, Vedantam*, Marouli*, et al. 2022

Estimating individual-level liability of complex traits

Population-level inference vs. individual-level inference



Population-level inference (GWAS)



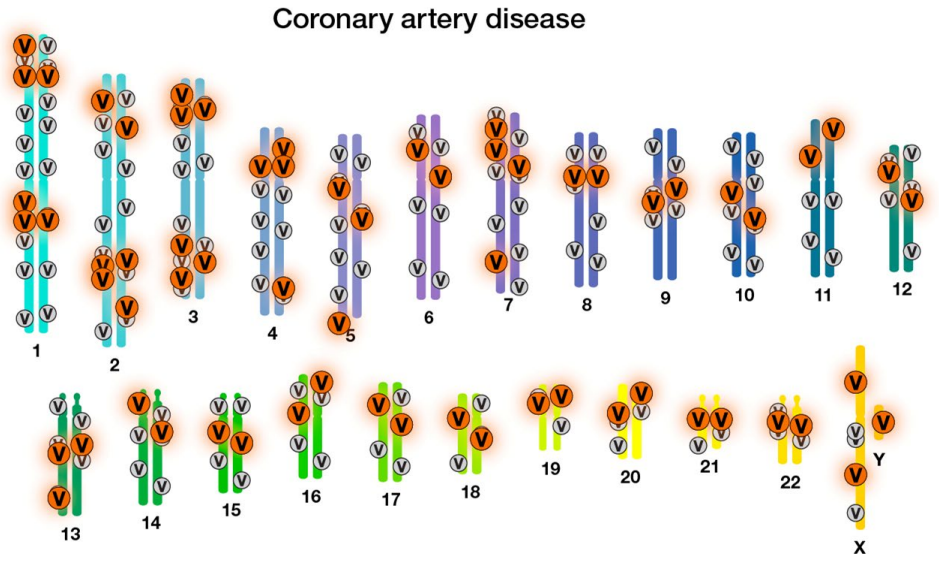
Individual-level inference (???)

How do we inform population-level insights into individuals?

https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores

Challenges in polygenic complex traits

- Monogenic traits (e.g. cystic fibrosis)
  - "Carrier" or "non-carrier"
  - *CFTR* (cystic fibrosis transmembrane conductance regulator)
  - high penetrance, high effect size, often coding variants
- Polygenic complex traits (e.g. coronary artery disease, height, etc.)
  - Different individuals have a different subset of "risk" alleles
  - Lower penetrance, lower effect size, many non-coding variants



https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores

Polygenic scores combine effects of disease-associated alleles for each individual

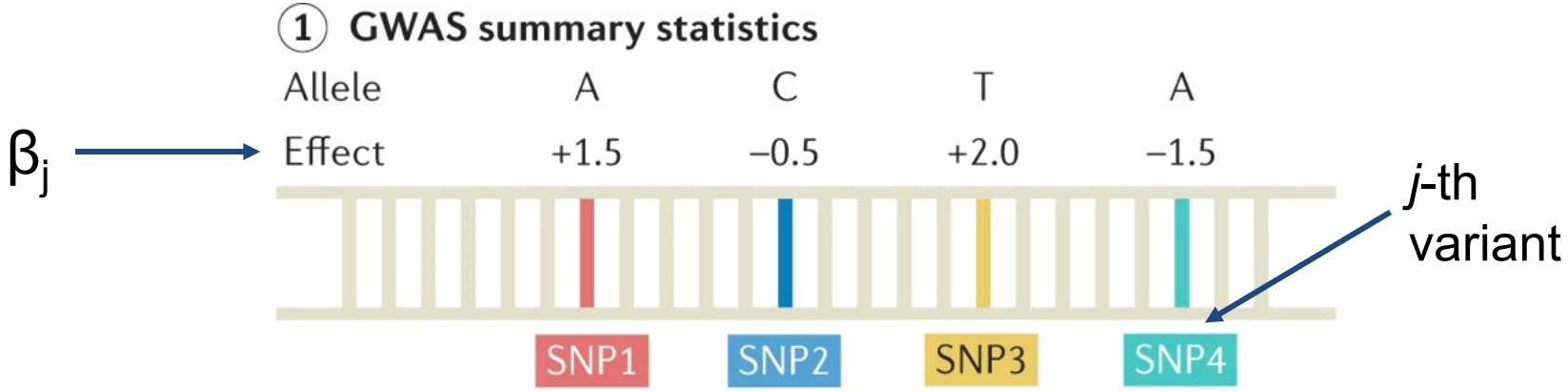
- Polygenic scores (PGS)
  - aka. Genetic risk score (GRS), Polygenic risk score (PRS), etc.
    - "risk" → disease risks
    - "Polygenic" → statement of the genetic architecture of a trait

- Polygenic score := weighted sum of disease-associated alleles

$$\mathrm{PRS}_i = \sum_{j \in J} \beta_j \, G_{ij}$$

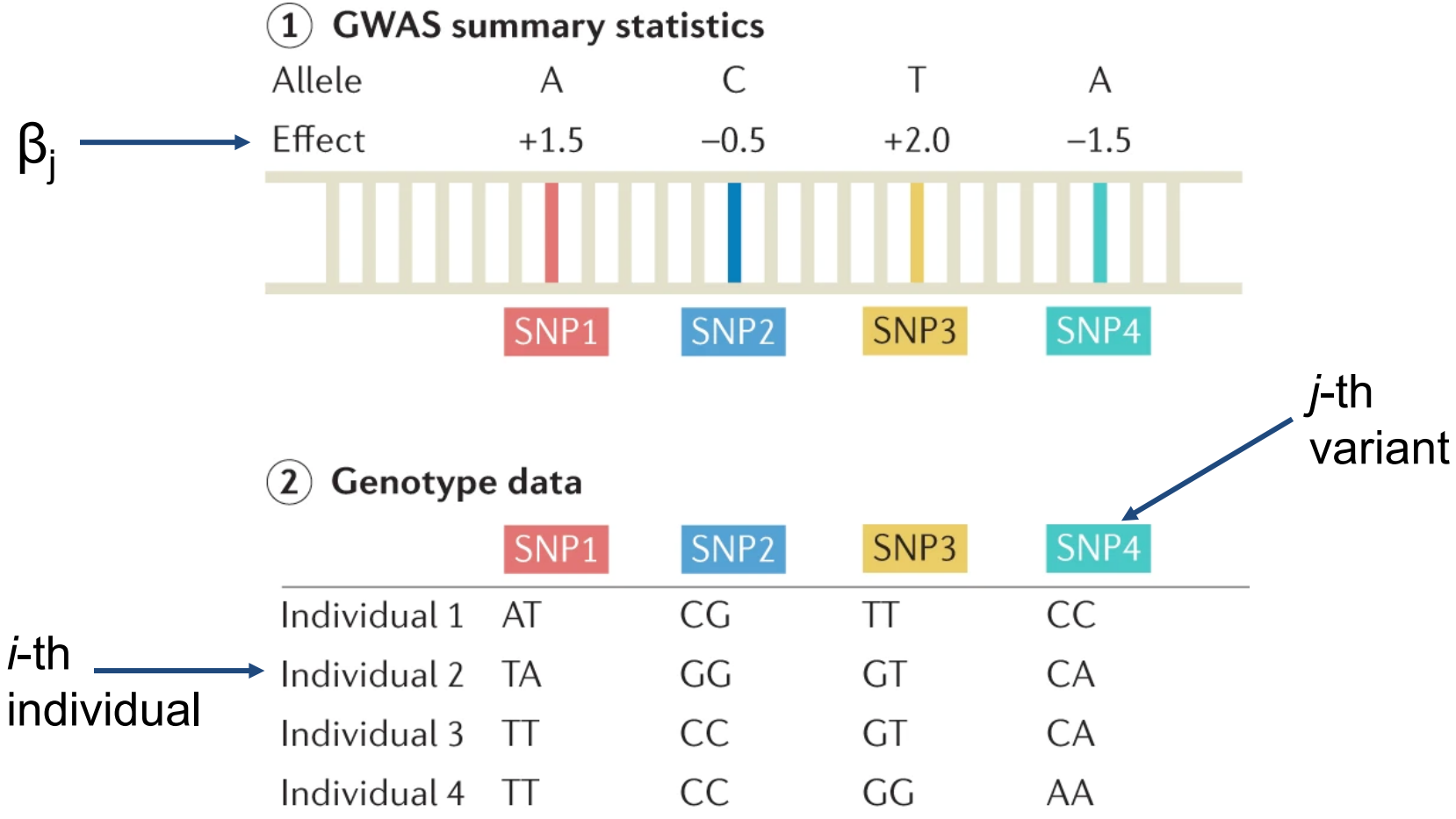
i-th individual *G*: genotype

j-th variant β: effect size



Polygenic scores combine effects of disease-associated alleles for each individual

- Polygenic score $\mathrm{PRS}_i = \sum_{j \in J} \beta_j \, G_{ij}$ *i*-th individual *j*-th variant *G*: genotype *β*: effect size



Polygenic scores combine effects of disease-associated alleles for each individual

1 GWAS summary statistics

Allele	A	C	T	A
Effect	+1.5	−0.5	+2.0	−1.5

β_j

2 Genotype data

	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA

$$\mathrm{PRS}_i = \sum_{j \in J} \beta_j \, G_{ij}$$

i-th individual genotype

j-th variant

G:

β: effect size

3 Polygenic risk score

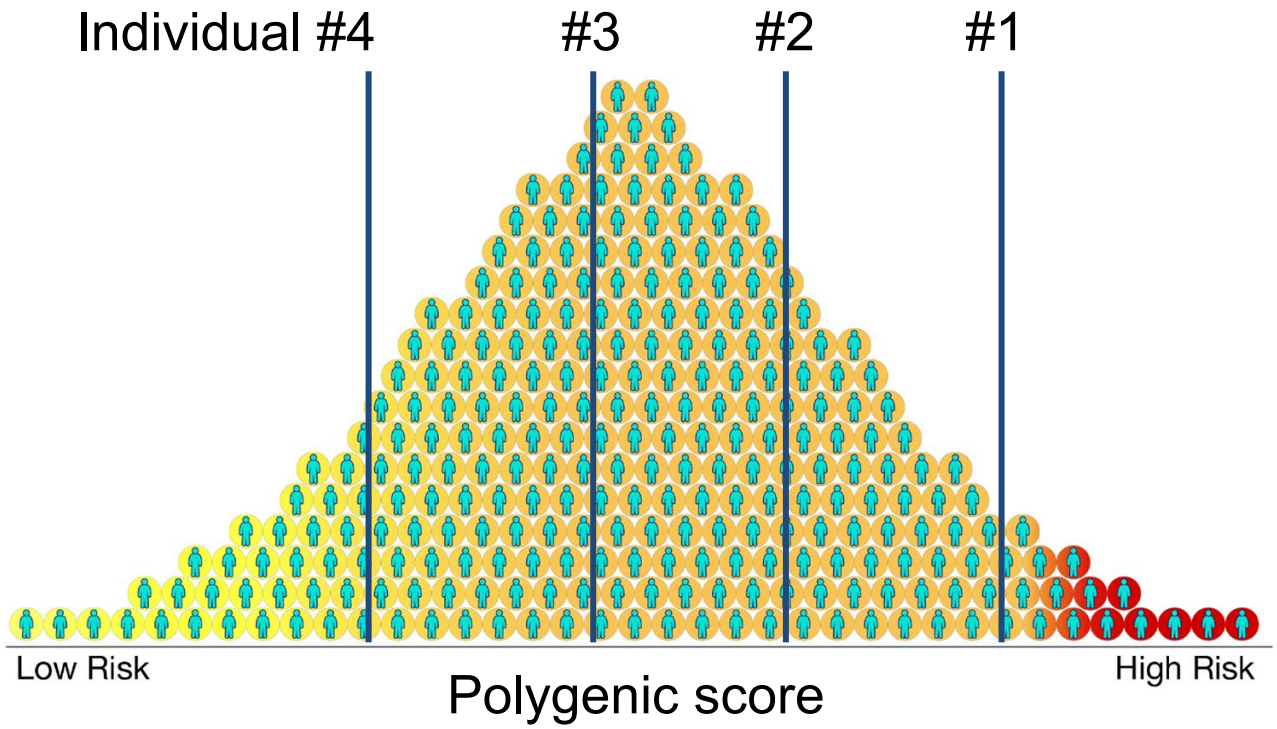
Individual 1	1.5	−	0.5	+	4.0	−	0.0	= **5.0**
Individual 2	1.5	−	0.0	+	2.0	−	1.5	= **2.0**
Individual 3	0.0	−	1.0	+	2.0	−	1.5	= **−0.5**
Individual 4	0.0	−	1.0	+	0.0	−	3.0	= **−4.0**

Uffelmann et al., *Nat Rev Methods Primers* (2021)

Polygenic scores combine effects of disease-associated alleles for each individual

③ Polygenic risk score

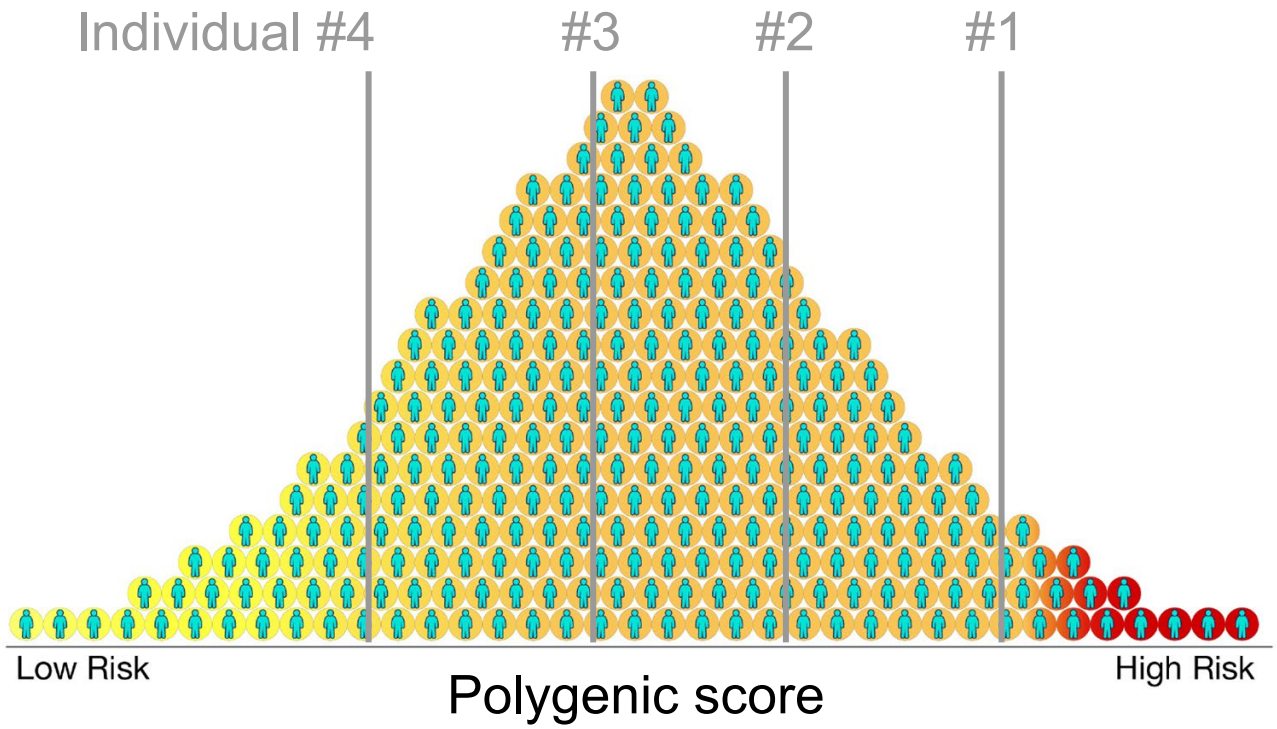
Individual 1	1.5	−	0.5	+	4.0	−	0.0	= **5.0**
Individual 2	1.5	−	0.0	+	2.0	−	1.5	= **2.0**
Individual 3	0.0	−	1.0	+	2.0	−	1.5	= **−0.5**
Individual 4	0.0	−	1.0	+	0.0	−	3.0	= **−4.0**



Uffelmann et al., *Nat Rev Methods Primers* (2021)
https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores

Polygenic scores estimate the relative genetic liability of disease

- **Genetic** liability of the disease – complex traits are influenced by genetics, environmental factors, and their interactions
- "**Relative**" – baseline risk factors (age, biological sex, comorbidity, …) are not part of the picture
- "**Estimate**" – sample size & statistical power, model misspecification



Uffelmann et al., *Nat Rev Methods Primers* (2021)
https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores

Potential of PRS in clinical practice

AHA SCIENTIFIC STATEMENT

Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association

Jack W. O'Sullivan, MBBS, DPhil, Chair; Sridharan Raghavan, MD, PhD; Carla Marquez-Luna, PhD; Jasmine A. Luzum, PharmD, PhD; Scott M. Damrauer, MD, FAHA; Euan A. Ashley, MBChB, DPhil, FAHA; Christopher J. O'Donnell, MD, MPH; Cristen J. Willer, DPhil; Pradeep Natarajan, MD, MMSc, Vice Chair; on behalf of the American Heart Association Council on Genomic and Precision Medicine; Council on Clinical Cardiology; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Radiology and Intervention; Council on Lifestyle and Cardiometabolic Health; and Council on Peripheral Vascular Disease

"*These observations point to the **possibility of using genetic profiling to inform clinical practice** in significantly larger groups of individuals than for whom monogenic cardiovascular variants are considered. As a result of exponential increases in the proportion of individuals with broad genetic profiling, **cardiovascular PRSs are beginning to enter clinical practice**. Such PRSs may be appropriately considered in select scenarios, given the current evidence base.*"

O'Sullivan, et al. *Circulation* (2022);

Potential relevance of PRS in clinical practice

Example: coronary artery disease

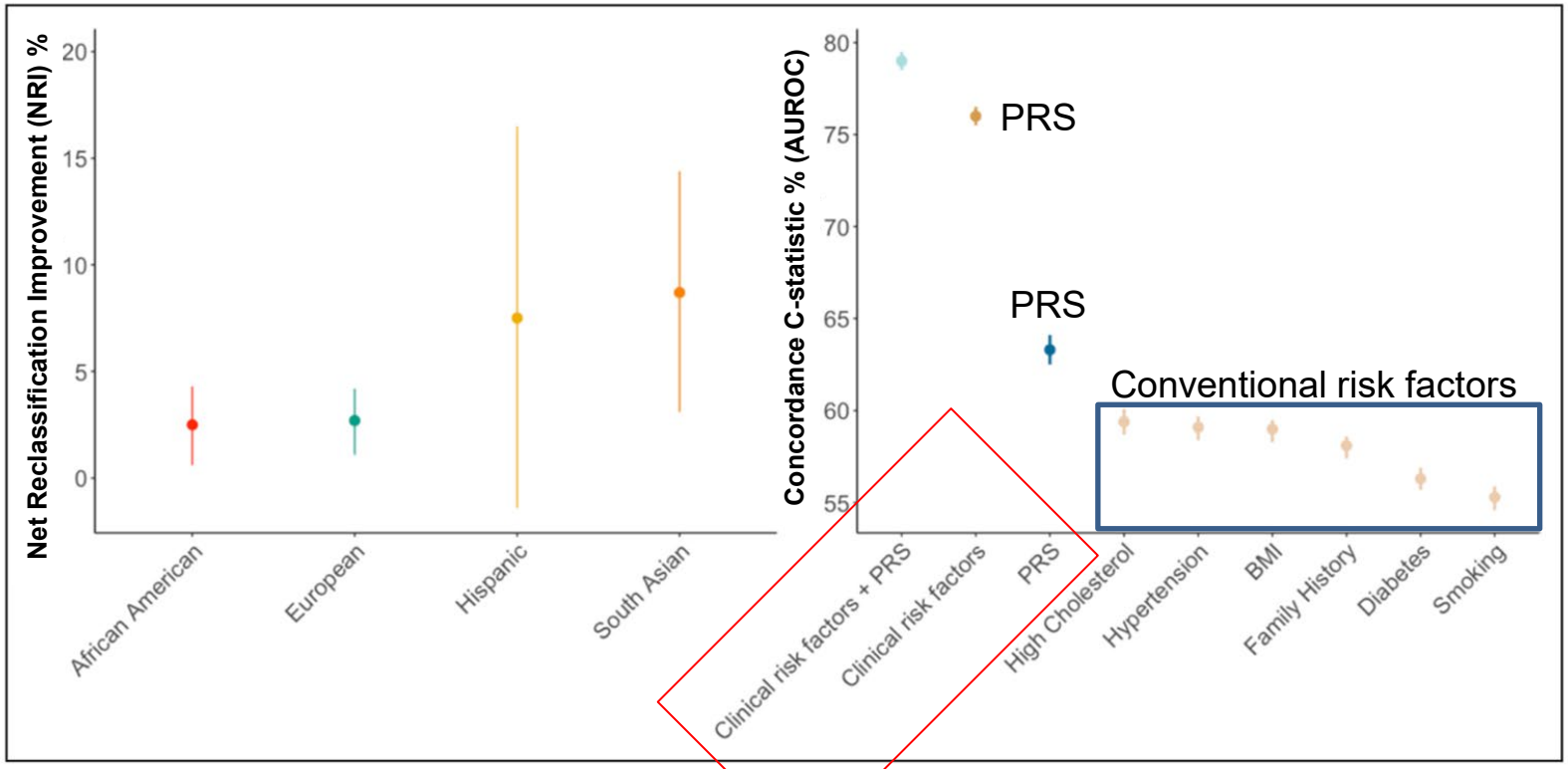


Figure 3. Predictive ability of polygenic risk scores for coronary artery disease.

- PRS has higher risk stratification ability than conventional risk factors
- PRS & conventional risk factors leads to improvement

O'Sullivan, et al. *Circulation* (2022);

Potential clinical utility of PRS for cardiovascular disease

Disease/risk factor	Potential clinical utility of PRS
CAD	Earlier identification for lifestyle therapies and statins, potentially for those with very high CAD PRSs Earlier screening for subclinical atherosclerosis to time the initiation of pharmacotherapies Use as a risk-enhancing factor for primary prevention in middle-aged patients at borderline-intermediate 10-y ASCVD risk
AF	Earlier AF detection and resultant prophylactic anticoagulation, potentially with monitoring devices Rigorous control of additive clinical risk factors for AF
T2D	Earlier lifestyle modification Potential consideration of prophylactic hypoglycemic medications with concomitant additional T2D clinical risk factors Genomic stratification may optimize hypoglycemic choice
VTE	Rigorous VTE risk-reducing strategies in the context of high-risk scenarios (prolonged travel, major surgery, etc)
Hypercholesterolemia	Earlier institution and earlier uptitration of lipid-lowering pharmacotherapies analogous to FH
Pharmacogenomics	Personalized drug therapy regimens that increase drug efficacy and decrease toxicities, eg, personalized β-blocker target dose in patients with HFrEF or the prevention of drug-induced QT prolongation

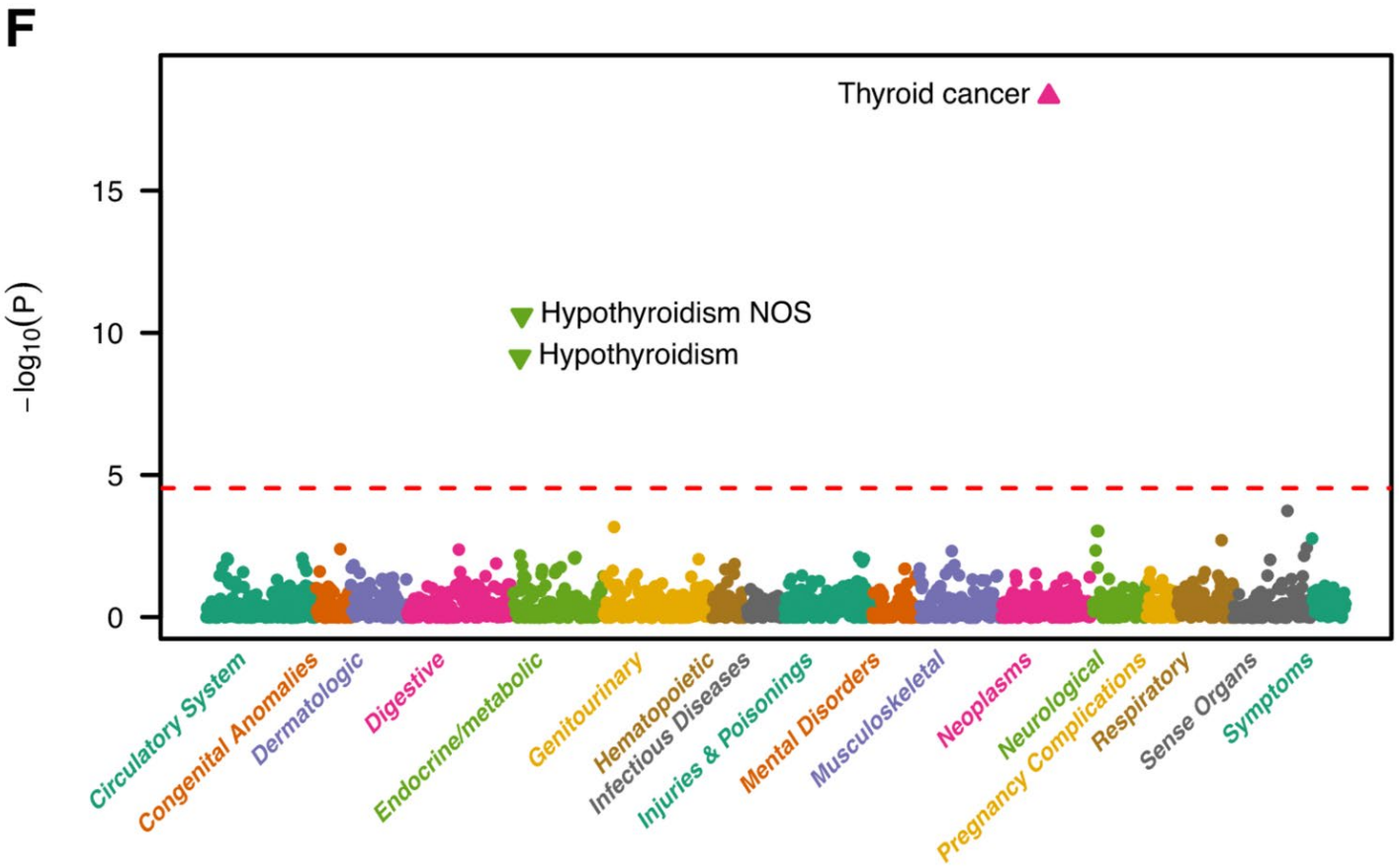
AF indicates atrial fibrillation; ASCVD, atherosclerotic cardiovascular disease; CAD, coronary artery disease; FH, familial hypercholesterolemia; HFrEF, heart failure with reduced ejection fraction; PRS, polygenic risk score; T2D, type 2 diabetes; and VTE, venous thromboembolism. Lone AF refers to AF in the absence of other cardiovascular risk factors (typically in young adults).

- Early-stage identification/intervention, Risk stratification, …

O'Sullivan, et al. *Circulation* (2022);

PGS is a useful tool for research

Cancer PRS model shows pleiotropic association with non-cancer traits



Evaluate the observed phenotypic enrichments of all patients with high cancer PRS

1. Start with PRS score
2. Rank patients
3. Find phenotypic enrichments for those patients
4. Method: ROC
   - x-axis: Cancer PRS score
   - y-axis: %people with trait
5. Take significance, plot it on this graph here

PRS-PheWAS analysis, assessing genetic correlation between traits

Fritsche, et al., *AJHG* (2018)

PGS is a useful tool for research

LETTERS
https://doi.org/10.1038/s41591-020-0785-8
nature medicine
Check for updates

Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan

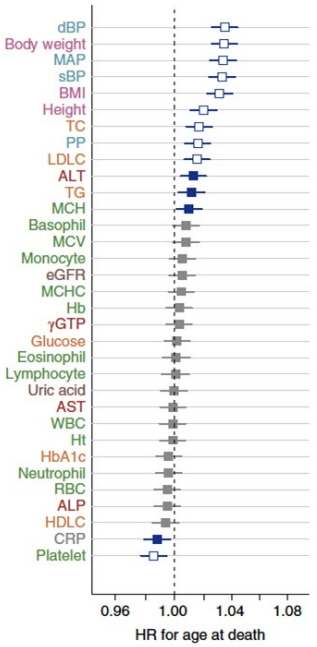
Saori Sakaue[1,2,3,18], Masahiro Kanai[1,2,4,5,6,7,8,18], Juha Karjalainen[4,5,6,8], Masato Akiyama[1,9], Mitja Kurki[4,5,6,8], Nana Matoba[1], Atsushi Takahashi[1,10], Makoto Hirata[11], Michiaki Kubo[12], Koichi Matsuda[13], Yoshinori Murakami[14], FinnGen, Mark J. Daly[4,5,6,8], Yoichiro Kamatani[1,15] and Yukinori Okada[2,16,17] ✉

Association of PRS with lifespan

Lifespan ~ PRS

$\Sigma_{i=1}^{m} \hat{b}x_i$

Genotype



PGS(biomarker) associations with lifespan (age at death)

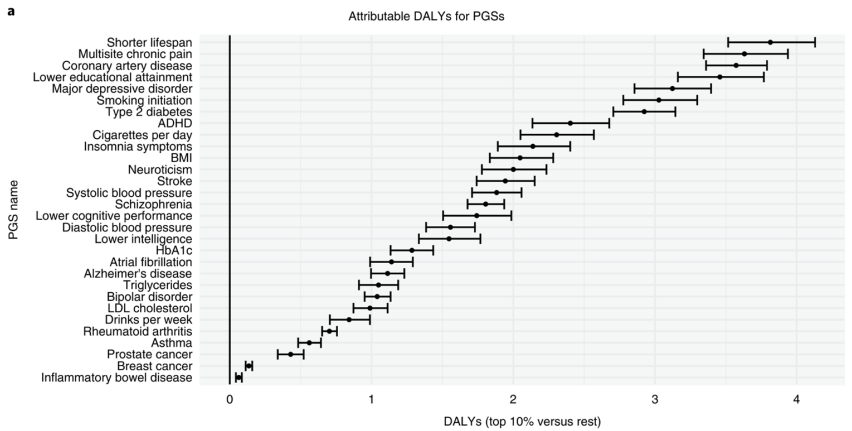
Death might affect phenotypes measured, but PRS of those phenotypes can correlate with age at death more 'cleanly'

nature medicine
ARTICLES
https://doi.org/10.1038/s41591-022-01957-2
Check for updates

OPEN

Genetic risk factors have a substantial impact on healthy life years

Sakari Jukarainen[1] ✉, Tuomo Kiiskinen[1,2,3], Sara Kuitunen[1,2], Aki S. Havulinna[1,2], Juha Karjalainen[1,3,4], Mattia Cordioli[1], Joel T. Rämö[1], Nina Mars[1], FinnGen[*,**], Kaitlin E. Samocha[3,5], Hanna M. Ollila[1,3,5,6], Matti Pirinen[1,7,8] and Andrea Ganna[1,3,4] ✉



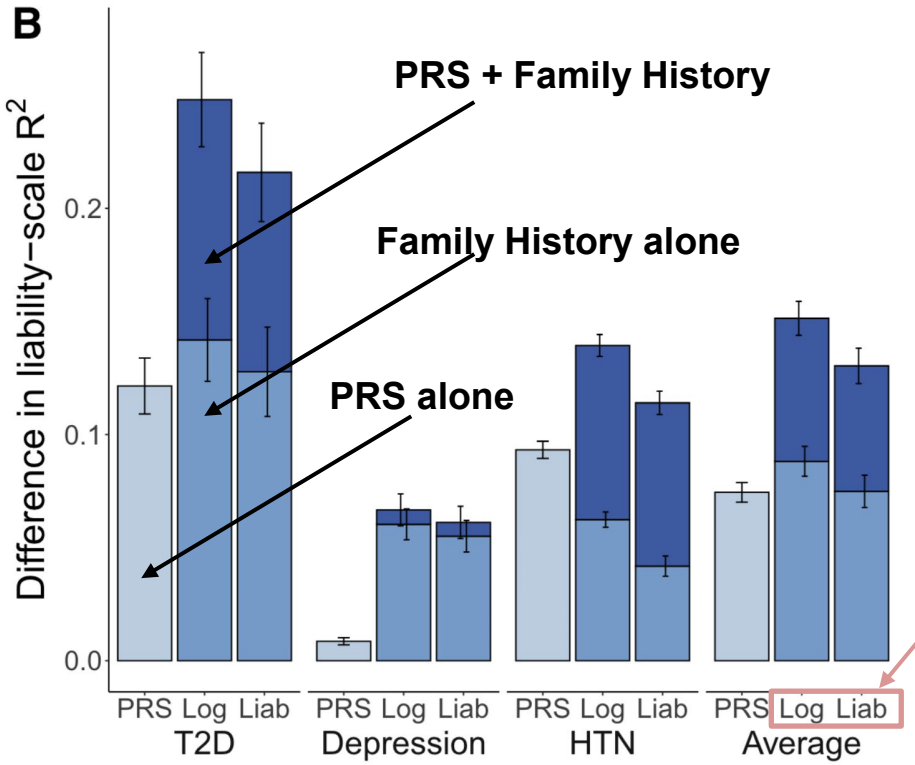
PGS associations with disability adjusted life years (DALY)

Sakaue*, Kanai*, et al., *Nat Med* (2020).; Jukarainen, et al., *Nat Med* (2022)

Family history (FH) complements PGS

Risk factors not captured in PGS

- Rare variants with large effects
  - Sample size & statistical power limitation in PGS
- Environmental factors



Two measures of liability:
- Log: logistic regression model
- Liab: liability threshold model

See Hujoel, et al. Cell Genom 2022

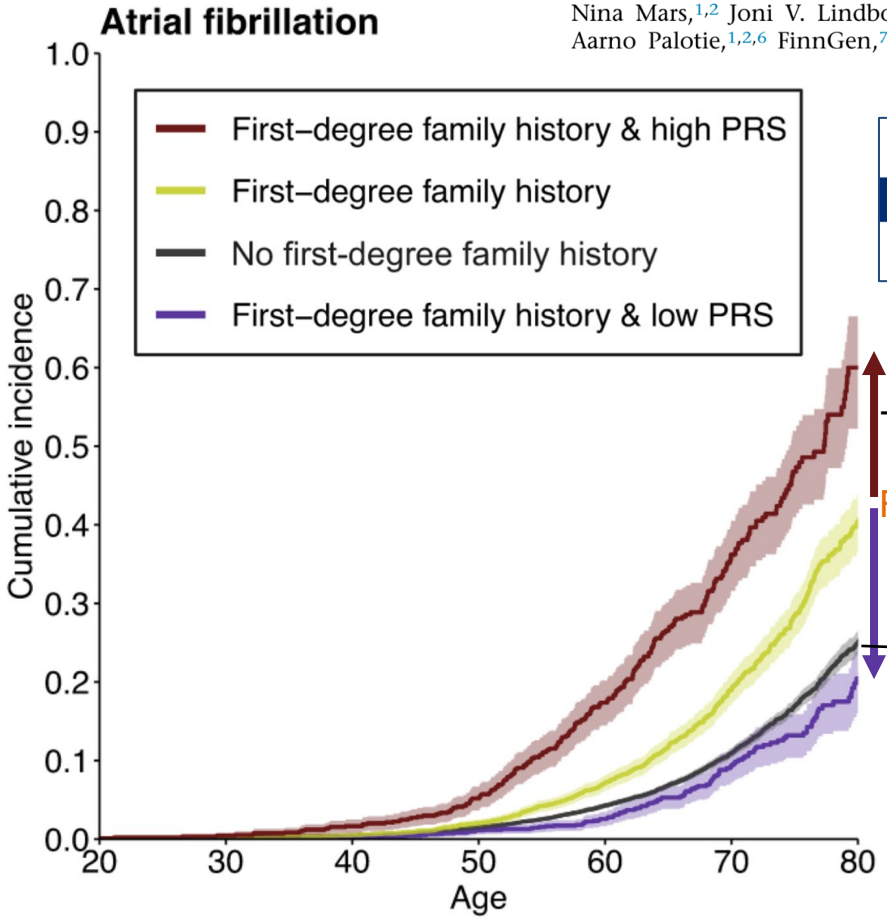
Hujoel, et al. Cell Genom 2022

Family history (FH) complements PGS

ARTICLE

Systematic comparison of family history and polygenic risk across 24 common diseases

Nina Mars,[1,2] Joni V. Lindbohm,[3,4,5] Pietro della Briotta Parolo,[1] Elisabeth Widén,[1] Jaakko Kaprio,[1,3] Aarno Palotie,[1,2,6] FinnGen,[7] and Samuli Ripatti[1,3,5,*]



FINNGEN

+ Positive PRS (increased risk)

Family history (high risk) (baseline, no PRS info)

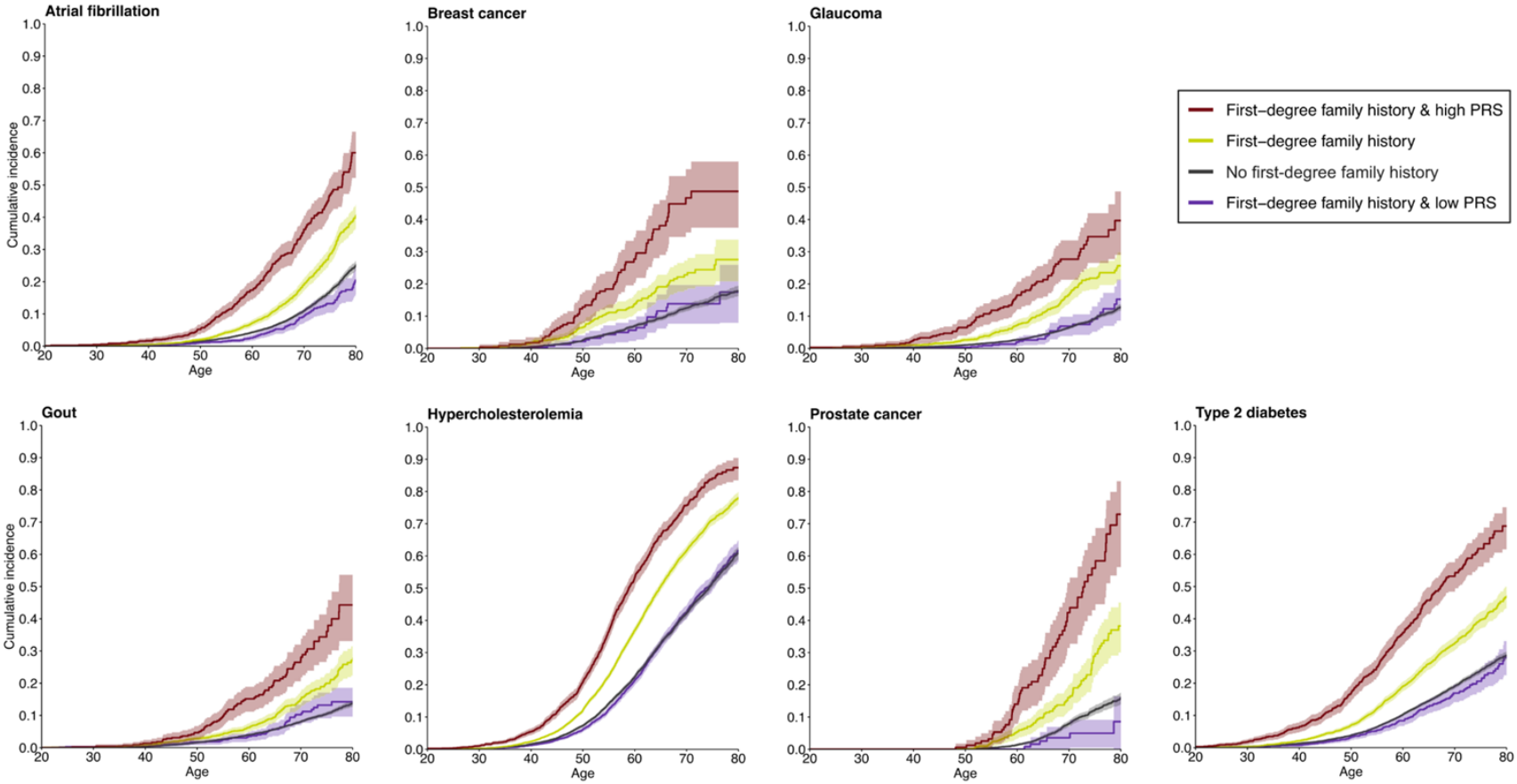
- Negative PRS (low risk)

No family history (low risk)

Family history (high risk) but low PRS!!

Mars, et al. AJHG 2022

Family history (FH) complements PGS



Mars, et al. AJHG 2022

Polygenic score (PGS) summary

- GWAS revealed large number of common variants contribute to complex traits; the individual effects of variants are small
- Polygenic scores (PGS) combine effects of disease-associated alleles for each individual
- PGS has potential relevance for clinical applications for some traits and for some populations
- PGS would be useful for research
- Current PGS models captures incomplete genetic liability of disease and PGS and family history are complementary to each other

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS

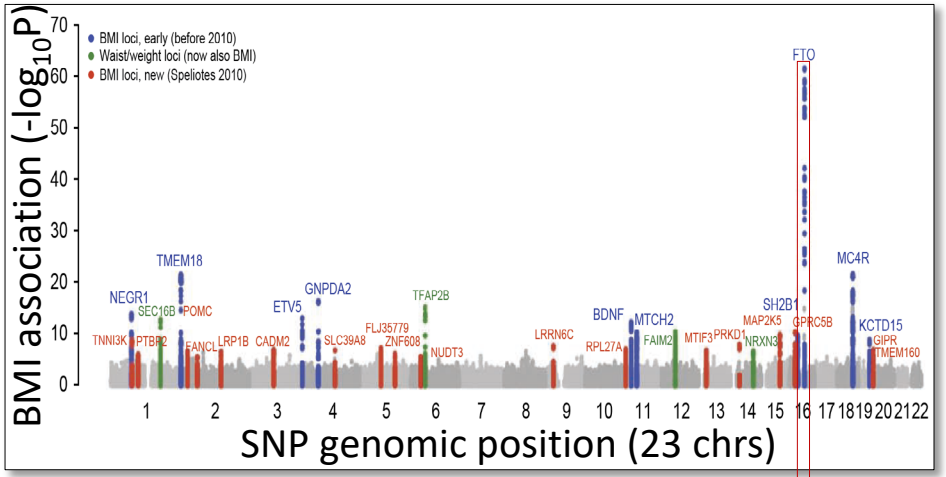
2. Polygenic scores (PGS)

3. From GWAS to biological insights

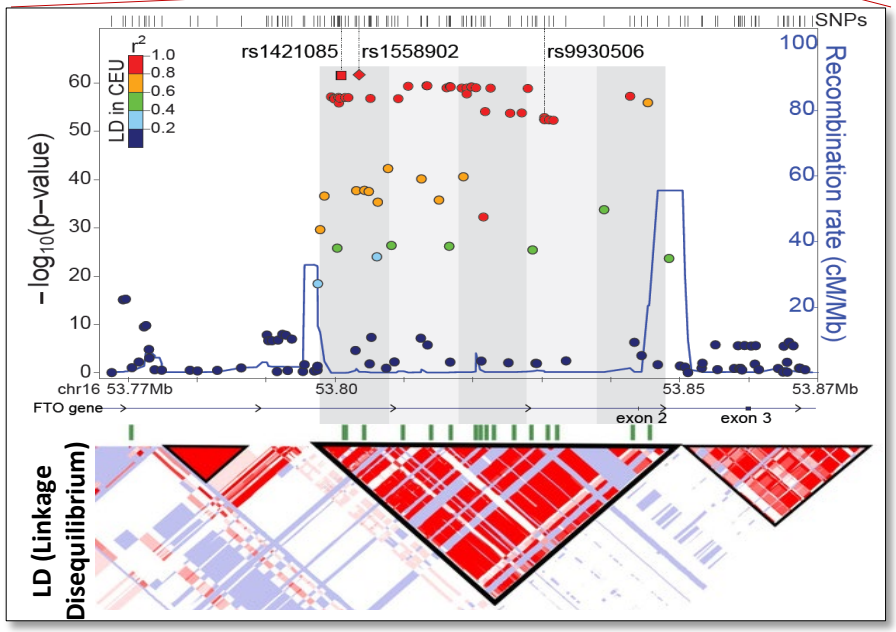
4. From Region to Mechanism / Circuitry

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

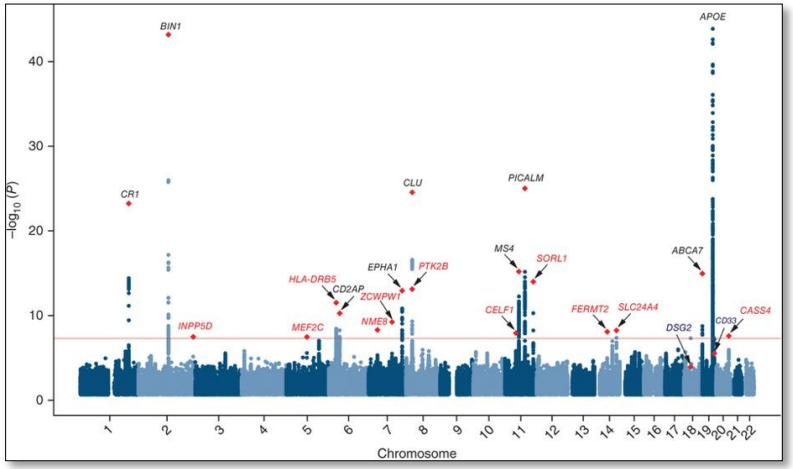
The promise of genetics

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

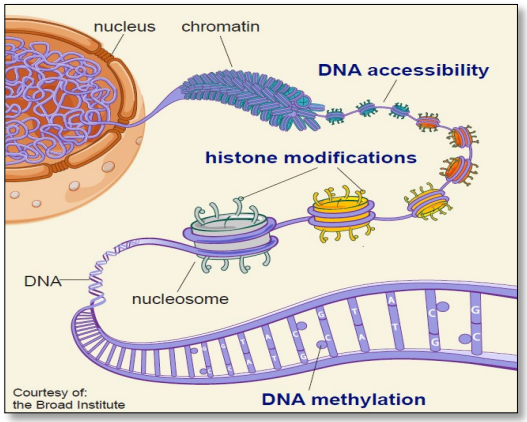
The challenge of mechanism

- **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

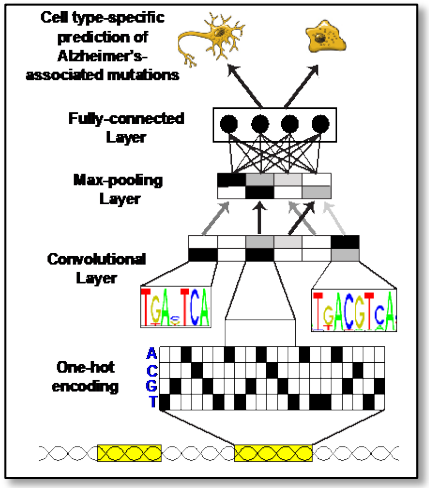
Dissect mechanisms of disease-associated regions



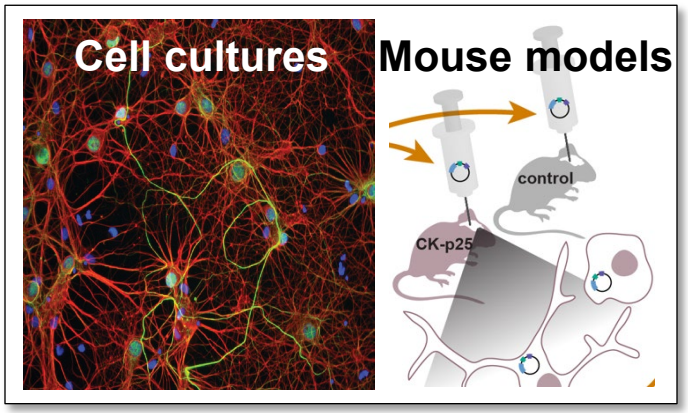
1. Disease genetics reveals common + rare variants/regions



2. Profile RNA + Epigenome in healthy + disease samples



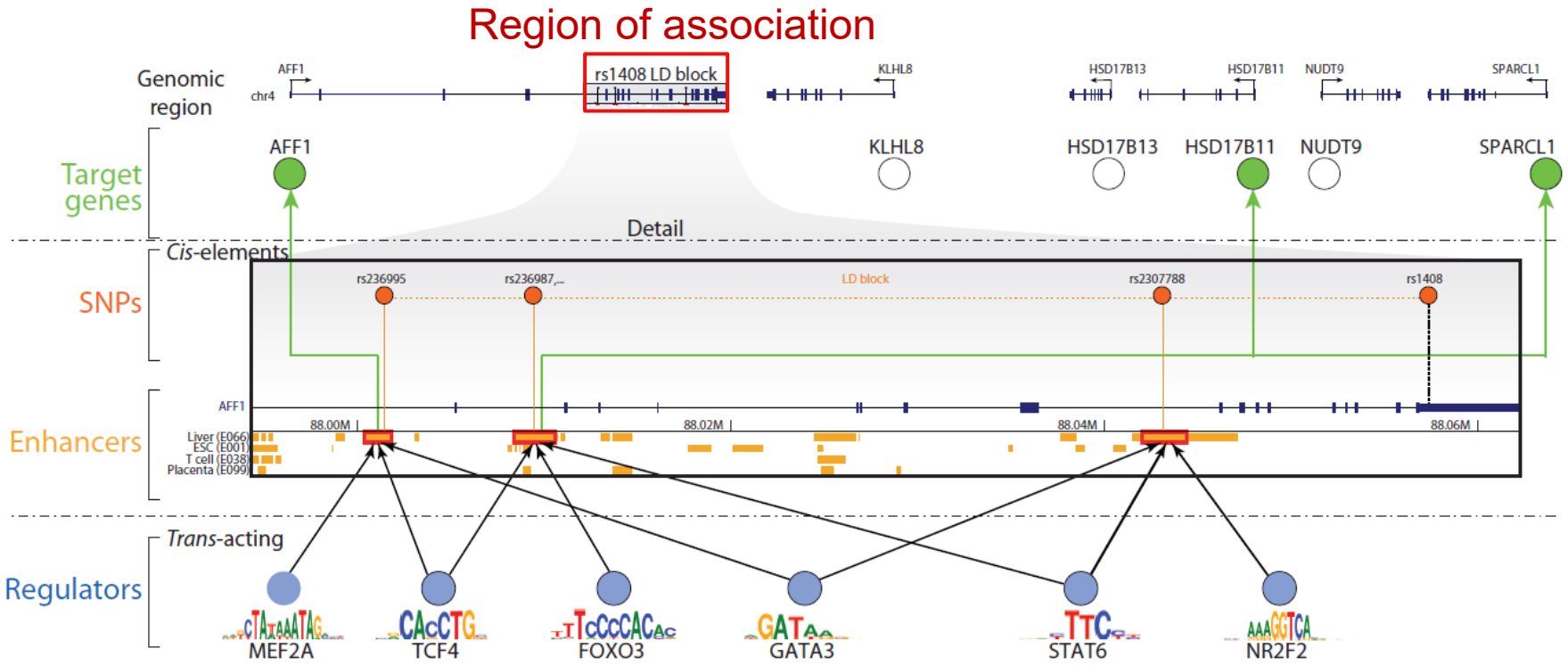
3. Integrate data to predict driver genes, regions, cell types



4. Validate predictions in human cells + mouse models

5. Disseminate results

Non-coding circuitry helps interpret disease loci

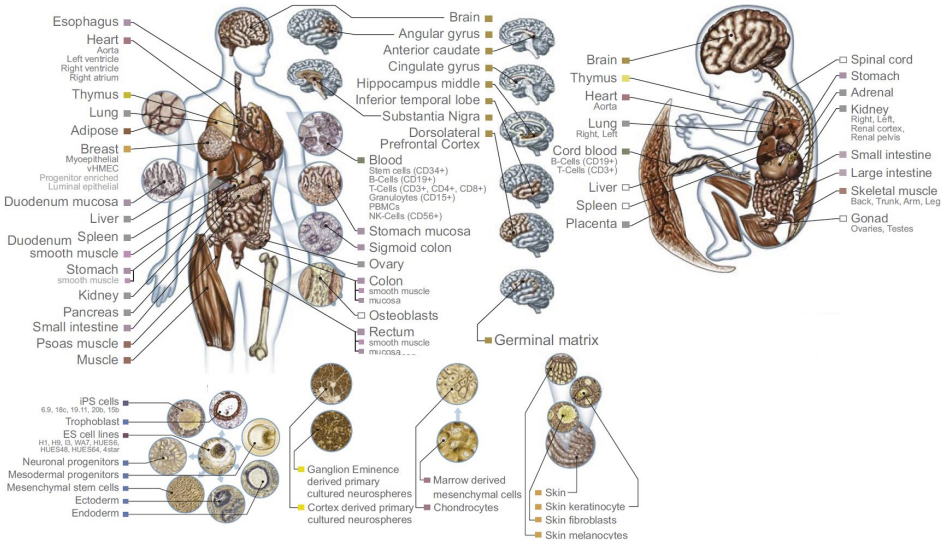


- Expand each GWAS locus using SNP linkage disequilibrium (LD)
  - Recognize **relevant cell types**: tissue-specific enhancer enrichment
  - Recognize **driver TFs**: enriched motifs in multiple GWAS loci
  - Recognize **target genes**: linked to causal enhancers

Quon bioRxiv 467852

Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



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

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

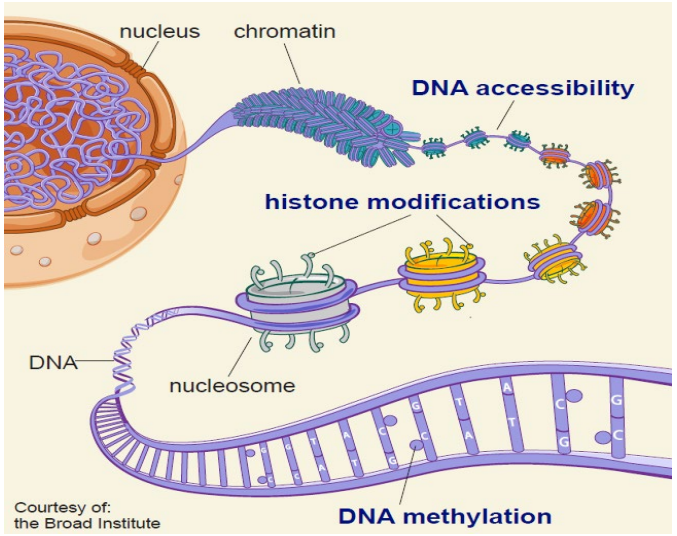
ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

X

Diverse epigenomic assays



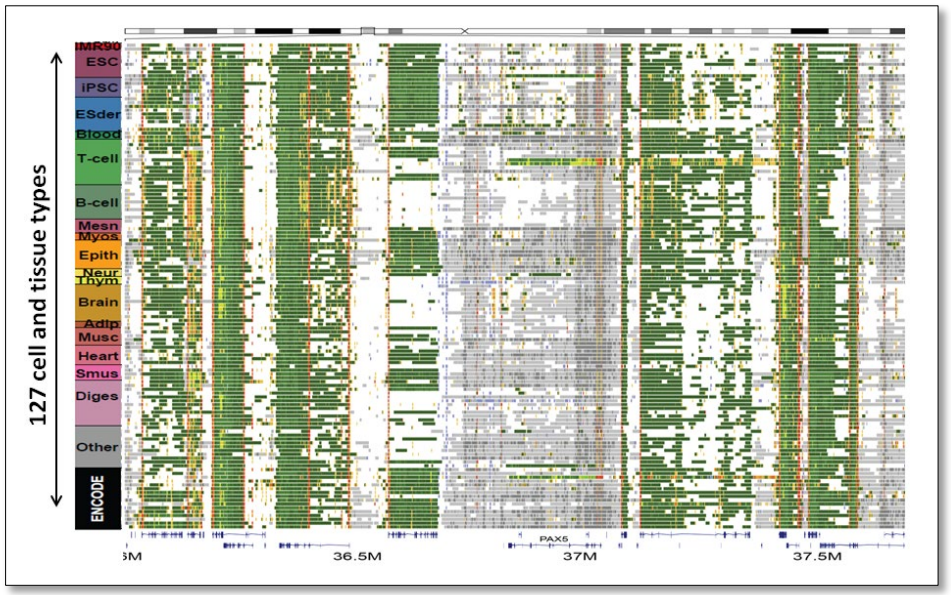
Histone modifications
- **H3K4me3, H3K4me1, H3K36me3**
- **H3K27me3, H3K9me3, H3K27/9ac**
- **+20 more**

Open chromatin:
- **DNA accessibility**

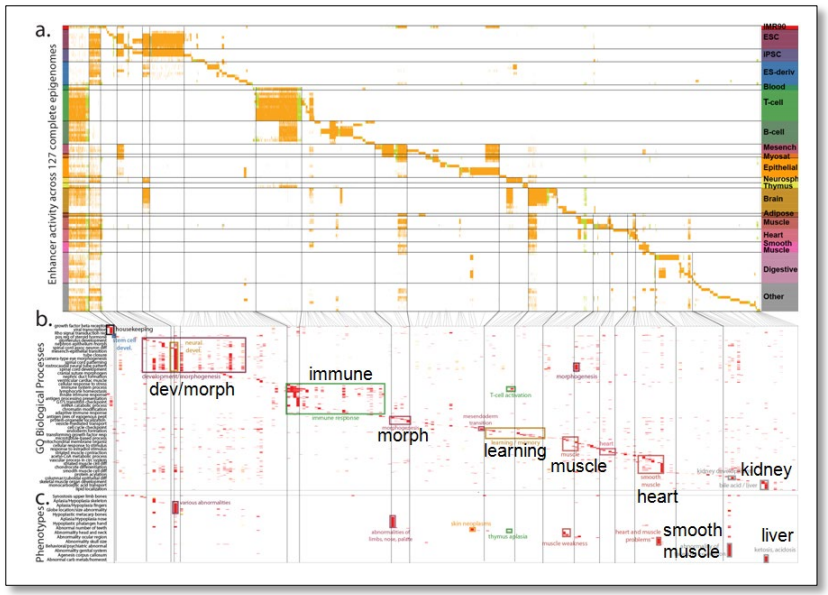
DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

Gene expression
- **RNA-seq, Exon Arrays**

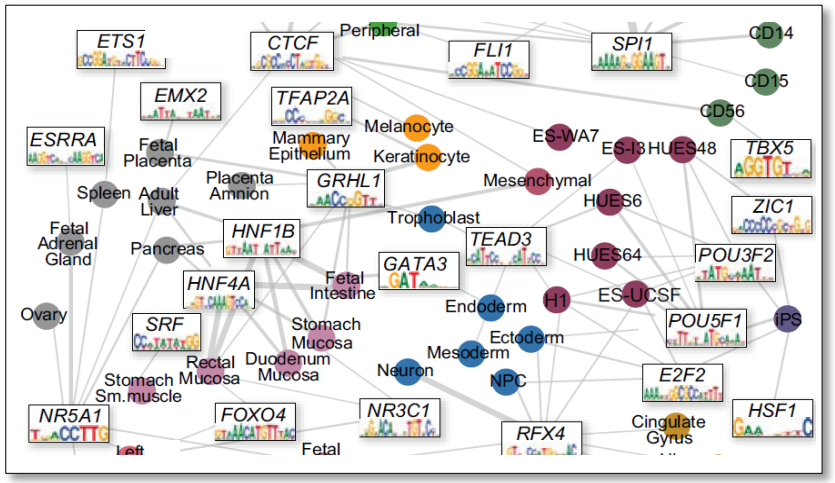
Enhancer modules, regulators, and target genes



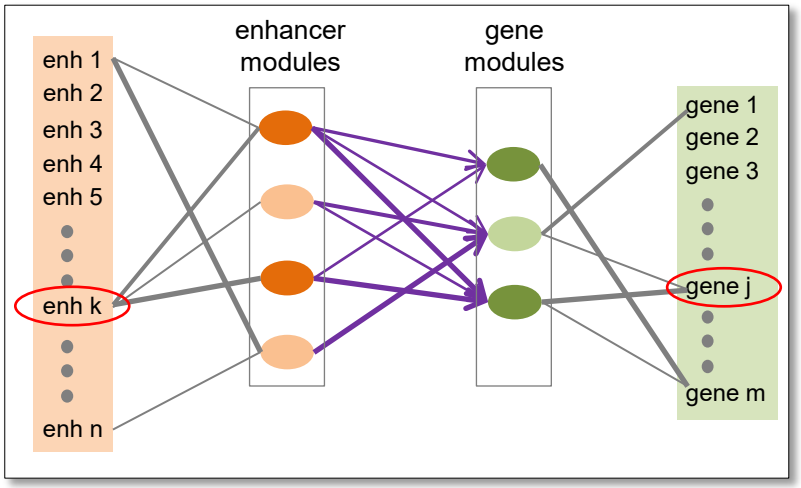
1. *Map chromatin states across 127 tissue/cells*



2. *Group enhancers into modules of common function*

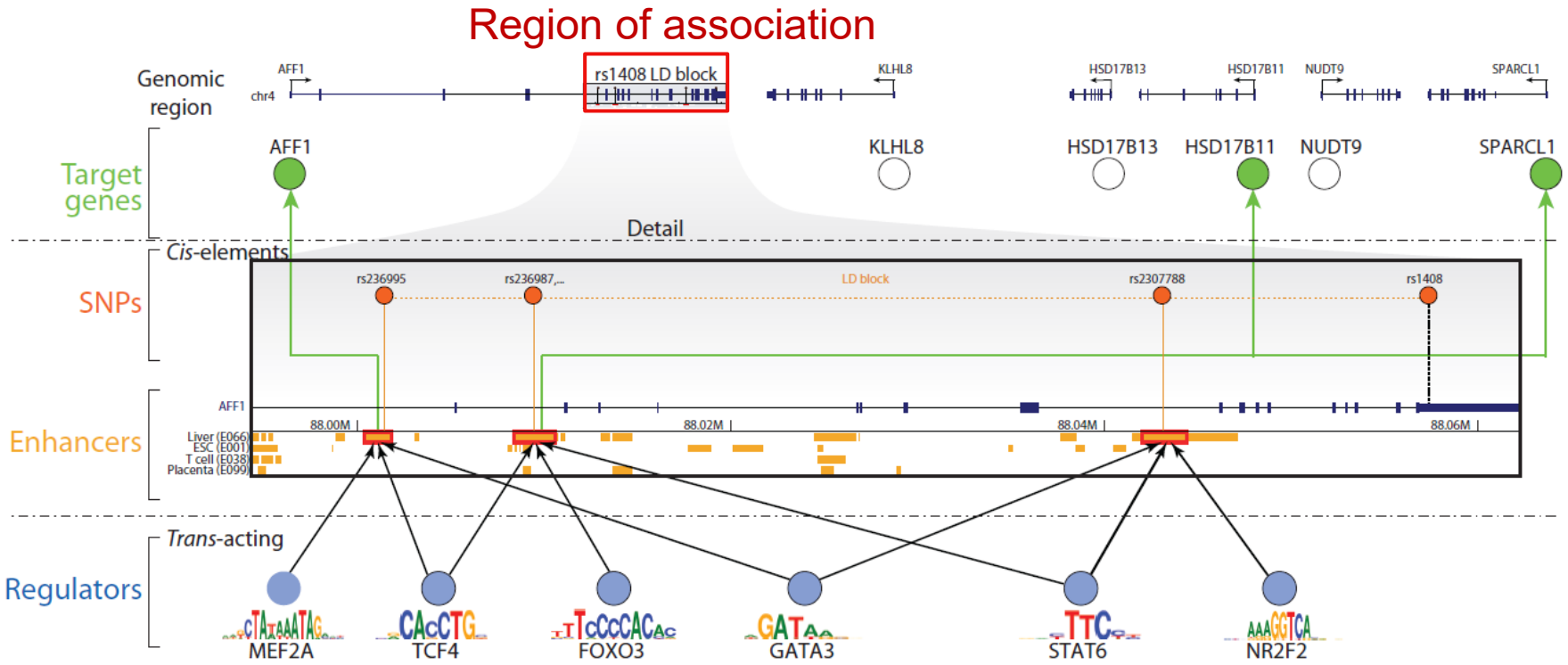


3. *Predict module regulators using motif enrichment*



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

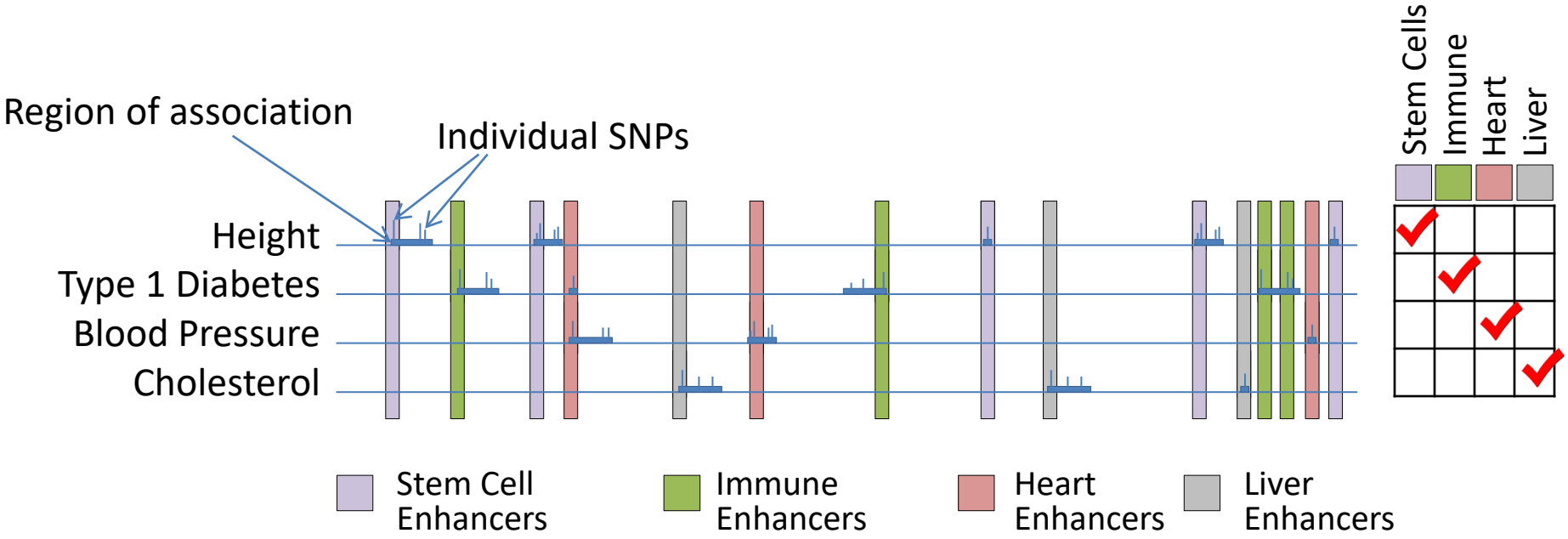
pos (hg19)	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	Drivers disrupted	GENCODE genes	dbSNP func annot
chr3:12329783	chr3:12288284	0.95	0.97	rs17036160	C	T	0.01	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 organs						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**

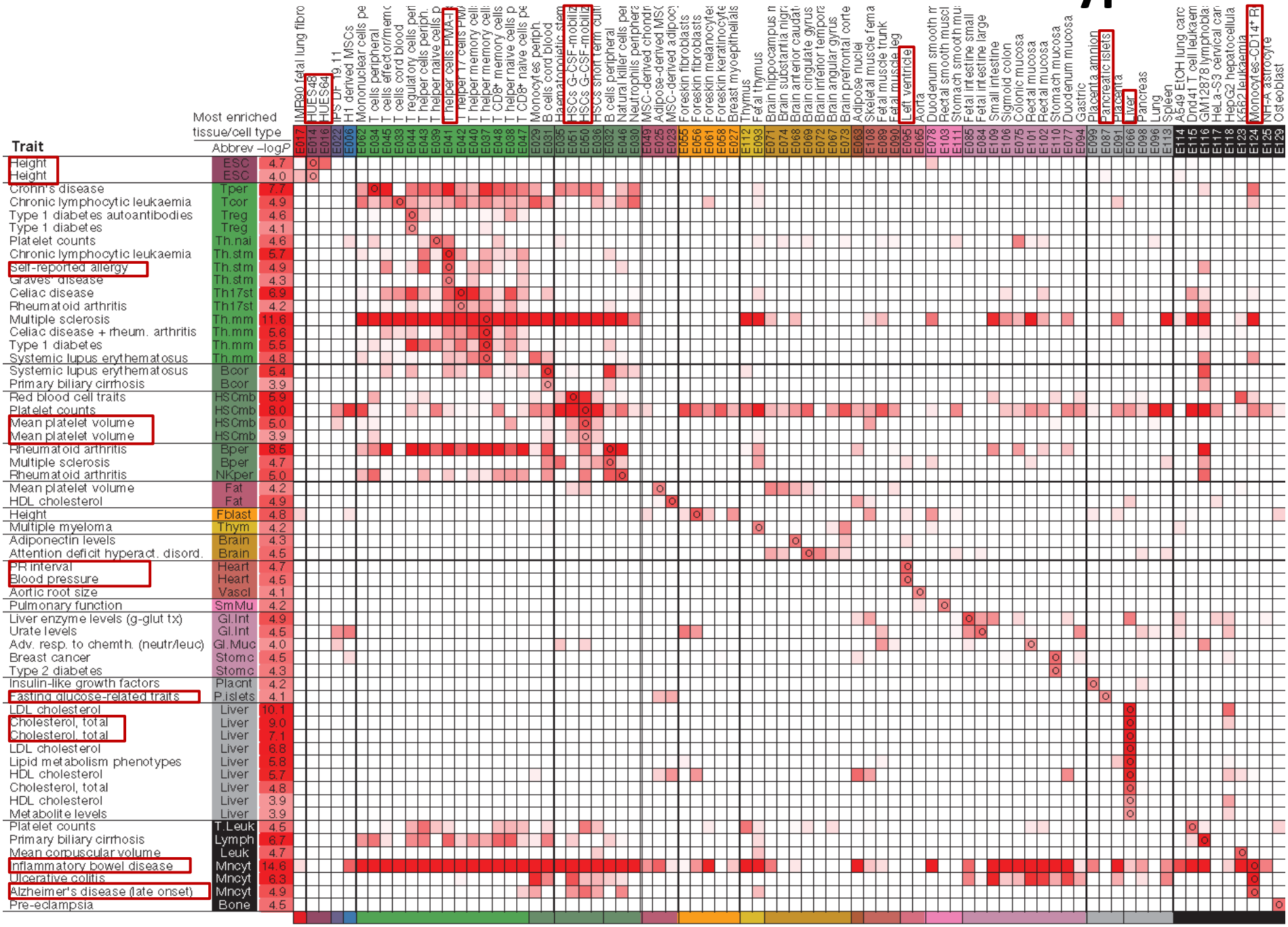
Ward, Kellis NAR 2011

Identifying disease-relevant cell types

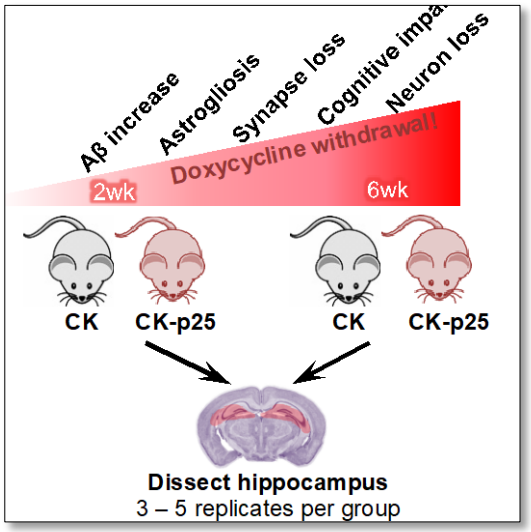


- For every trait in the GWAS catalog:
  - Identify all associated regions at P-value threshold
  - Consider all SNPs in credible interval ($R^2 \geq .8$)
  - Evaluate overlap with tissue-specific enhancers
  - Keep tissues showing significant enrichment ($P<0.001$)
- Repeat for all traits (rows) and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

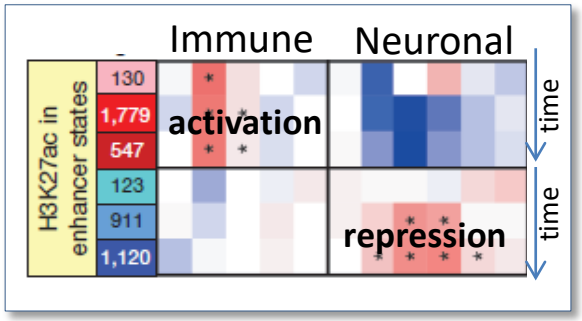
GWAS hits in enhancers of relevant cell types



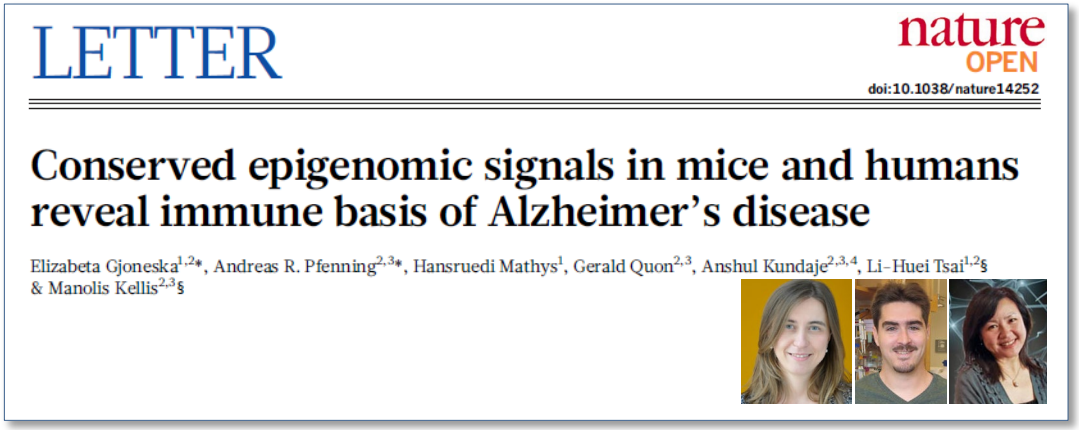
Immune activation + neural repression in human + mouse

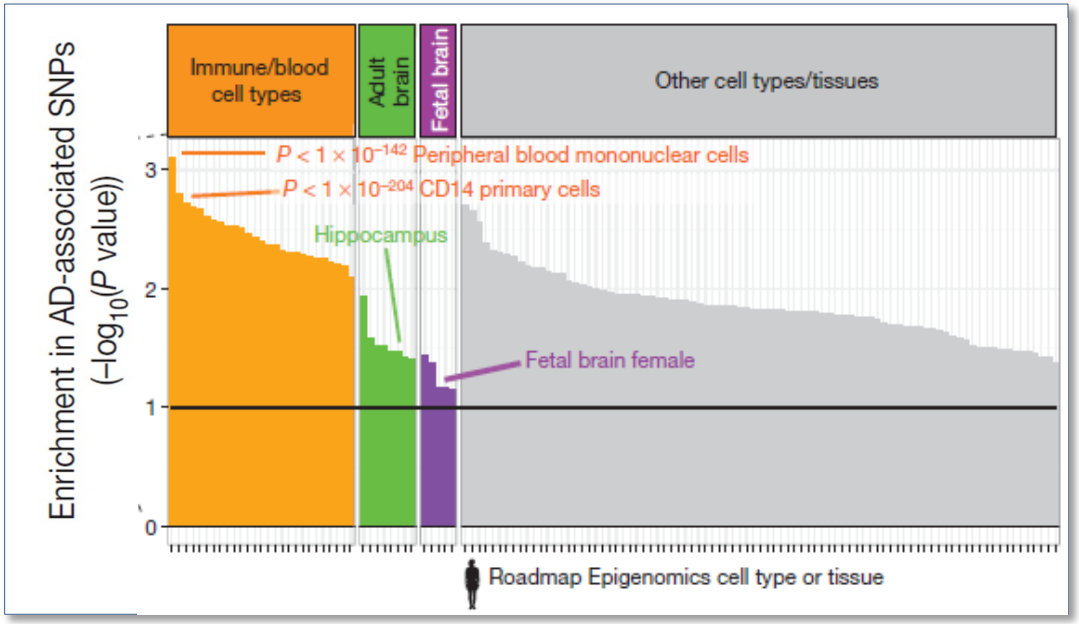


Epigenomics of AD progression



Immune activation precedes neuronal repression

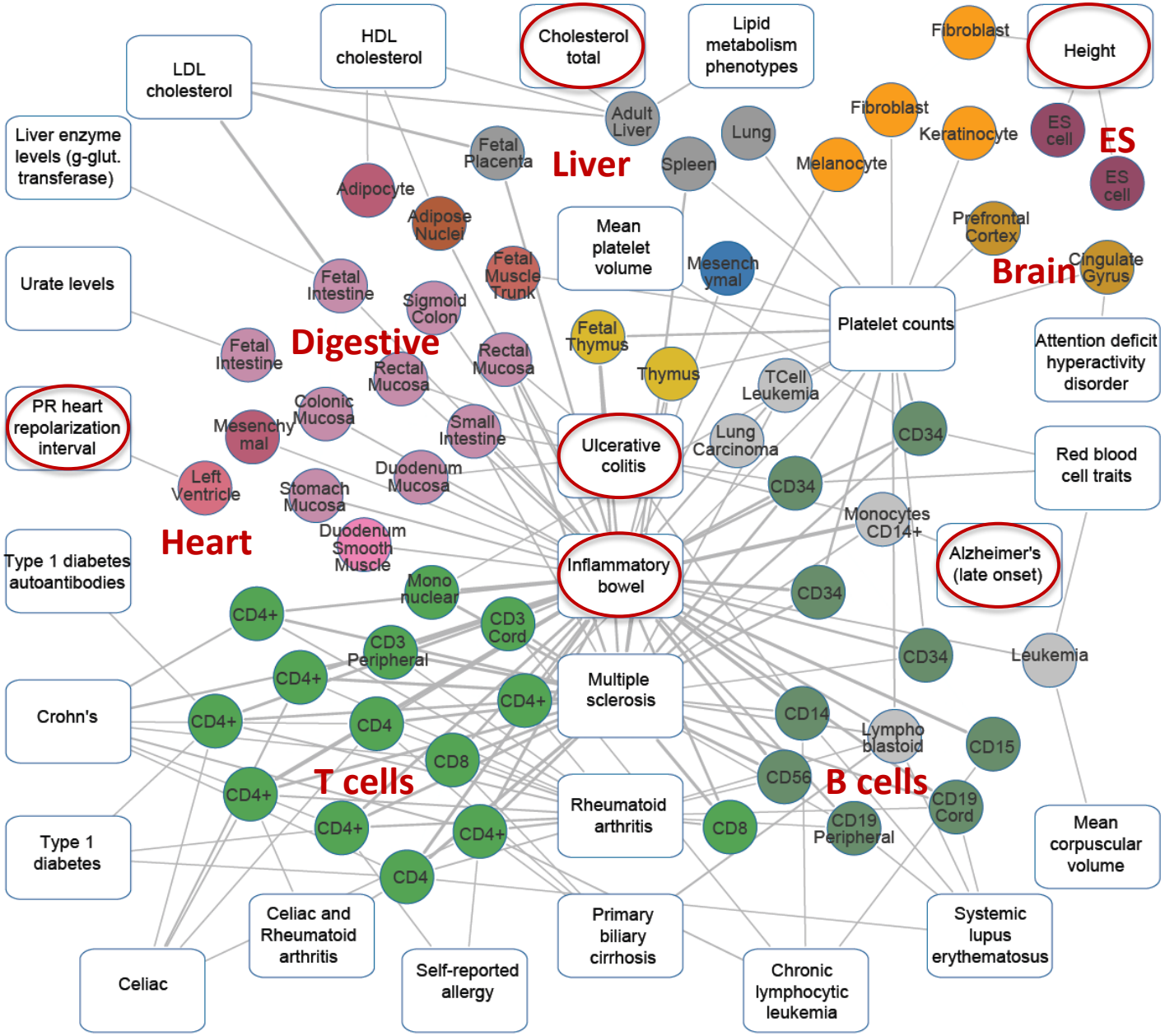




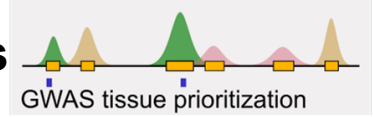
AD variants localize in immune cells, not neuronal

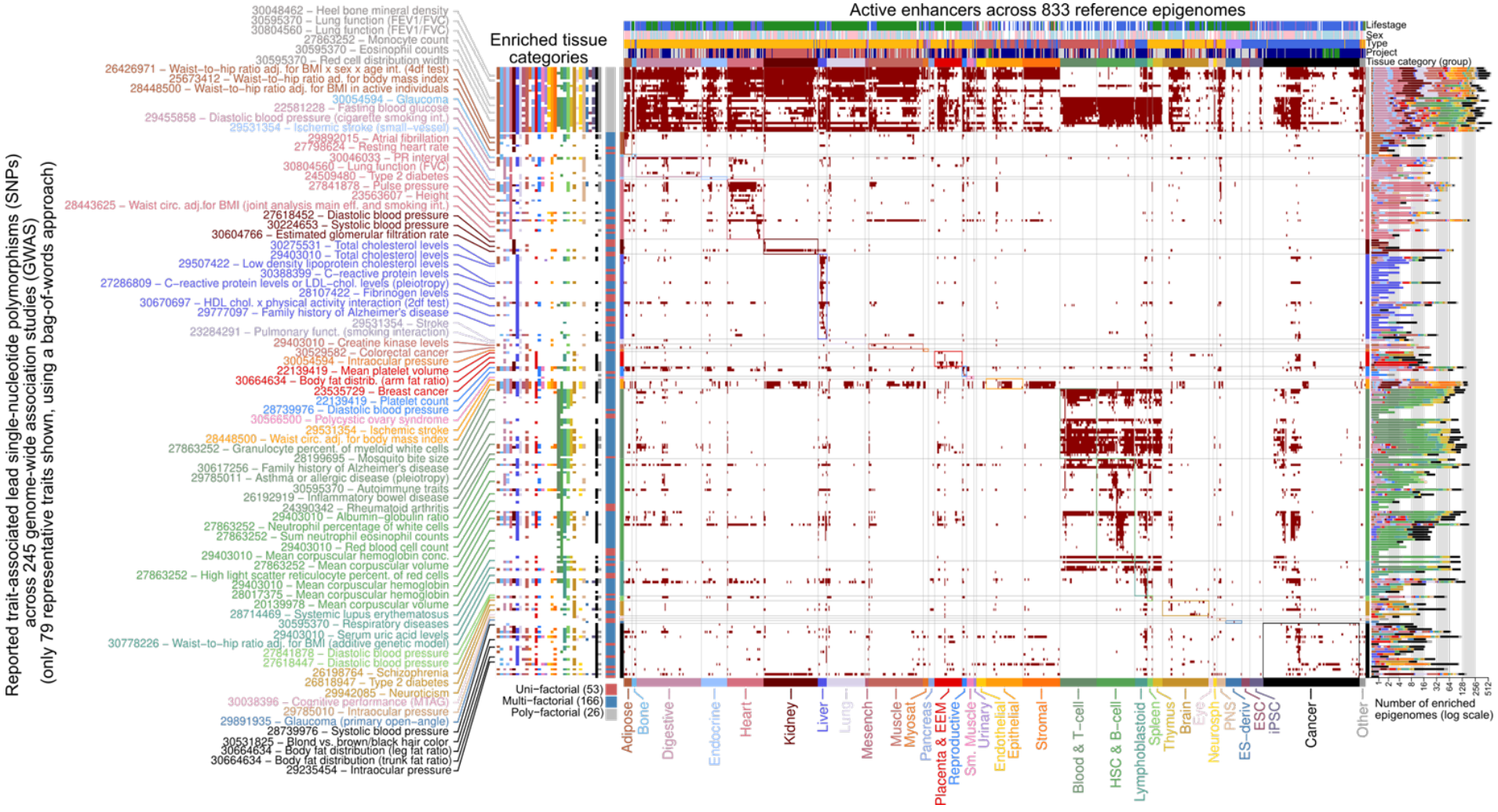
Inflammation as the causal component of Alzheimer's disease

Linking traits to their relevant cell/tissue types



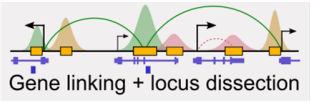
Predict tissues for 200+ traits by epigenome enrichments



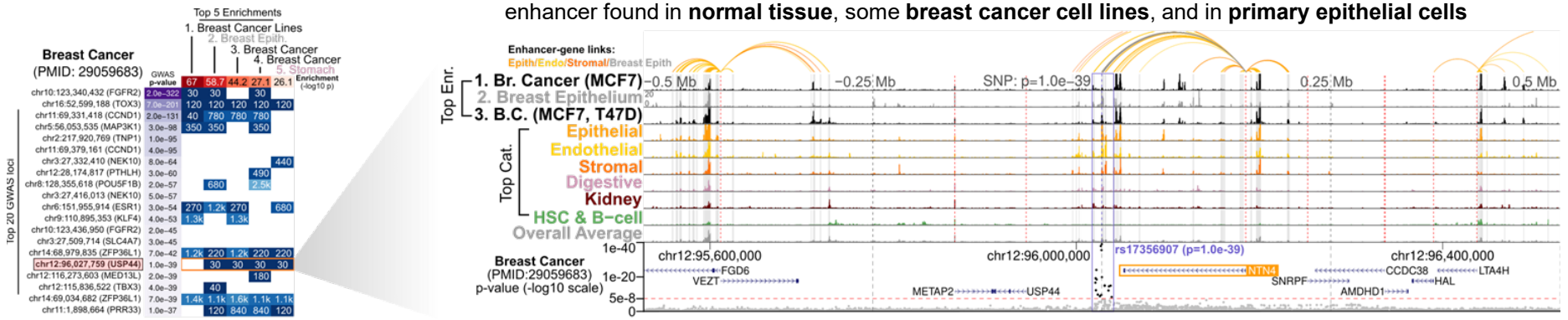


Tissue x GWAS enrichments ➔ 245 GWAS-tissue enrichments, most in novel epigenomes (shown)
Enhancer-tree enrichments ➔ 540 GWAS-tissue enrichments, focusing at right level of resolution on tree

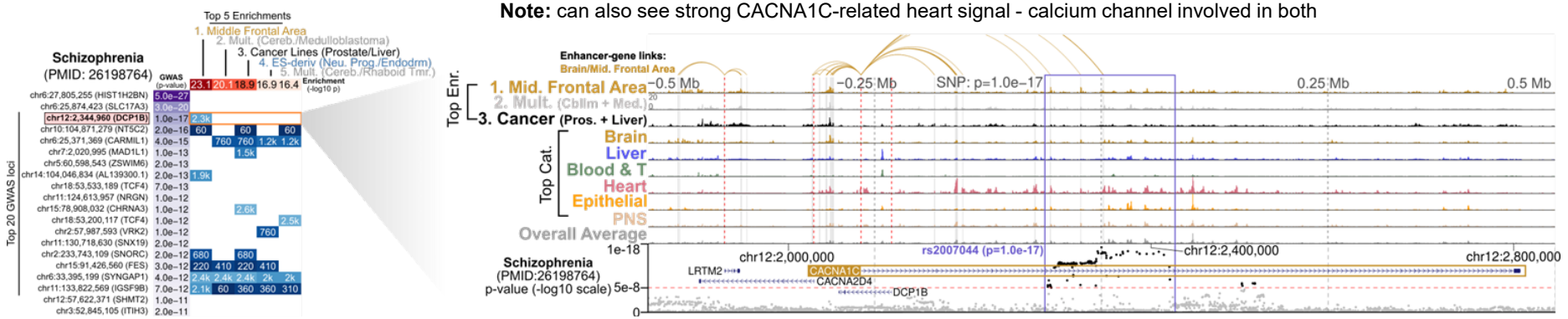
GWAS locus dissection: enriched tissues, driver SNPs, target genes



Example1: Localized breast cancer signal in USP44 locus links strongly to NTN4 gene (assoc. w/ prognosis, metastasis)
enhancer found in **normal tissue**, some **breast cancer cell lines**, and in **primary epithelial cells**

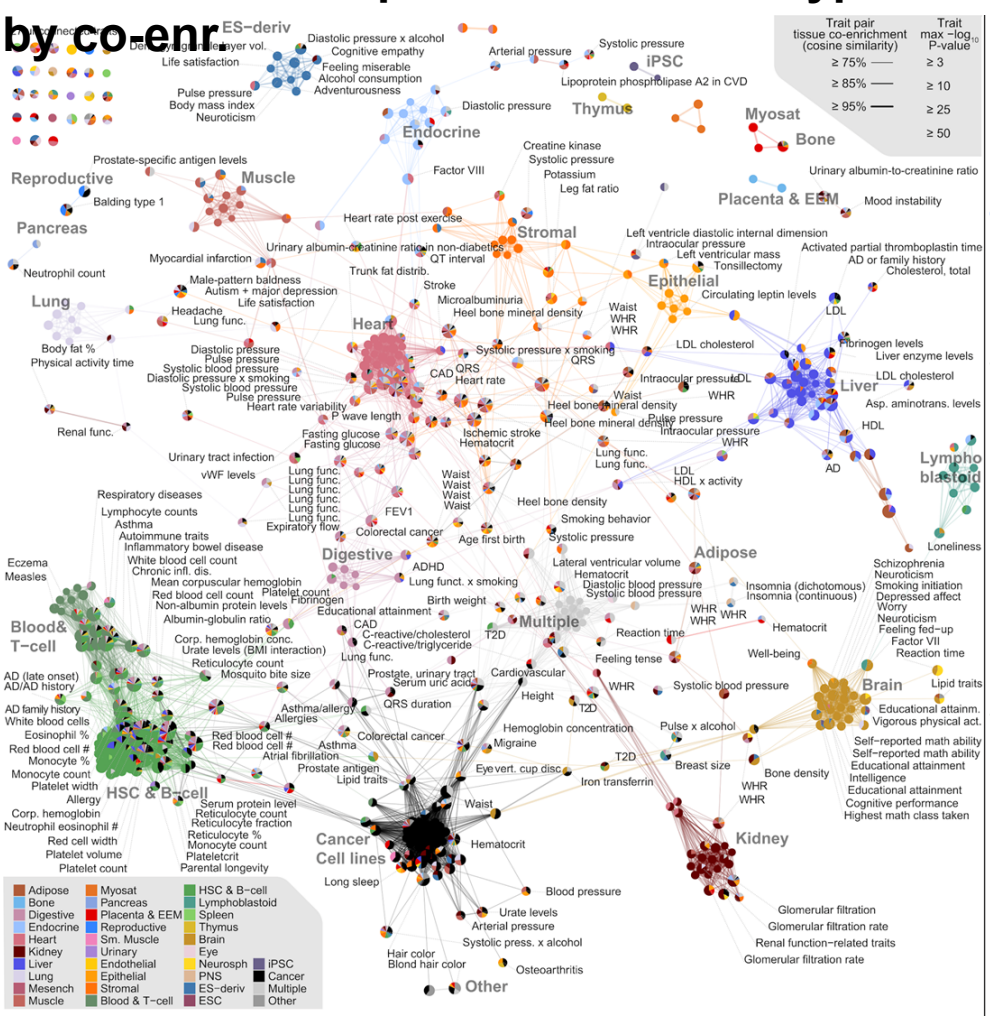


Example 2: Broad schizophrenia signal in the CACNA1C locus in USP44 locus links to CACNA1C gene through multiple enhancers
Note: can also see strong CACNA1C-related heart signal - calcium channel involved in both



Looking at enrichment-prioritized tissues for each GWAS shows **tissue-specific activity of enhancers in loci**.
Locus dissection for a **localized signal (NTN4 - breast cancer)** and a **broad signal (CACNA1C - schizophrenia)**

GWAS-Tissue enrichment: pred. disease cell type/tissue in 500+ traits + network of sim. traits by co-enr.



Range of trait complexity in epigenomic similarity network:

1. Uni-factorial traits (cores):

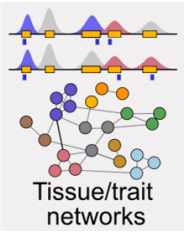
QT/PR intervals/QRS (heart)
C-reactive protein (liver)
TSH levels (endocrine)
Educational attain. (brain)
Schizophrenia (brain)
Life satisfaction (ES-deriv neur)
Glomer. filtration rate (kidney)
Autoimmune traits (T-cells)
Monocyte count (HSC & B-cell)

2. Multi-factorial (connect):

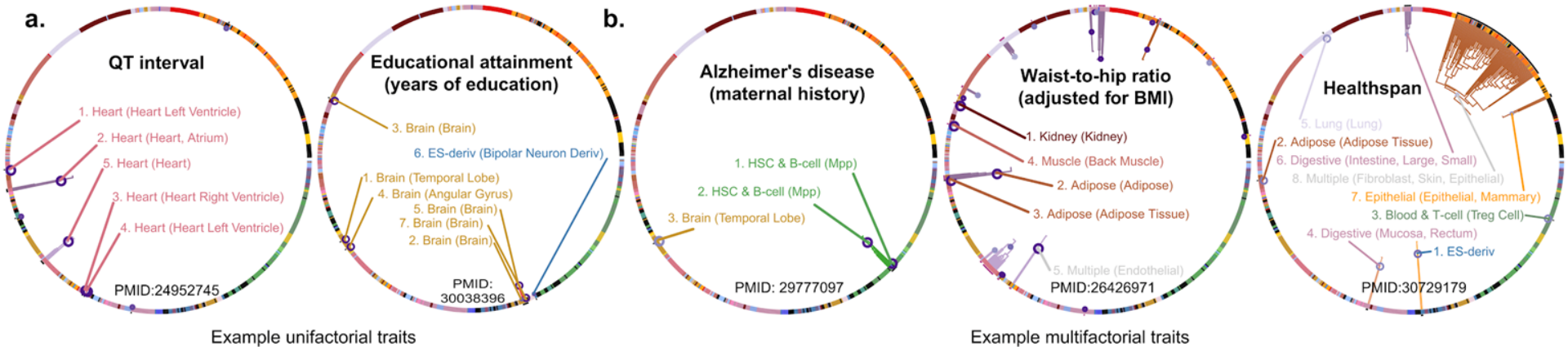
CAD (heart, endocrine, liver)
HDL/triglycerides (liver/ adipose)
Lung FEV1, FVC (lung, heart, digestive)
Blood pressure (heart with endocrine, endothelial, and liver)
Alzheimer's (immune and brain)
Blood cell fractions (principal blood with liver, digestive, other)

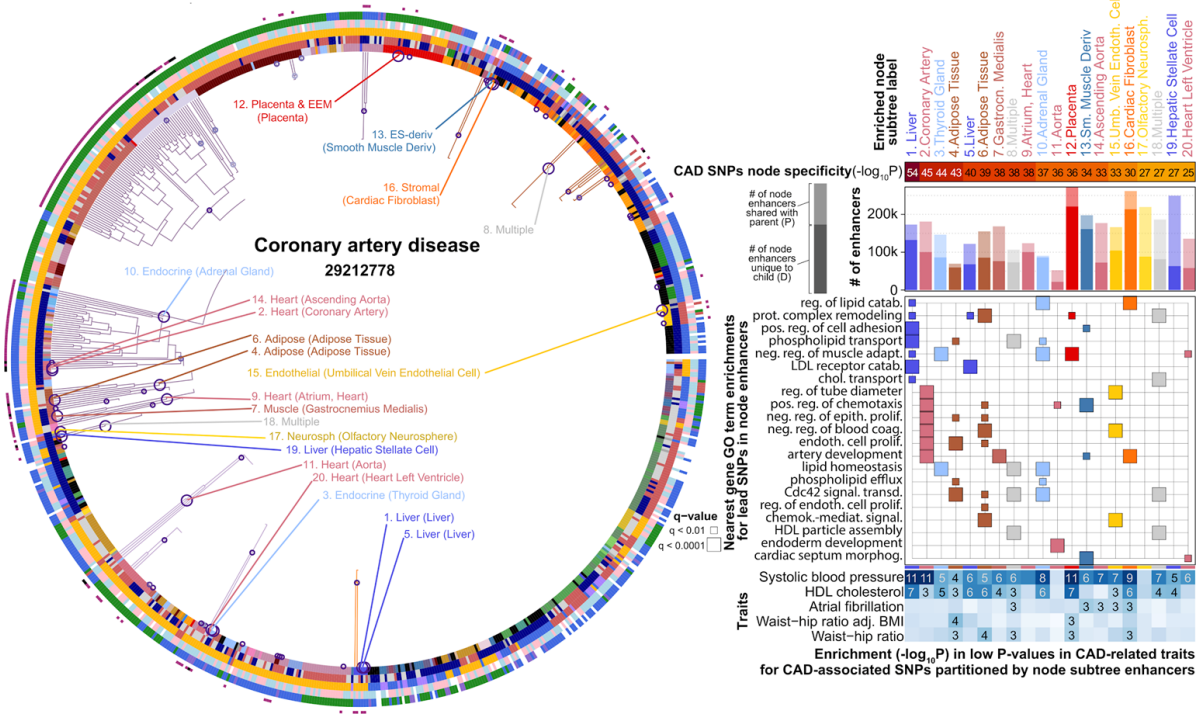
3. Poly-factorial:

Waist-hip ratio measures
Heel bone mineral density



1. **Range of unifactorial → poly-factorial traits**
2. **Partitioning multi-factorial traits into tissues + pathways of action**

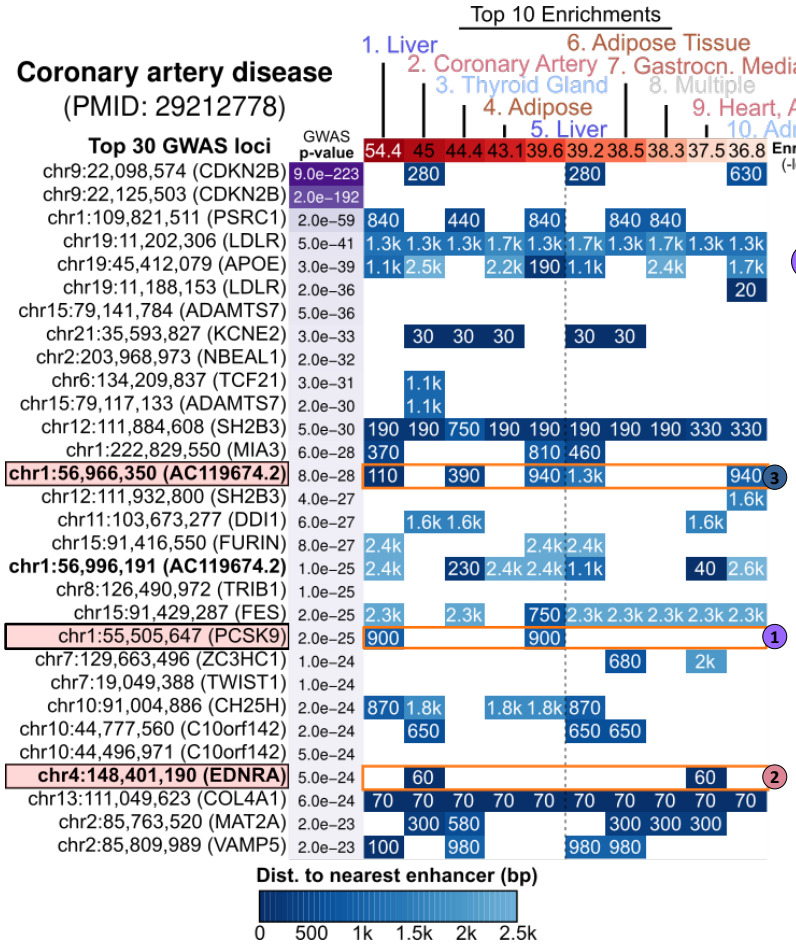




2. CAD is partitioned by top nodes:

Heart	Liver	Adipose	Endocrine
212 SNPs	169 SNPs	122 SNPs	121 SNPs
artery, cardiac, vessel morph.	chol./lipid metab. and transport	adipose tissue innervation	lipid homeostasis
Systolic BP, Atrial fib.	Systolic BP, HDL chol.	Waist-hip-ratio, HDL chol.	Systolic BP, HDL chol.

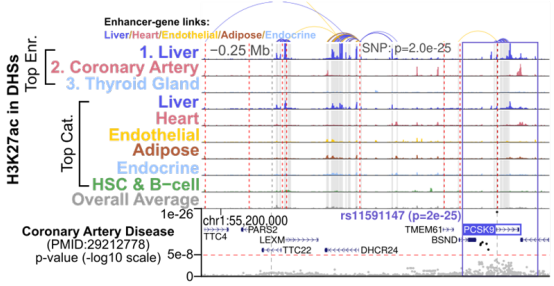
CAD locus analysis illustrates both GWAS-level and locus-level pleiotropy



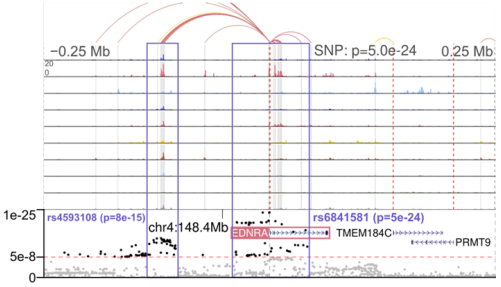
Lead SNP by tissue heatmap:

- SNPs only overlapping heart enhancers (eg. EDNRA, TCF21, ADAMTS7)
- SNPs only overlapping liver (eg. PCSK9)
- SNPs without overlaps (non-enhancer/conditions not captured?)
- SNPs with multiple tissue overlaps (LDLR, APOE, SH2B3, PLPP3)

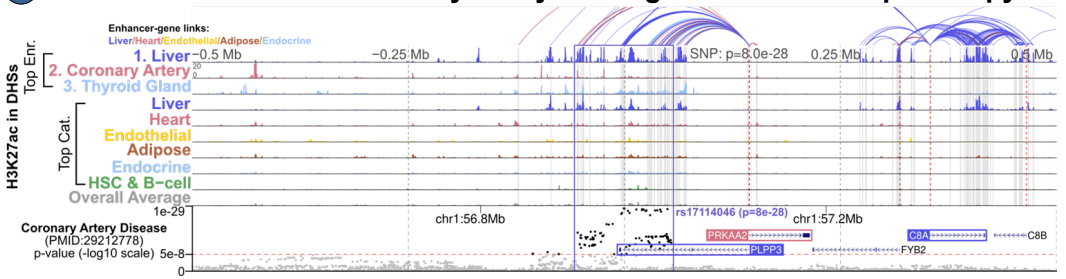
1 **PCSK9: Liver-only** mechanism, mediated through primarily one variant



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



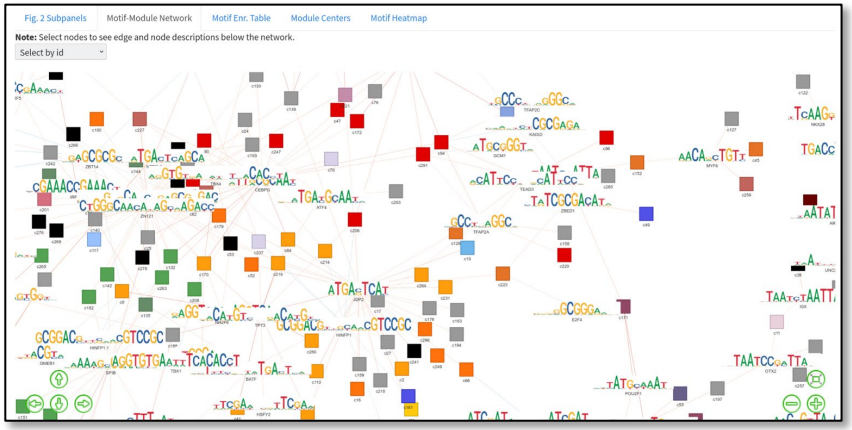
3 **PLPP3: Both liver and coronary artery: multi-gene/multi-tissue pleiotropy**



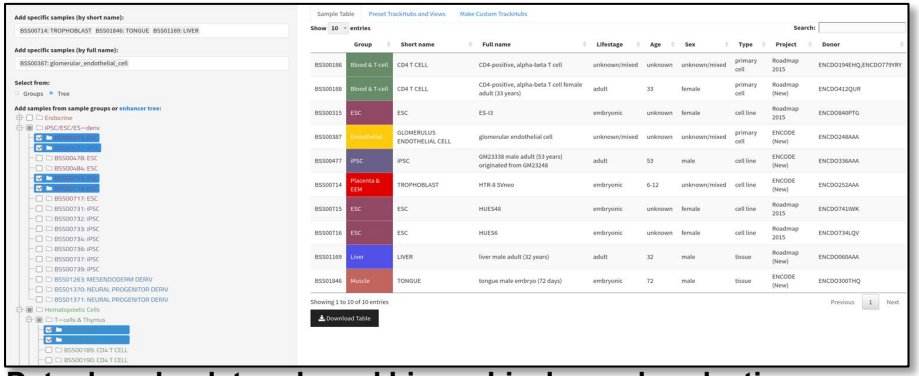
Exploration: Develop browser for: epimap data / gene regulation analysis / GWAS analysis

Interactive browser including:

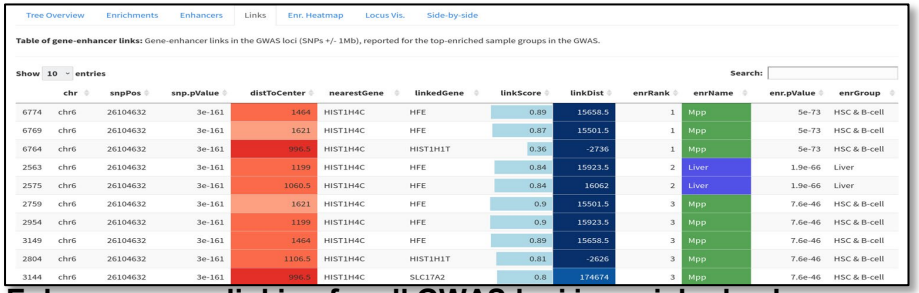
- Custom track hub creation
- Modules-motifs network
- GWAS enrichments
- Per-GWAS locus visualizations



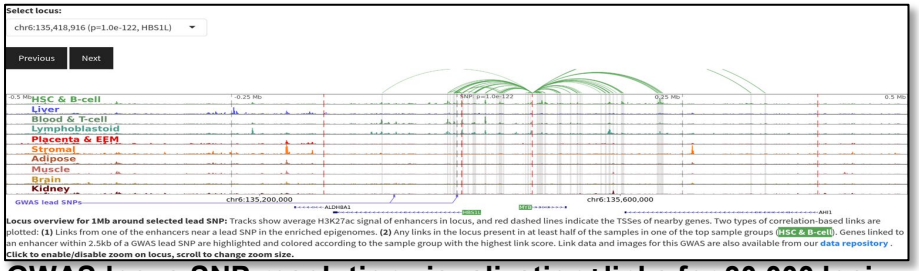
Gene-regulatory circuitry: TF-enhancer regulator linking



Data download: tree-based hierarchical sample selection

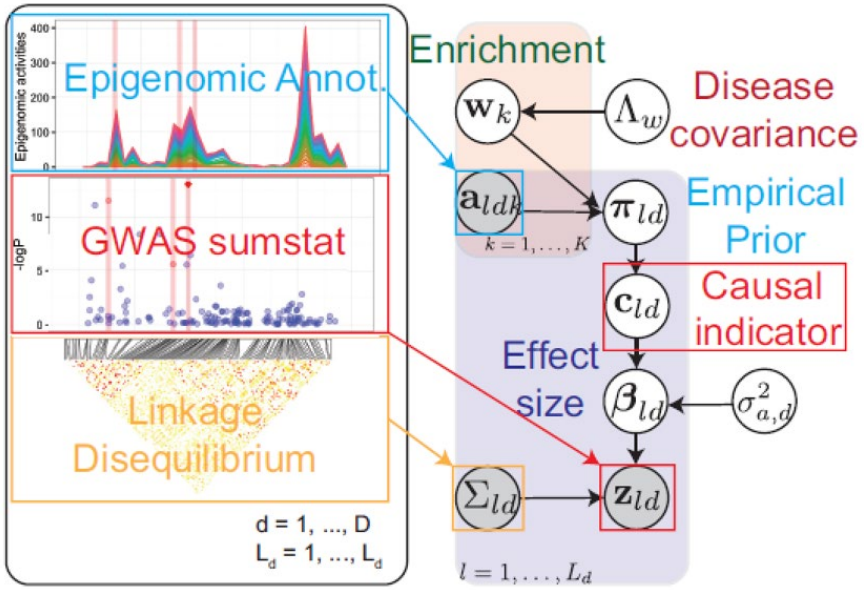


Enhancer-gene linking for all GWAS loci in enriched enhancers

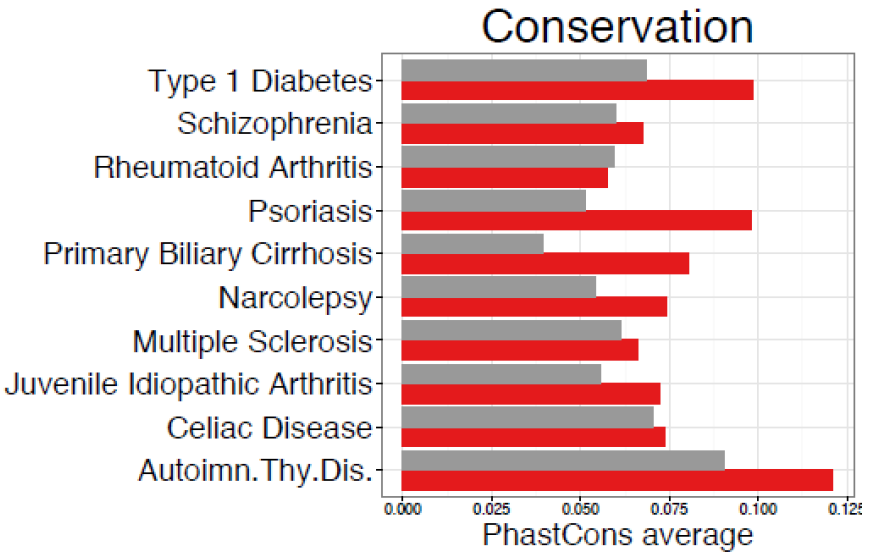


GWAS locus SNP-resolution visualization+links for 30,000 loci

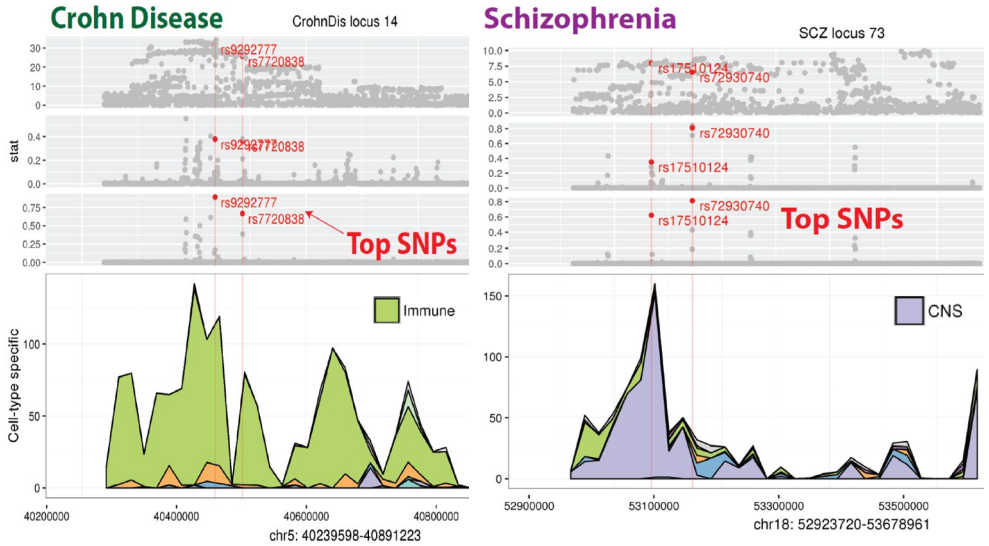
Bayesian fine-mapping: Predict causal variant and cell type



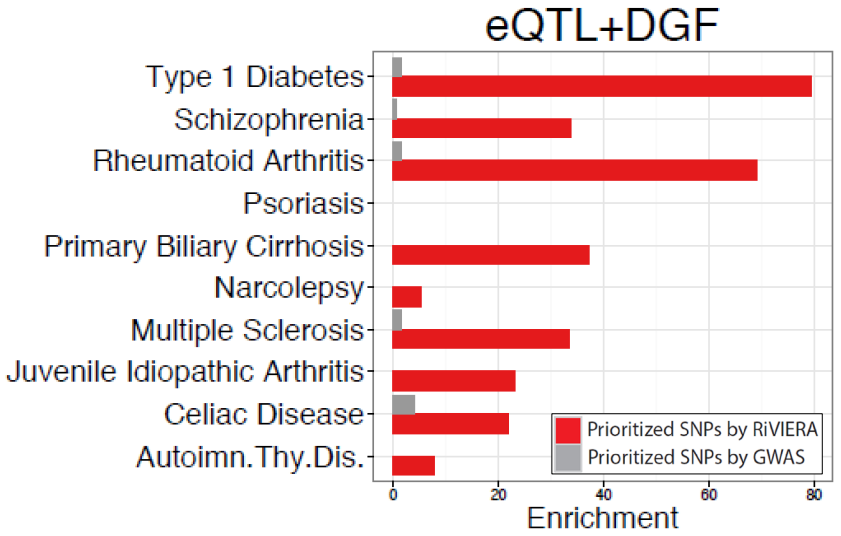
RiVIERA: multi-trait GWAS integration



Capture conserved elements

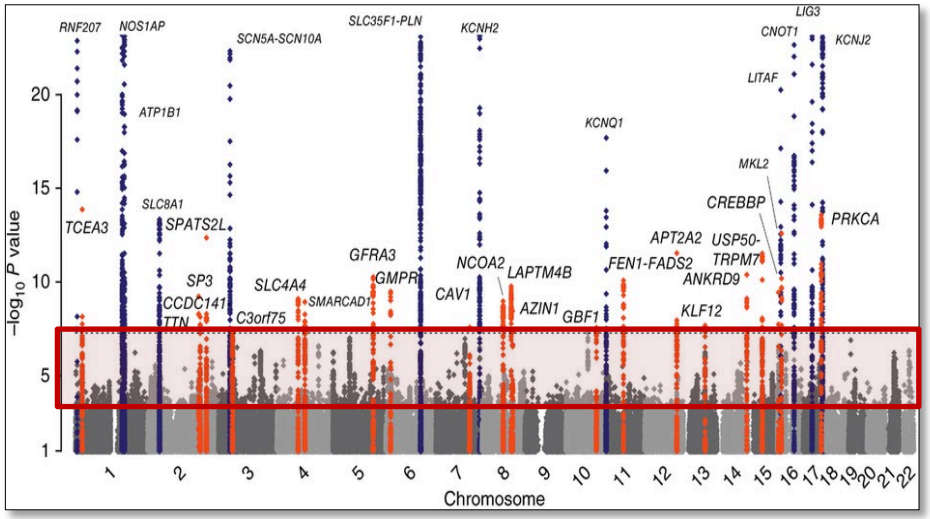


Predict causal variants and cell types



Capture eQTLs from GTEx

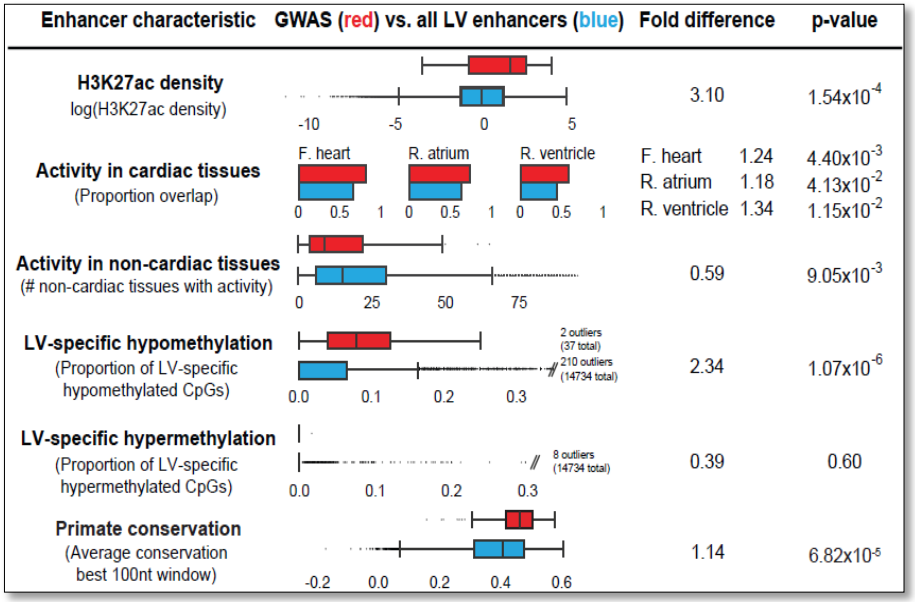
Combine GWAS+Epig to find new target genes/SNPs



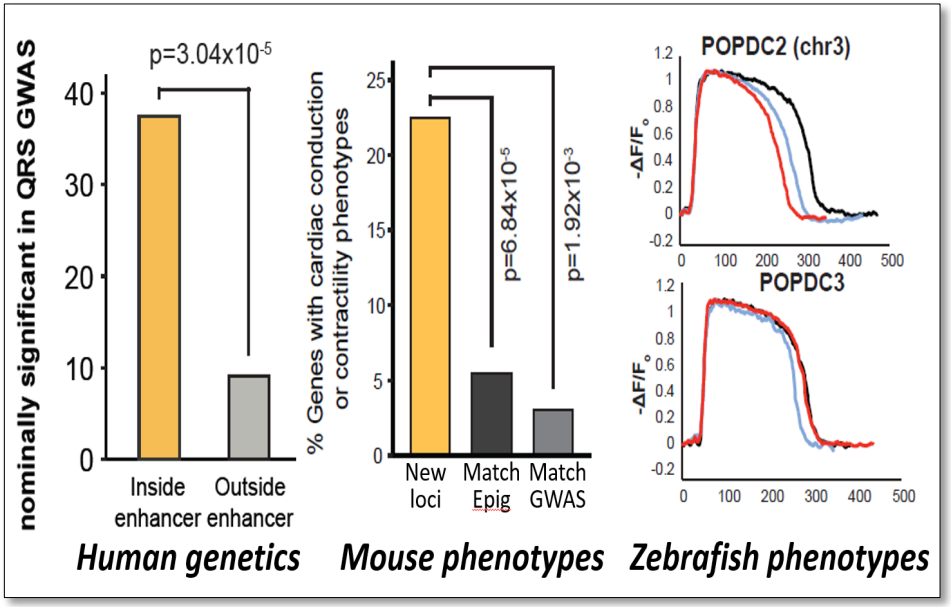
Lead SNP	p-value	Enhancer	1. Luciferase reporter	2. 4C-seq interactions
rs1886512	4.30×10^{-8}	chr13:74,520,000-74,520,400	0.015	No interactions
rs1044503	5.13×10^{-7}	chr14:102,965,400-102,972,000	4.70×10^{-9}	CINP, RCOR1
rs10030238	6.21×10^{-7}	chr4:141,807,800-141,809,600	1.35×10^{-14}	RNF150
		chr4:141,900,800-141,908,000	-	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

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS

2. Polygenic scores (PGS)

3. From GWAS to biological insights

4. From Region to Mechanism / Circuitry

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

Collaborators and contributors

MIT / Broad Institute


Melina Claussnitzer


Gerald Quon


Wouter Meuleman


Nezar Abdennur


Manolis Kellis

U Bergen, Norway


Simon Dankel


Gunnar Mellgren

U. Toronto


Chi-Chung Hui


Kyoung-Han Kim

Munich

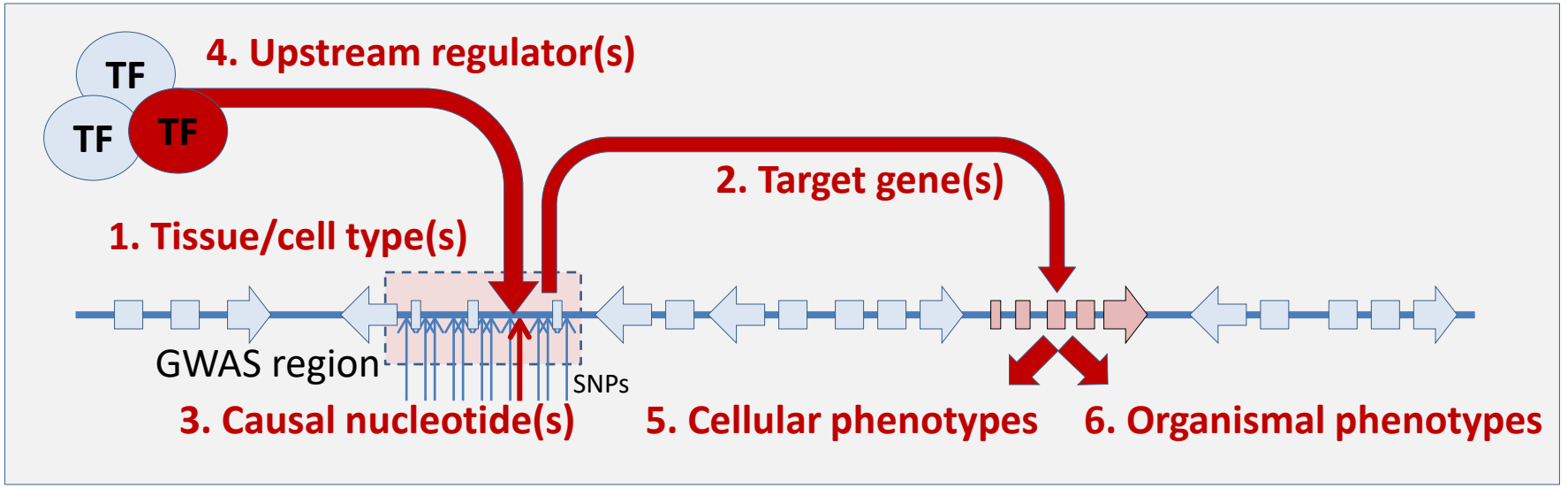

Hans Hauner

Harvard


Yi-Hsiang Hsu

Funding: NIH: NHGRI, Common Fund; Kroner-Fresenius

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

This talk:

Apply these to the FTO locus

The NEW ENGLAND JOURNAL of MEDICINE

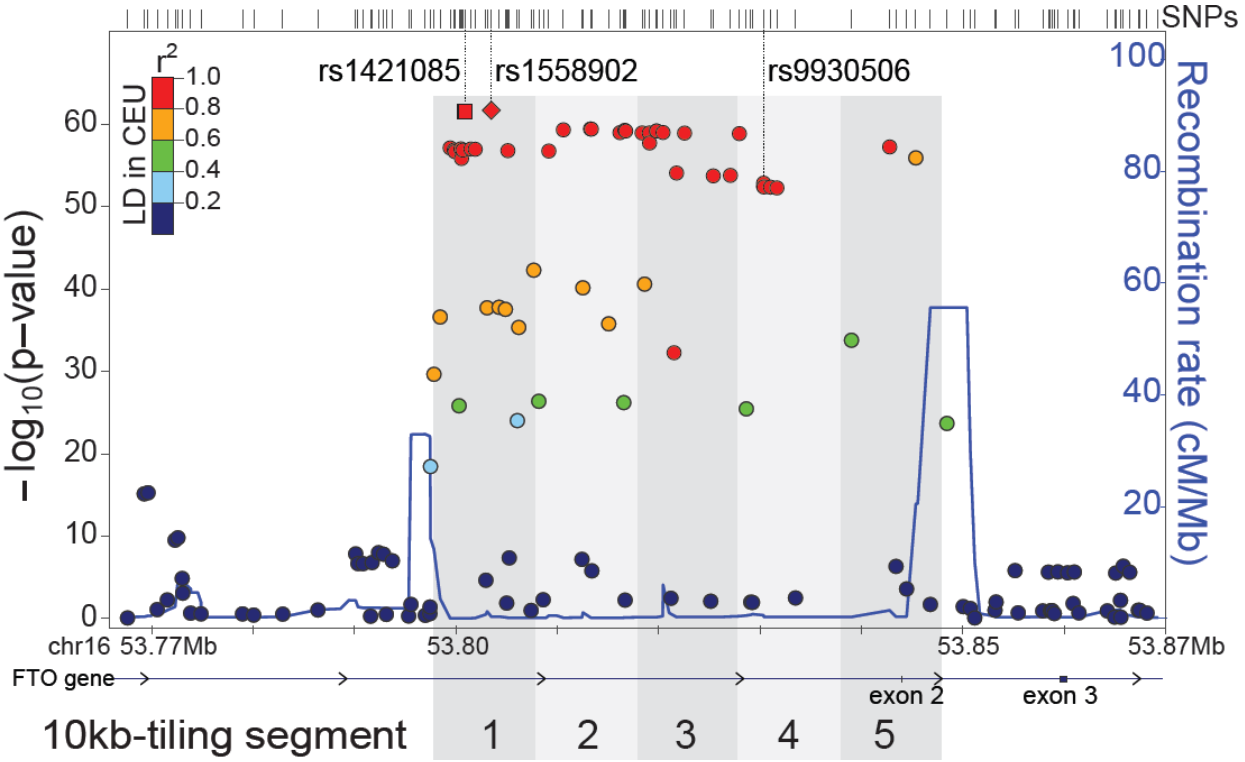
FTO Obesity Variant Circuitry and Adipocyte Browning in Humans

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.

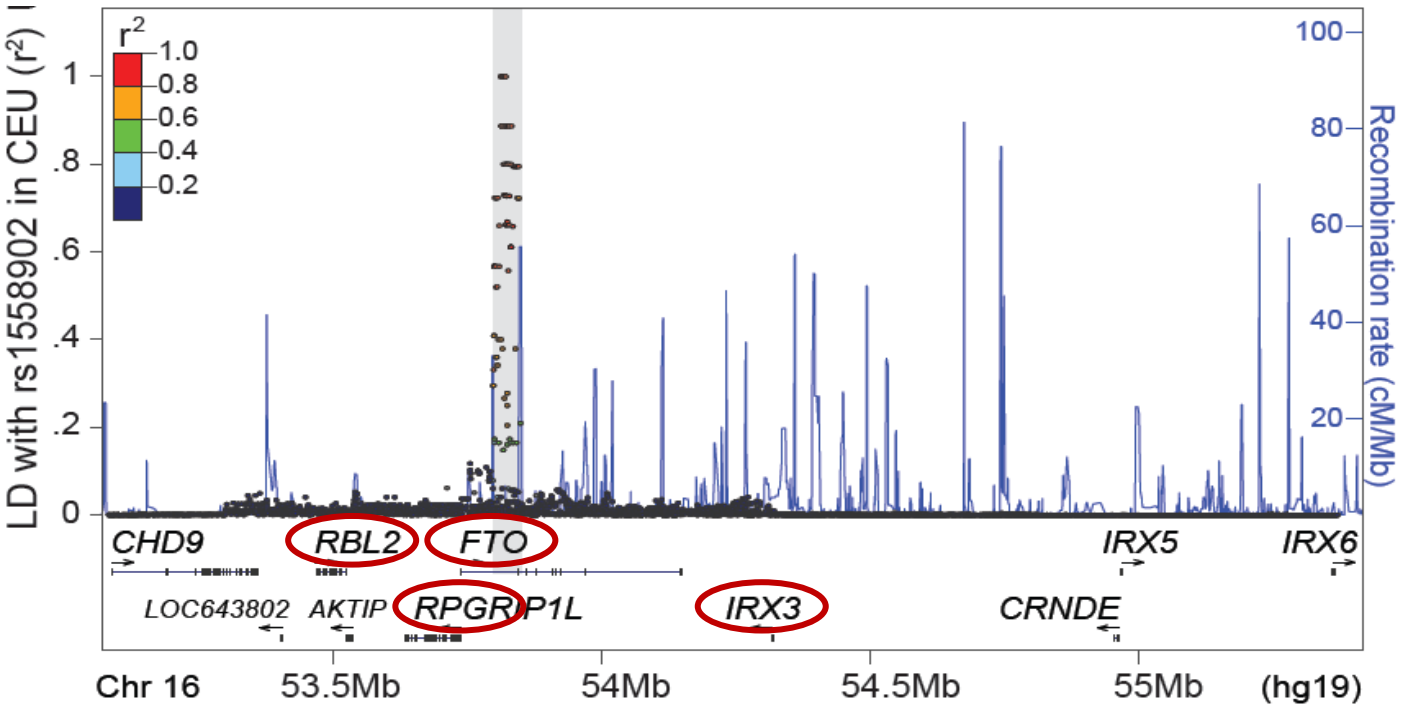
- Implications for obesity therapeutics
- Deep down, a model for dissecting GWAS

FTO region: strongest association with obesity



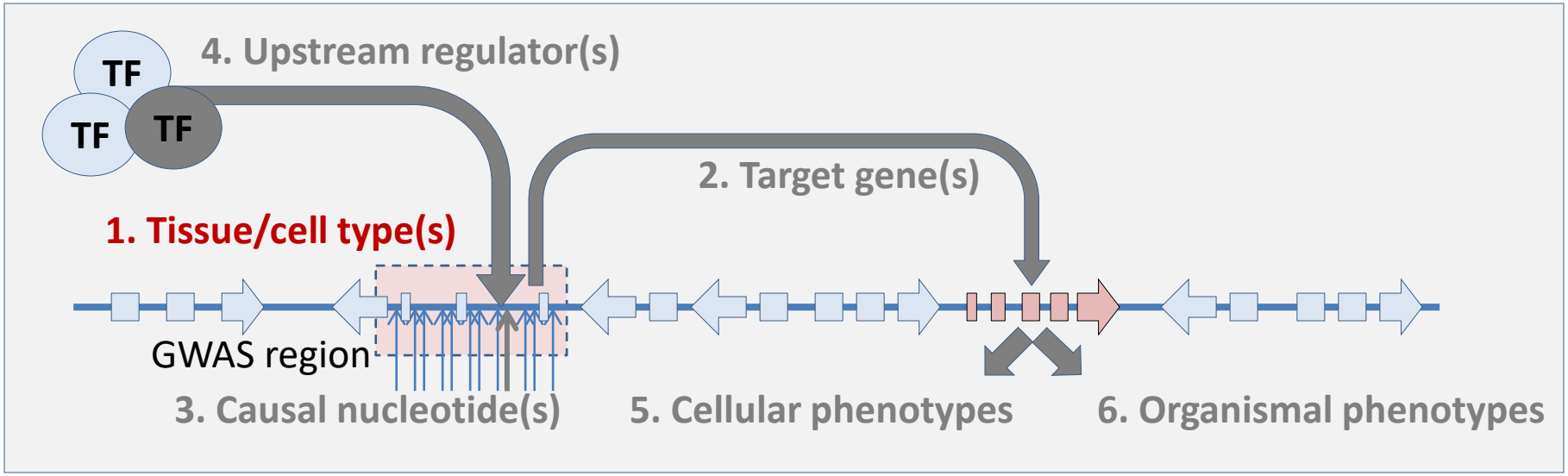
- Associated with **obesity**, Type 2 Diabetes, Cardiovascular traits
- 89 variants in LD, spanning 47kb, intron 1 of FTO gene
- No protein-altering variants: regulatory role? Target? Tissue?

Conflicting proposals of target gene, tissue



- Conflicting predictions: different targets/tissues/species:
  - **FTO** itself: Fischer Nature 09 (Overlap, Mouse **whole-body KO**)
  - **IRX3** in **pancreas**: Ragvin PNAS 10 (4C, Zebrafish KO)
  - **RBL2** in **lymphocytes**: Jowett Diabetes 2010 (Expression levels, eQTL)
  - **RPRGRIP1L** in **brain**: Stratigopoulos JBC 2014 (Leptin signaling, CUX binding)
  - **IRX3** in **brain**: Smemo Nature 14 (4C, Mouse brain DN)

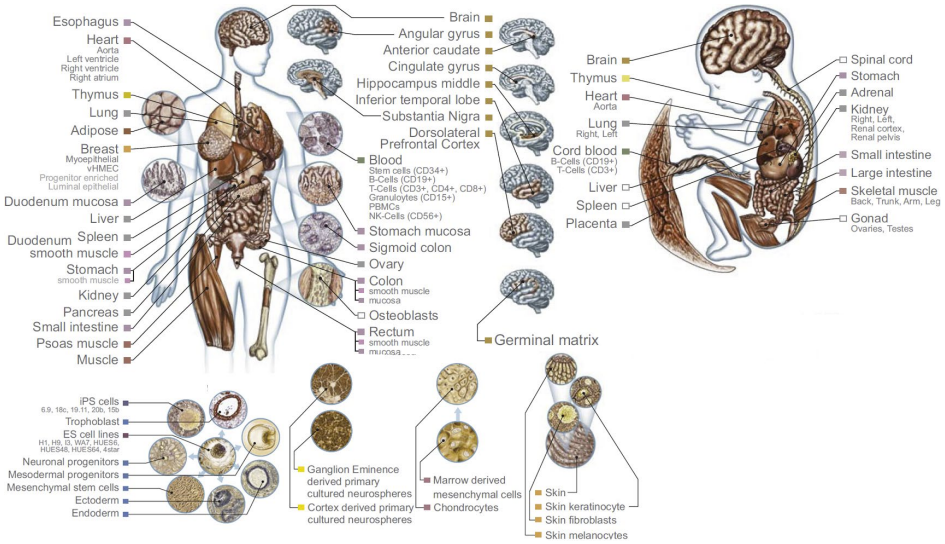
1. Establish relevant tissue/cell type



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

Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



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

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

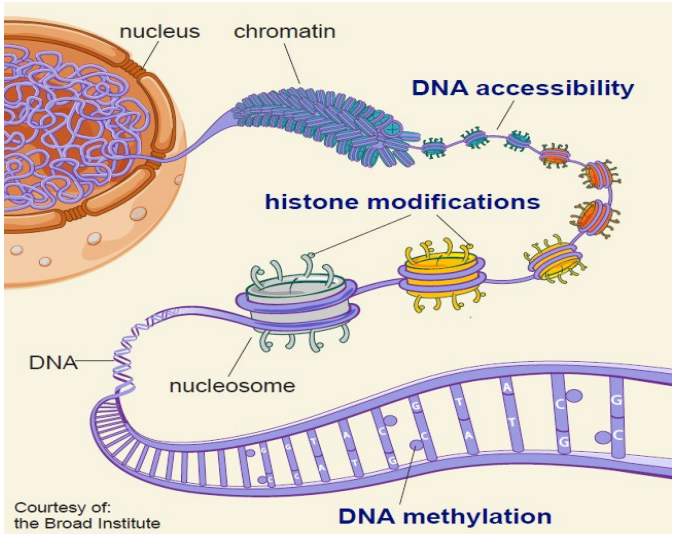
ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

X

Diverse epigenomic assays



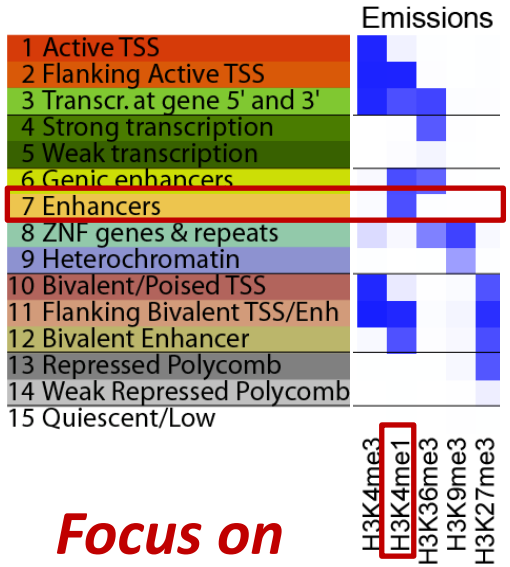
Histone modifications
- **H3K4me3, H3K4me1, H3K36me3**
- **H3K27me3, H3K9me3, H3K27/9ac**
- **+20 more**

Open chromatin:
- **DNA accessibility**

DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

Gene expression
- **RNA-seq, Exon Arrays**

Chromatin state annotations across 127 epigenomes

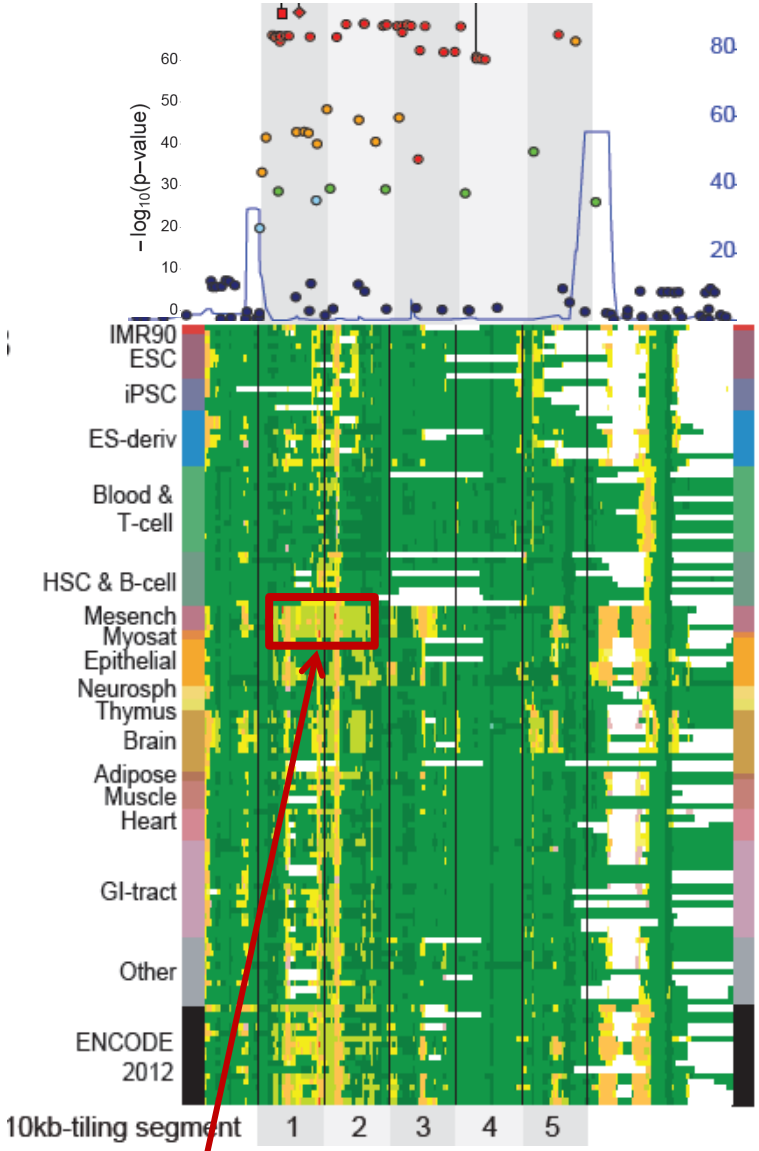


Focus on enhancer states

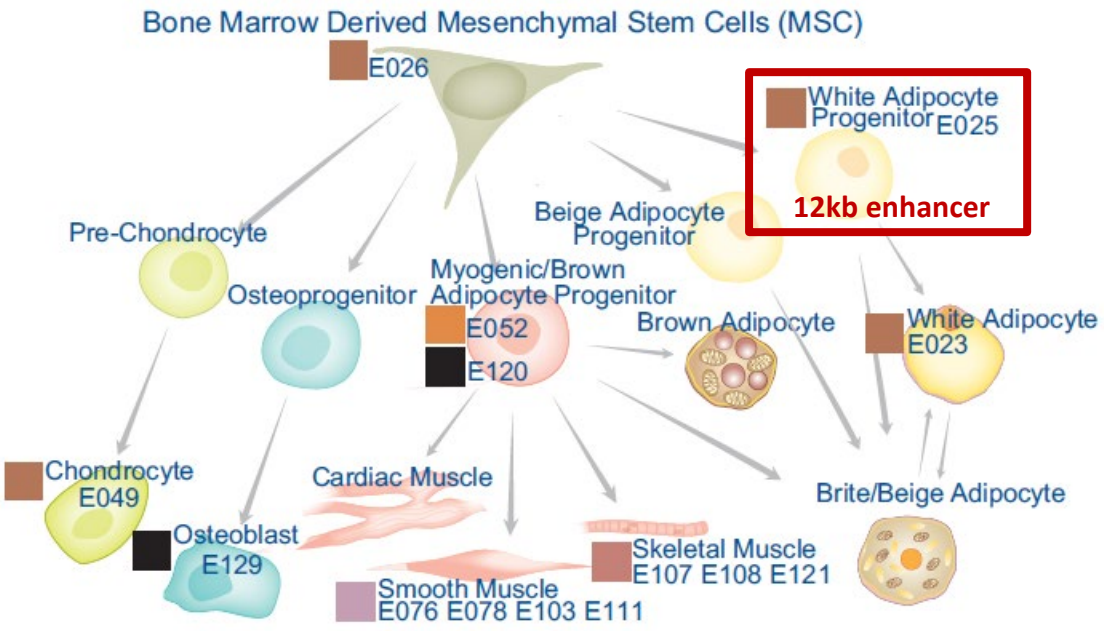
Roadmap Epigenomics, Nature 2015

Tissue-specific annotations of predicted enhancers, promoters, transcribed, repressed, quiescent regions

1. Tissue: Chromatin states predict adipocyte function

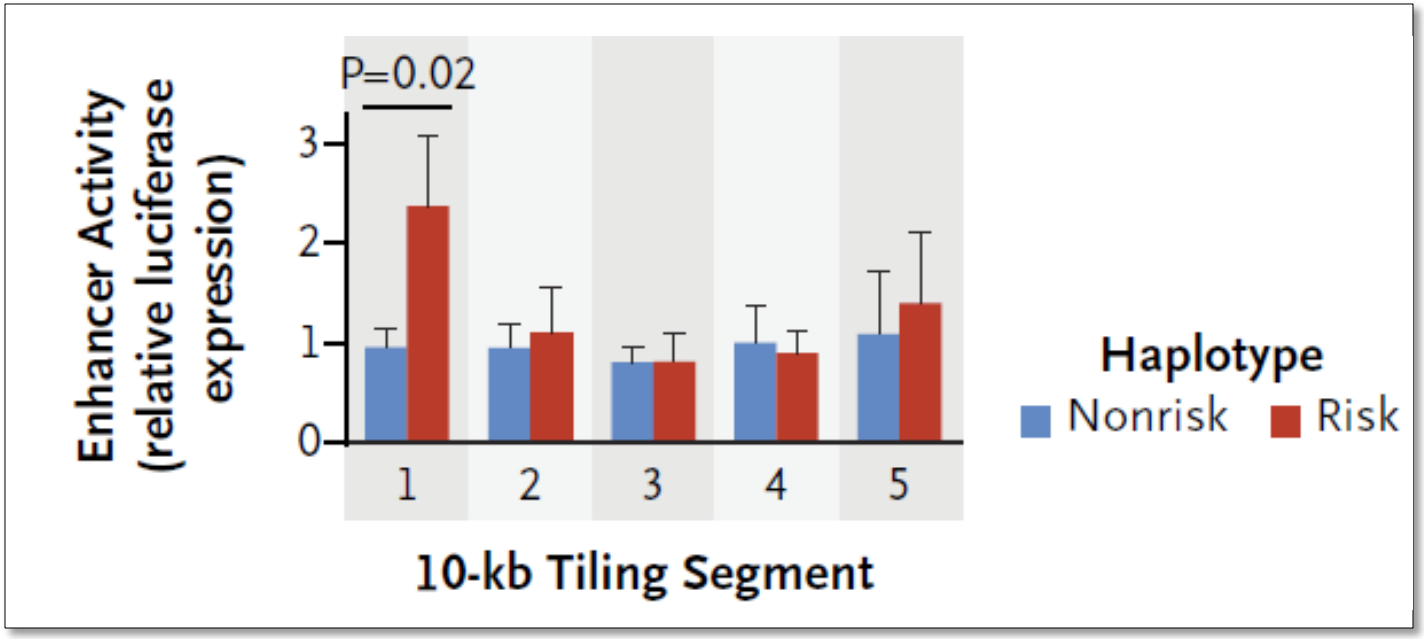


12 kb super-enhancer



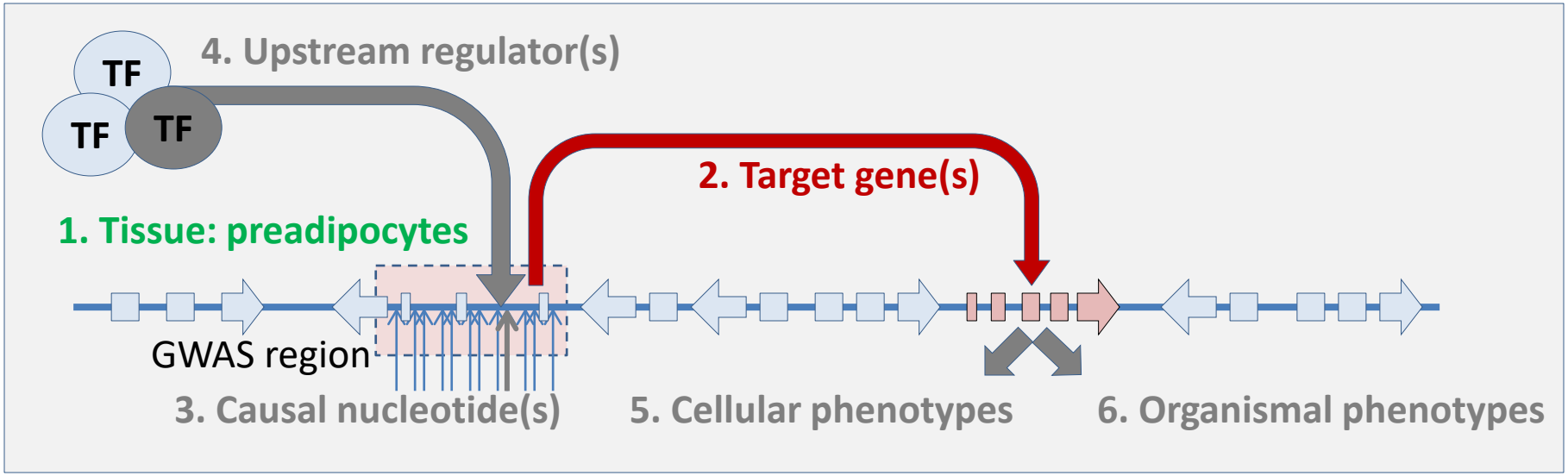
Epigenomic signatures point to progenitors of white/beige adipocytes

Enhancer tiling experiments confirm region, cell type



*Risk haplotype shows **increased** activity, gain-of-function*

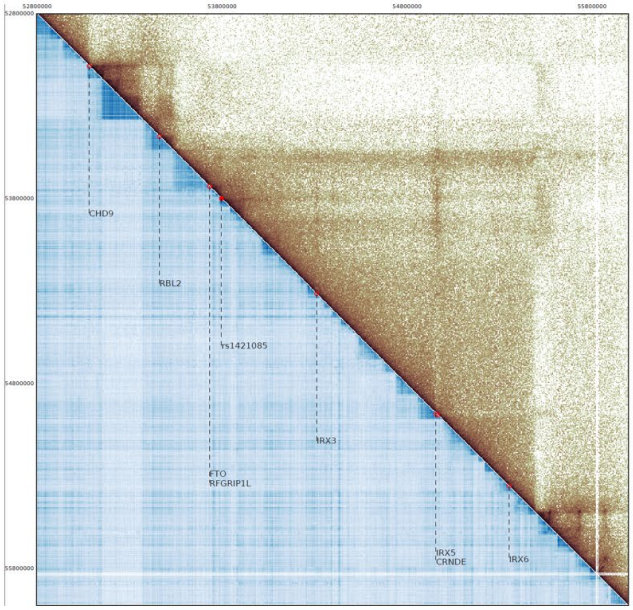
2. Establish downstream target genes



1. Establish relevant **tissue/cell type: pre-adipocytes**
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

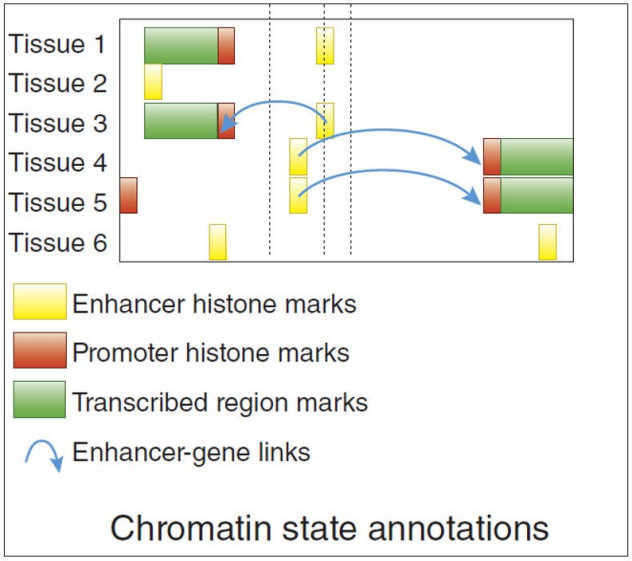
Link enhancers to their target genes 3 lines of evidence:

Physical



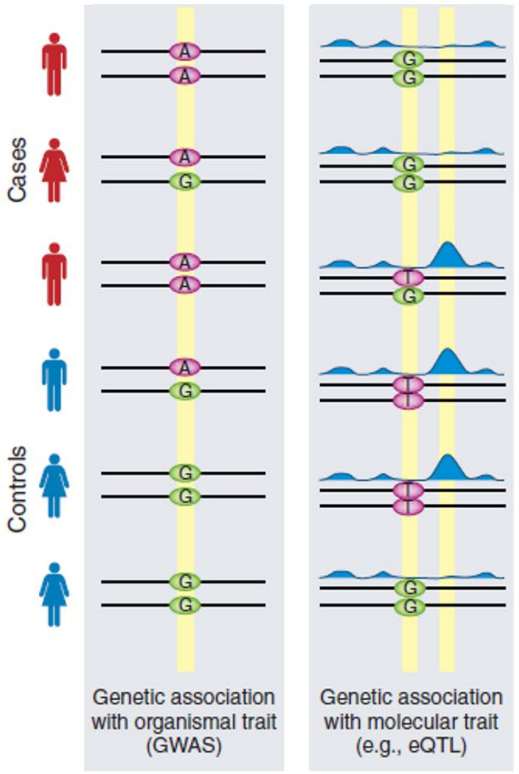
Hi-C: Physical proximity in 3D

Functional



Enhancer-gene activity correlation

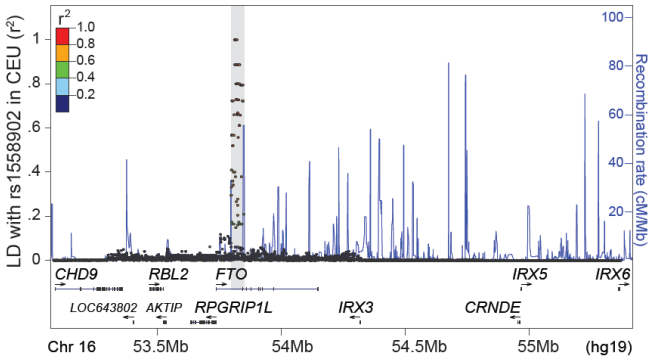
Genetic

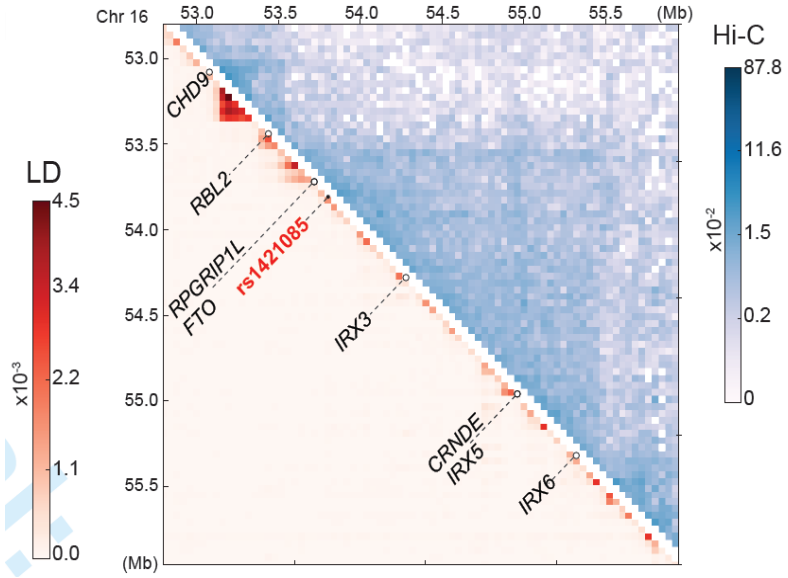


eQTL evidence: SNP effect on expression

Complementary evidence at physical, functional, genetic level

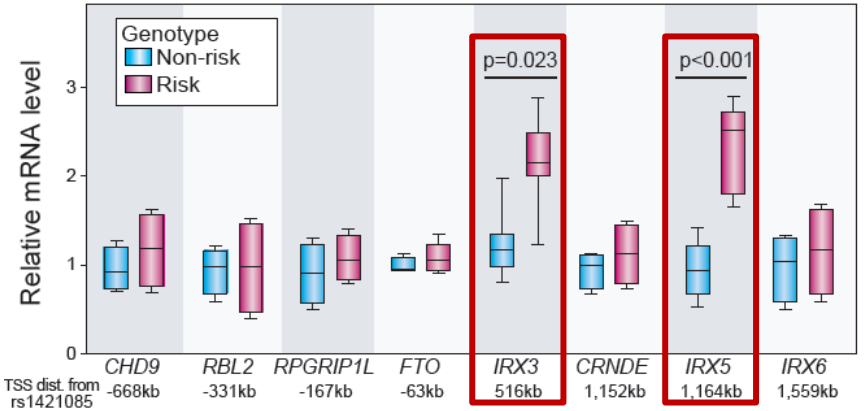
2. Targets: 3D folding and expr. genetics indicate IRX3+IRX5





Dixon, Nature 2012

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

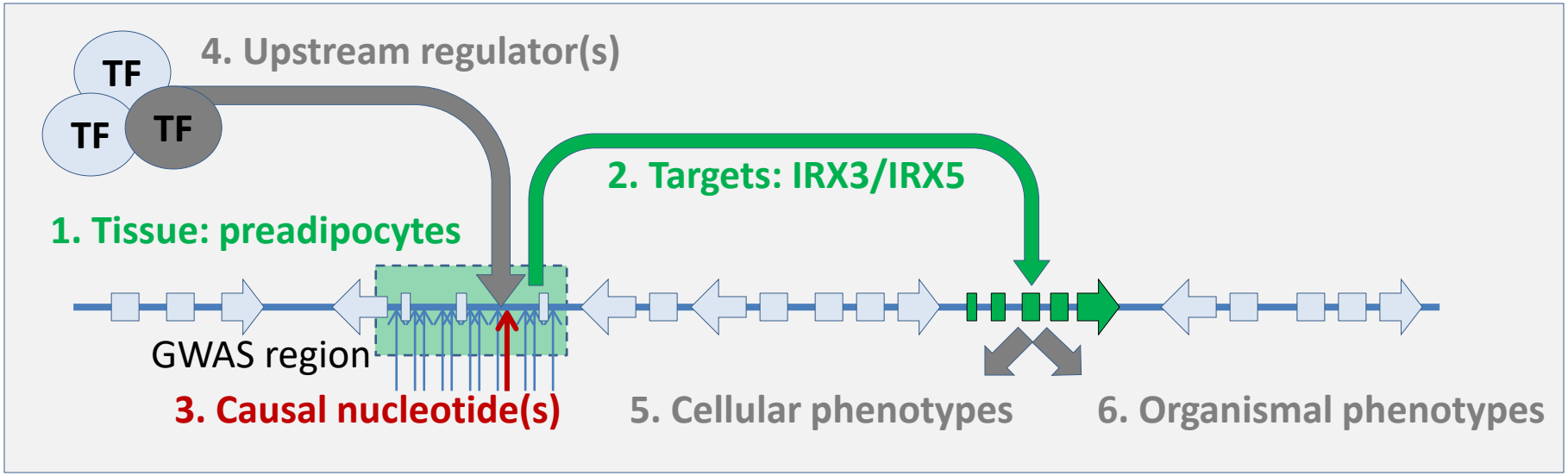


eQTL targets: IRX3 and IRX5

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

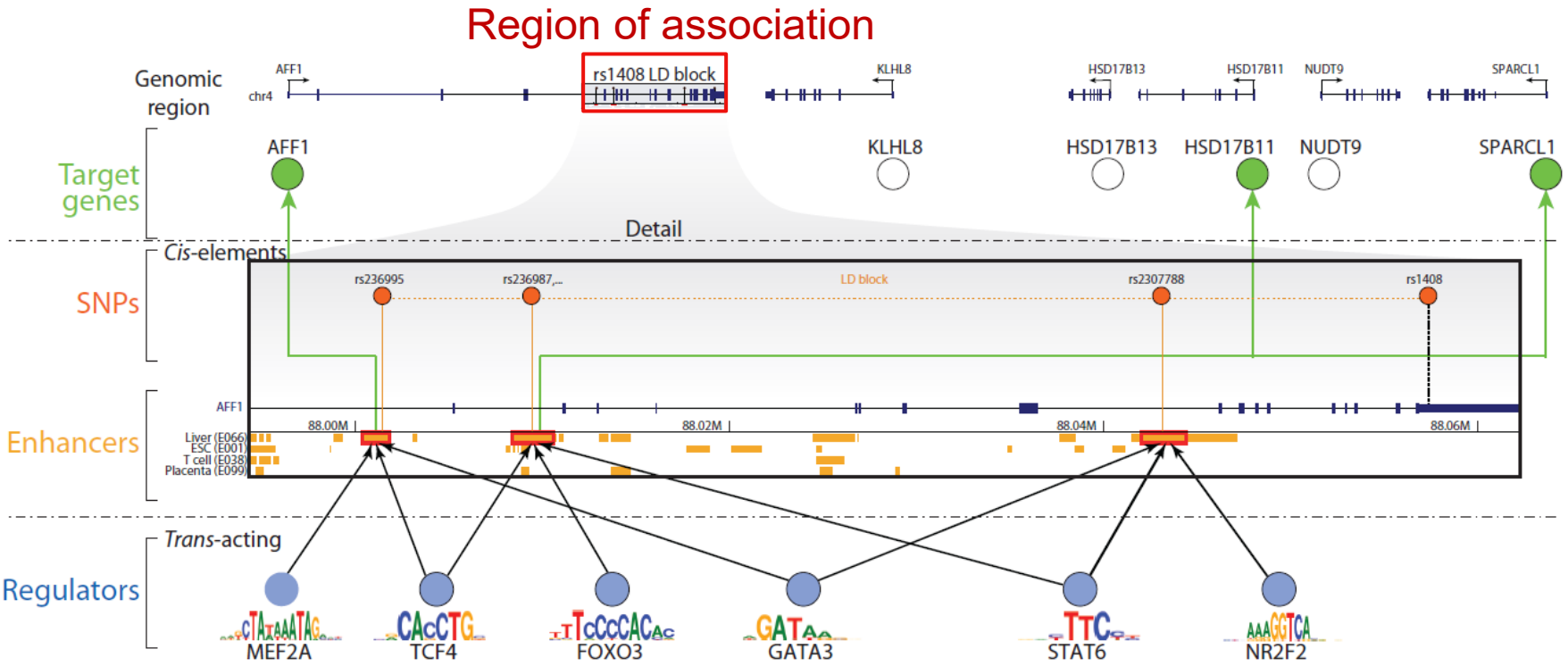
(1) Risk allele: increased expression
(gain-of-function by loss of repressor)
(2) Action in early adipocyte differentiation
(eQTL is not visible in whole-adipose tissue)

3. Establish causal variant



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

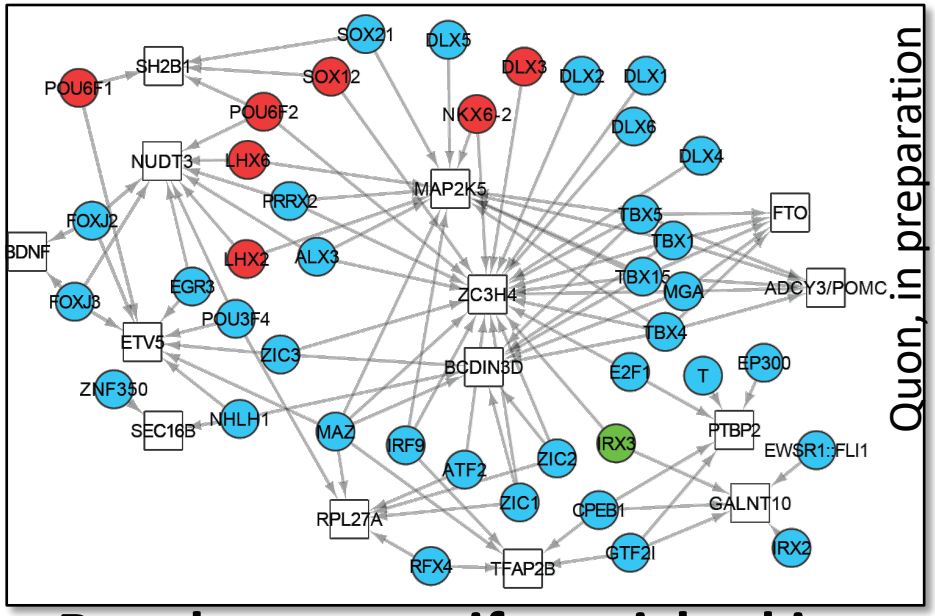
Non-coding circuitry helps interpret disease loci



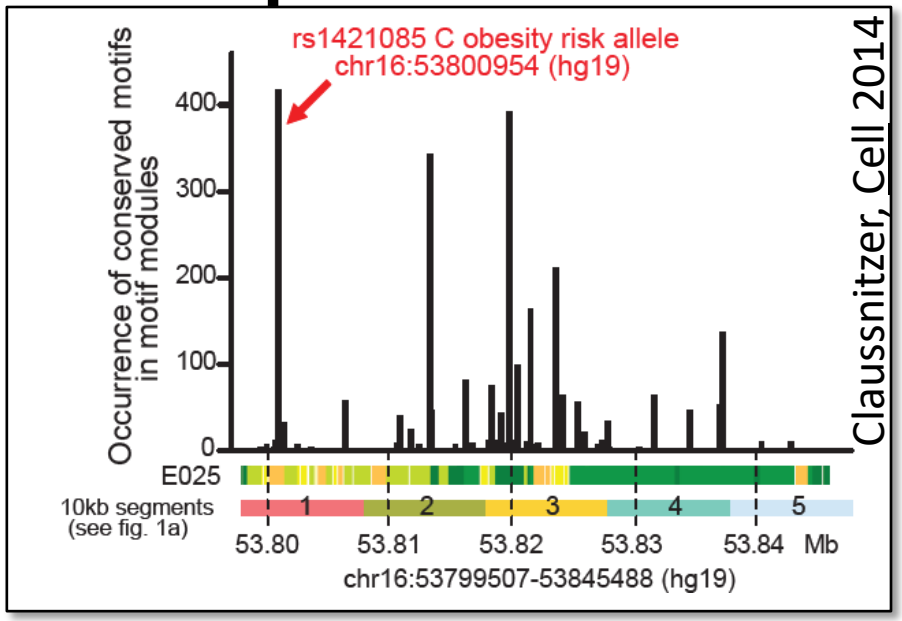
Study multiple GWAS loci to find commonalities/enrichments:

- Epigenomics: narrow down regulatory **regions**, relevant cell types
- Comparative genomics: prioritize **SNPs** over conserved nucleotides
- Regulatory genomics: match **motifs** to predict driver TFs/regulators
- Functional genomics: predict **target genes** in common pathways

Motif enrichment + conservation: predict causal SNP



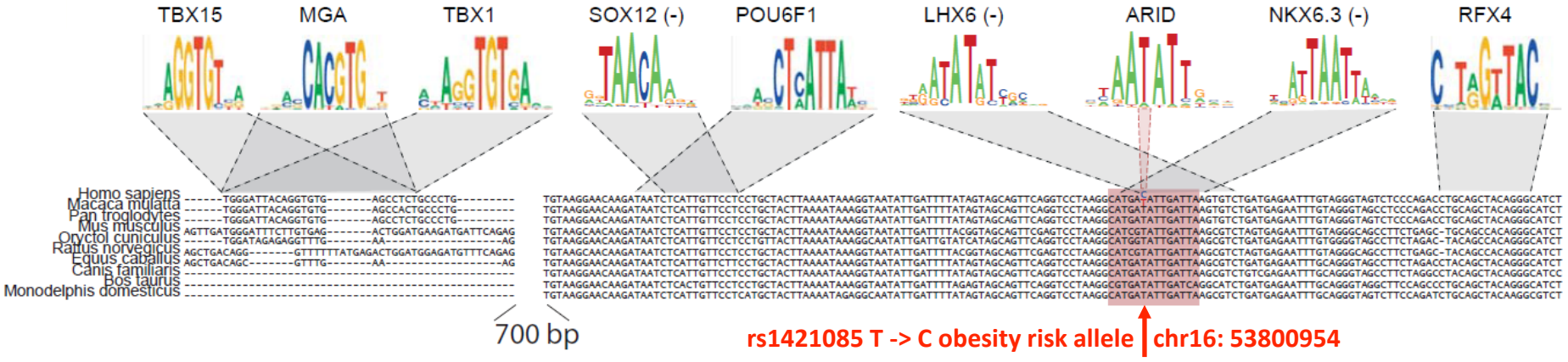
Quon, in preparation



Claussnitzer, Cell 2014

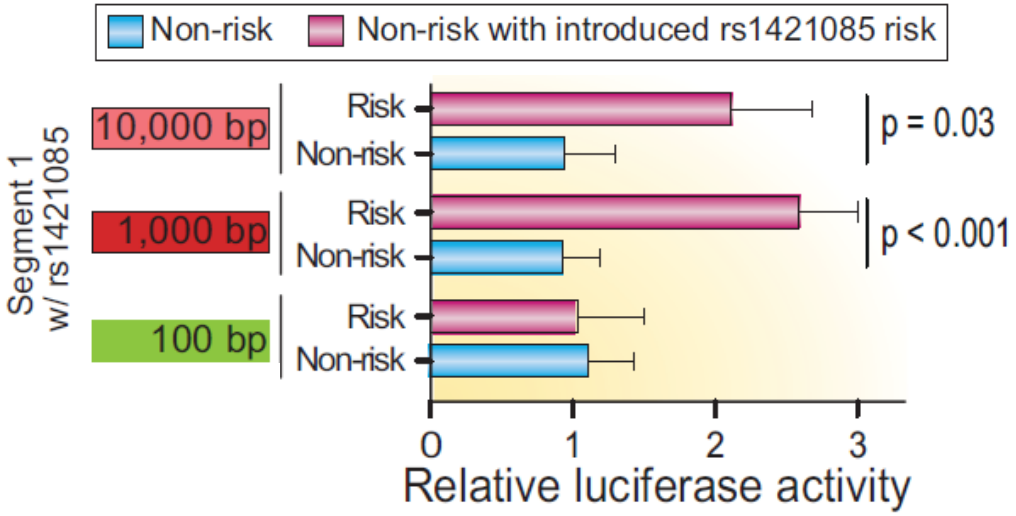
Regulatory motifs enriched in BMI GWAS hits

Regulatory motif combinations conserved across mammals



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

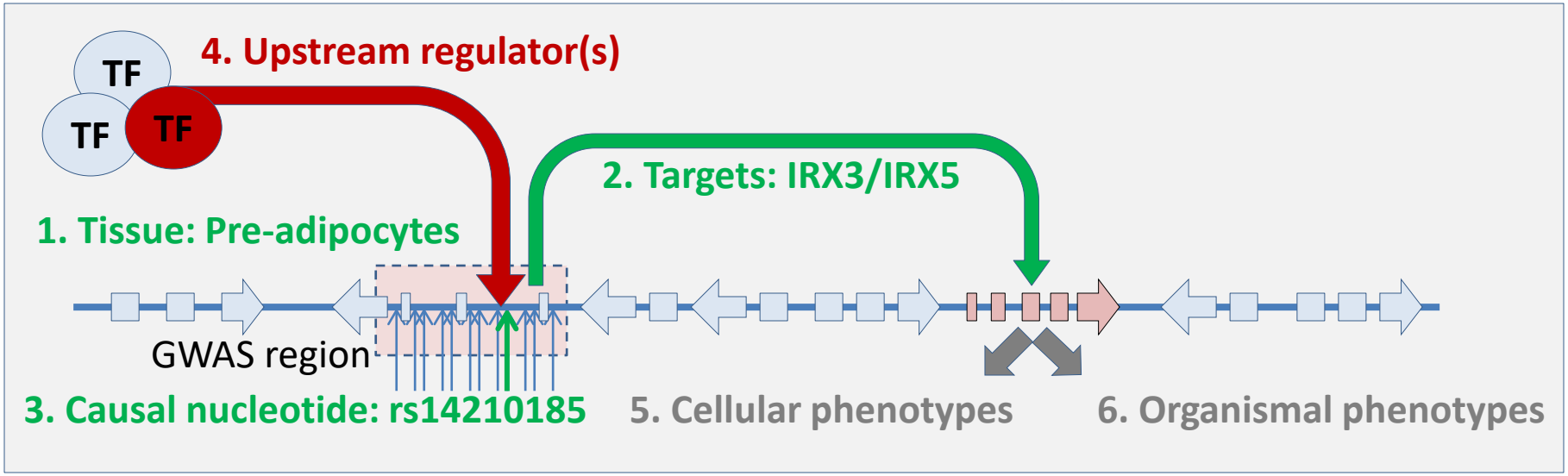
Single-nucleotide alteration alters enhancer activity



SNP recapitulates risk haplotype de-repression

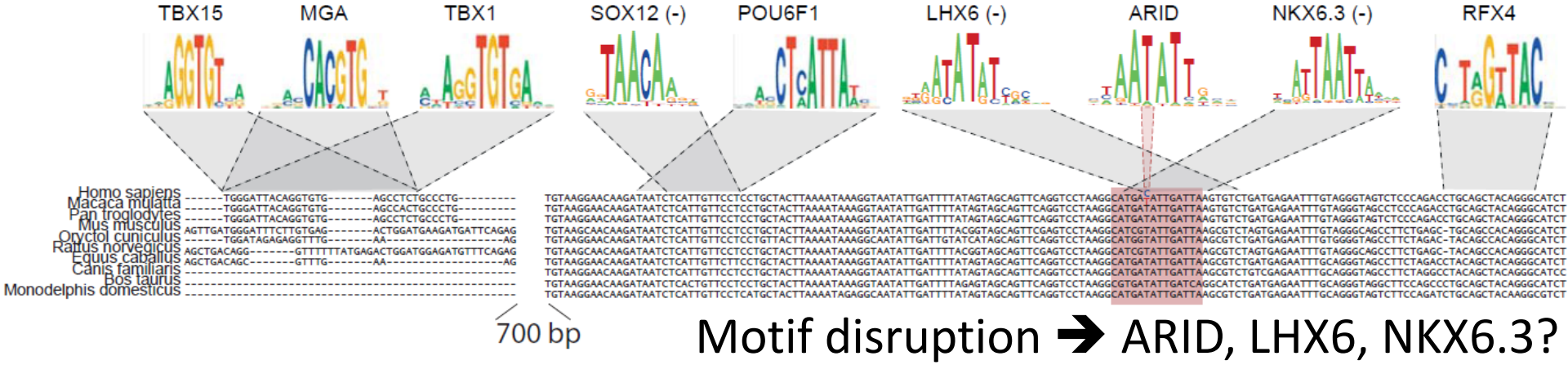
Acts at 10kb and 1kb (not at 100bp)

4. Establish upstream regulator

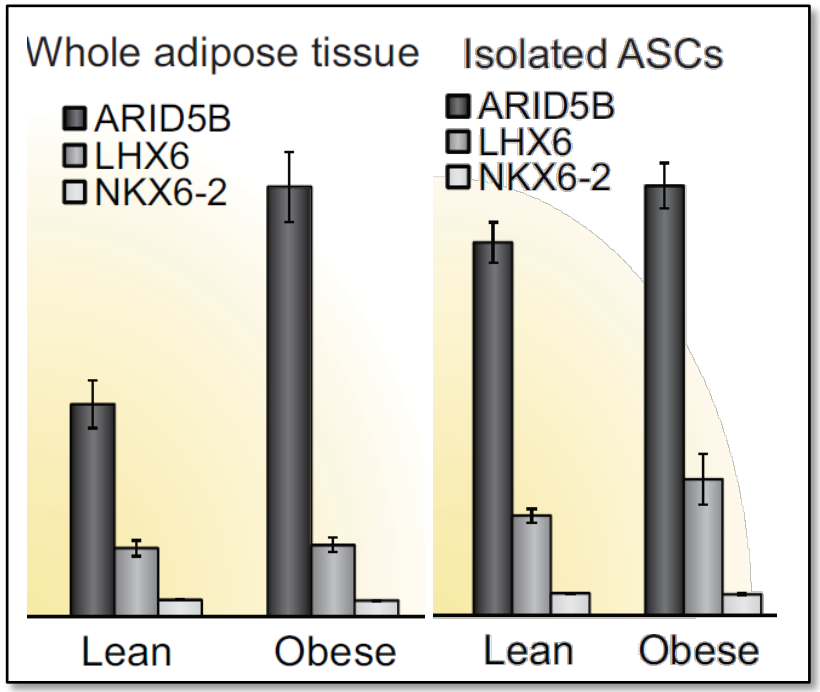


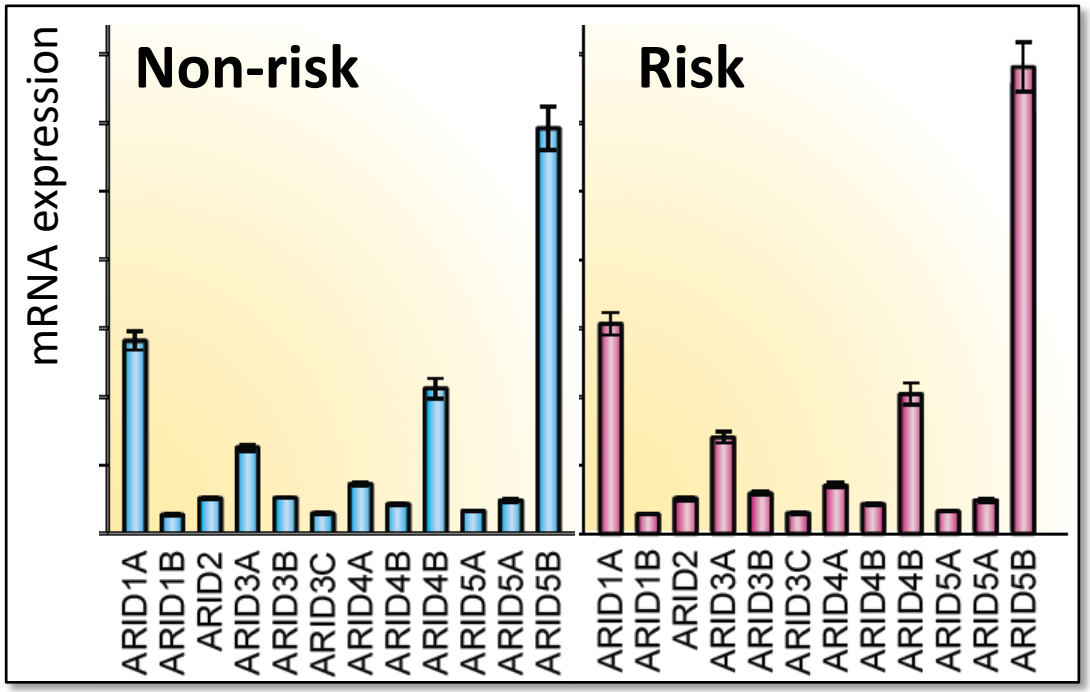
1. Establish relevant **tissue/cell type:** **pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant: **rs1421085**
4. Establish upstream **regulator** causality
5. Establish **cellular** phenotypic consequences
6. Establish **organismal**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

Regulator expression implicates ARID5B repressor



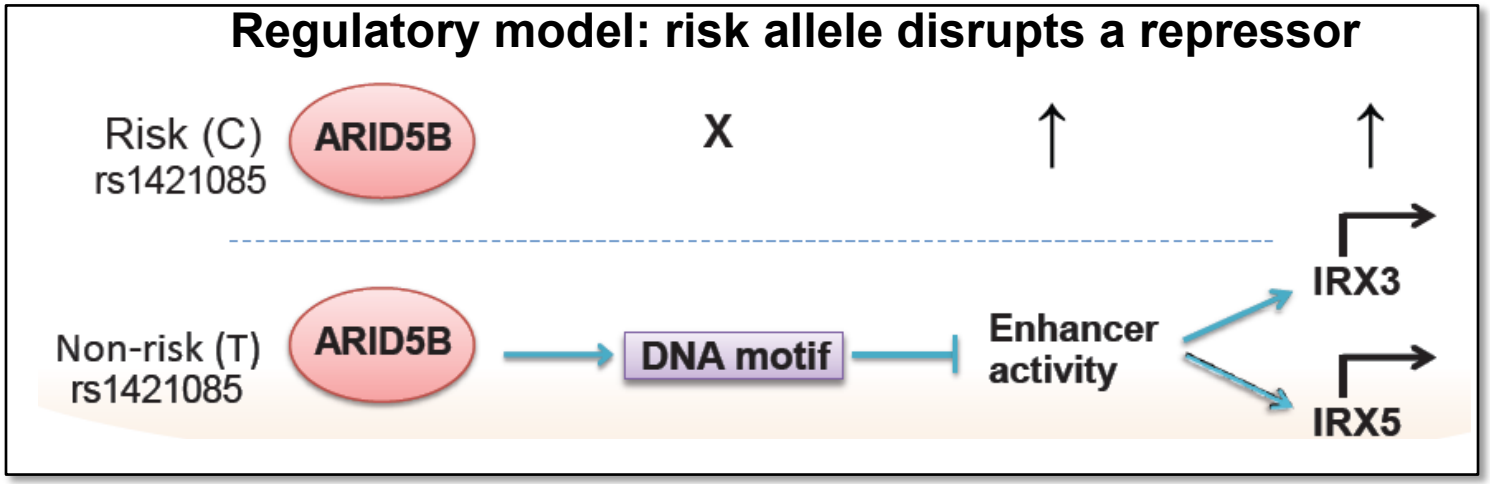
Motif disruption ➔ ARID, LHX6, NKX6.3?





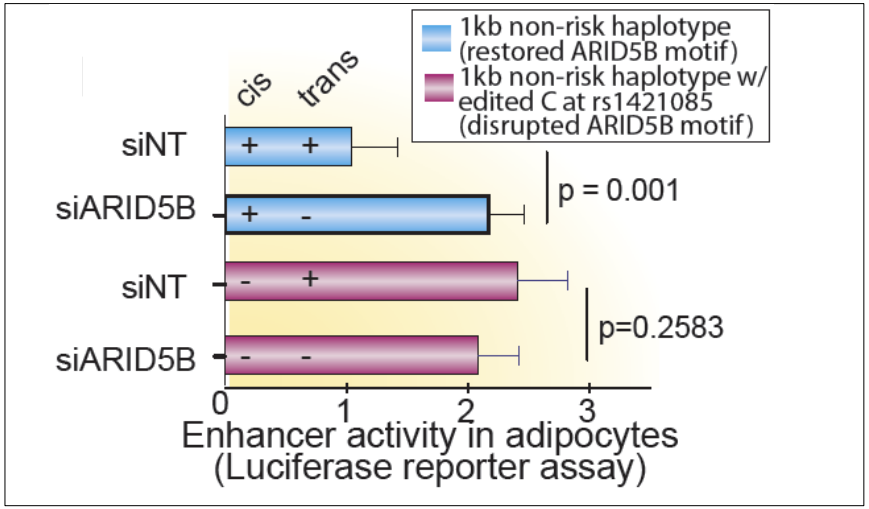
- Adipose/ASC expression suggests ARID family
- Highest expression implicates ARID5B upstream regulator

Causality and epistasis of ARID5B repression

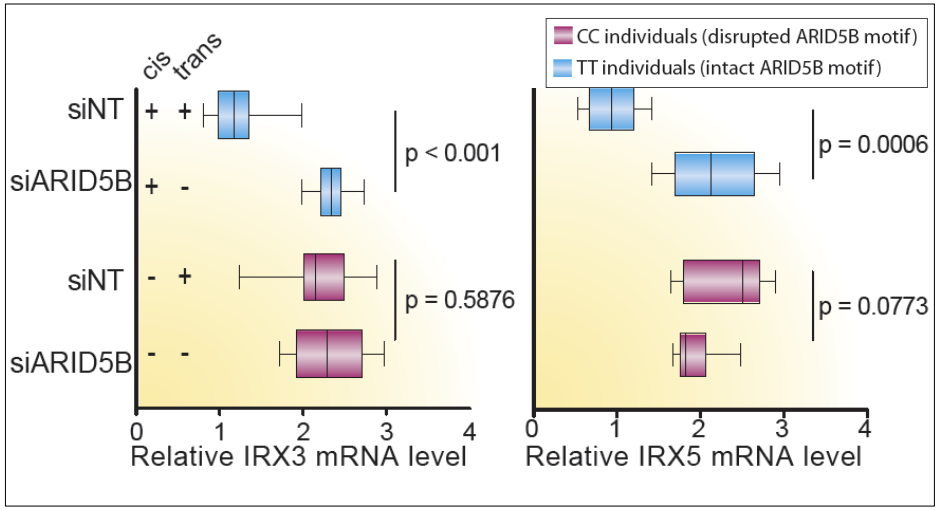


Cis/trans conditional analysis

Enhancer activity

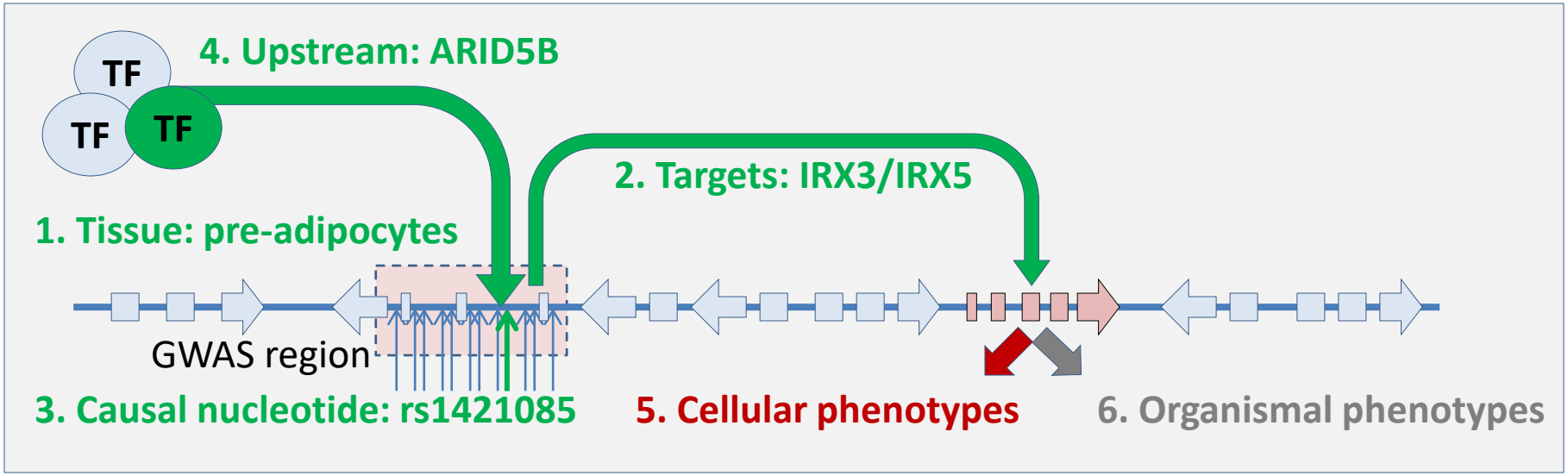


IRX3/5 expression

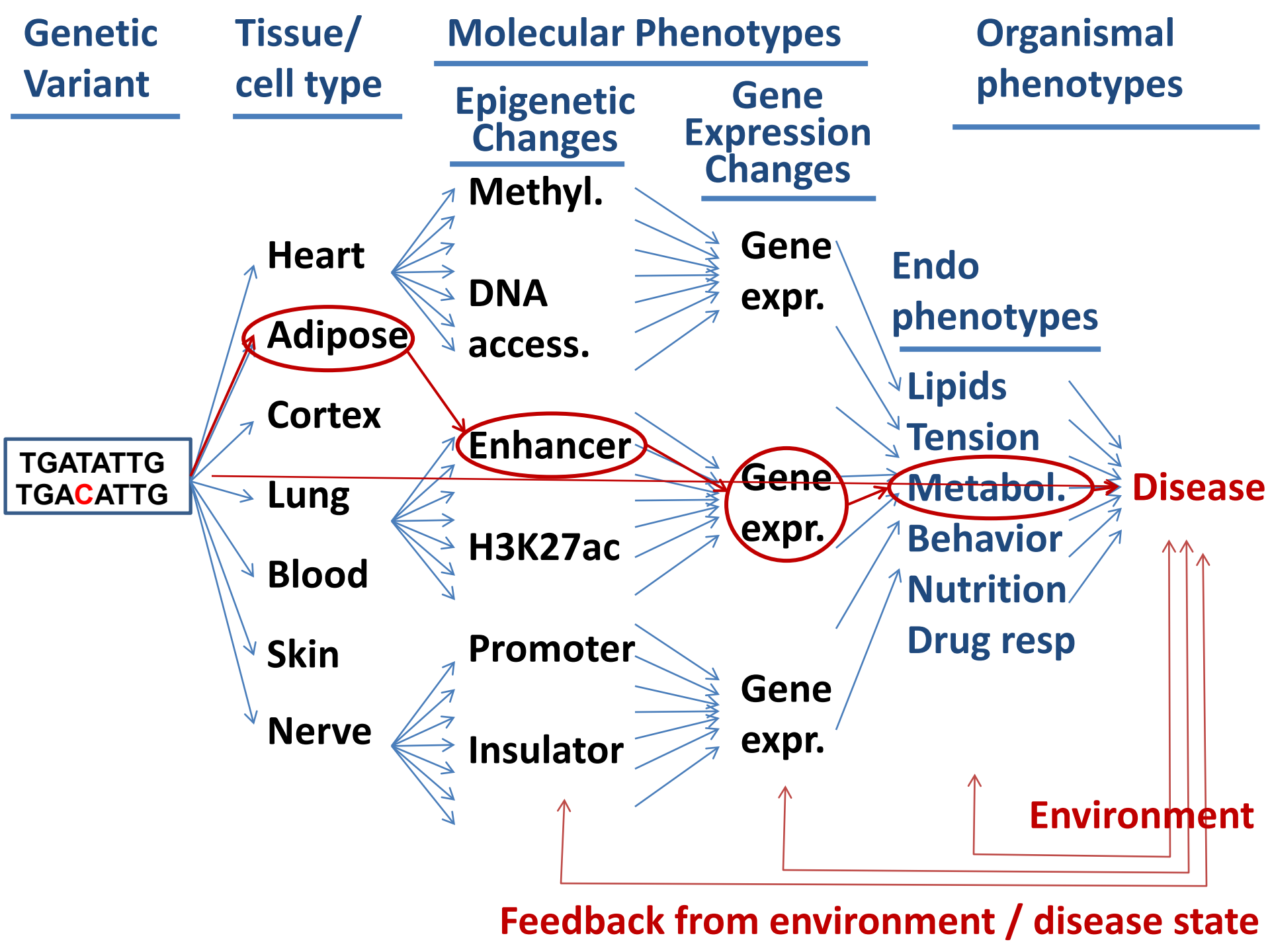


- ***Repression of enhancer, IRX3 and IRX5 all require both TF and motif***
- ***Disrupting motif (CC), or repressing ARID5B (siRNA) ➔ de-repression***

5. Establish cellular phenotypic consequences

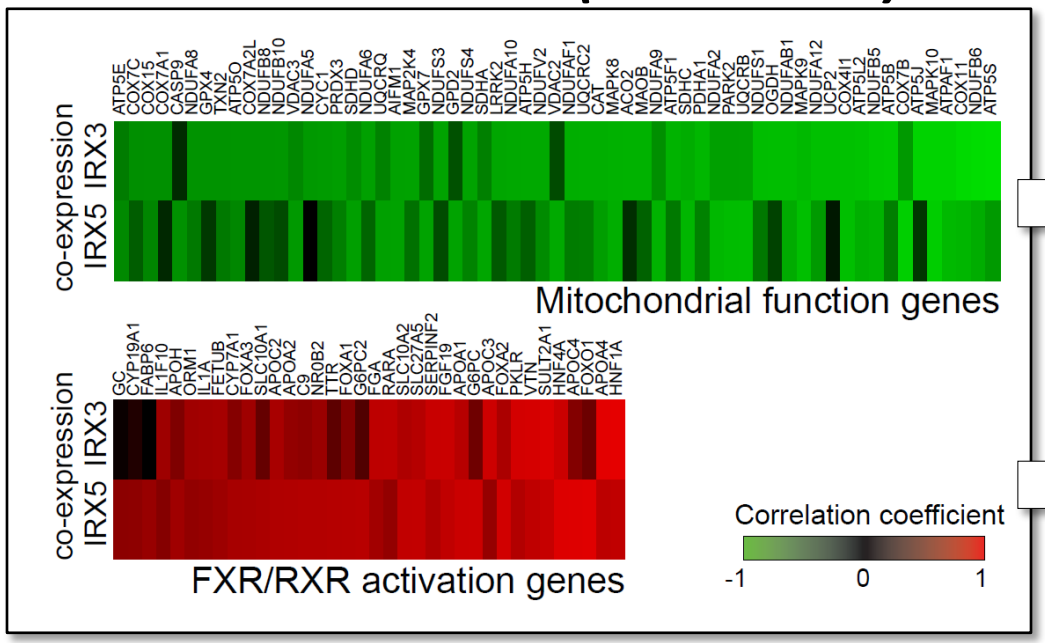


1. Establish relevant **tissue/cell type:** **pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant: **rs1421085**
4. Establish upstream **regulator** causality: **ARID5B**
5. Establish **cellular** phenotypic consequences
6. Establish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

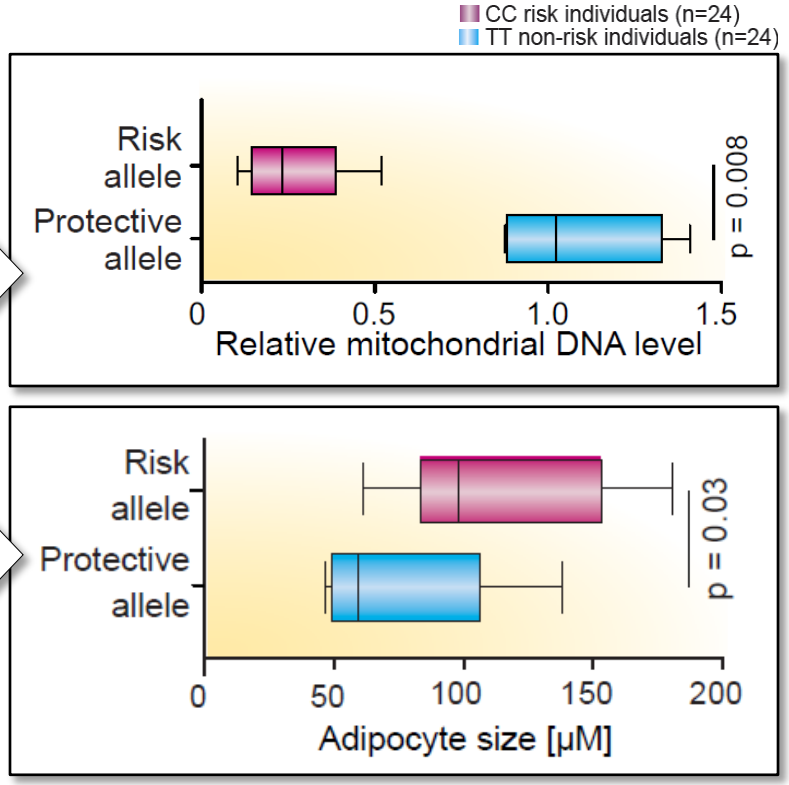

Genetic Variant
Tissue/ cell type
Molecular Phenotypes
Epigenetic Changes
Gene Expression Changes
Organismal phenotypes
TGATATTG
TGACATTG
Heart
Adipose
Cortex
Lung
Blood
Skin
Nerve
Methyl.
DNA access.
Enhancer
H3K27ac
Promoter
Insulator
Gene expr.
Gene expr.
Gene expr.
Endo phenotypes
Lipids
Tension
Metabol.
Behavior
Nutrition
Drug resp
Disease
Environment
Feedback from environment / disease state

Expression analysis to recognize target processes

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



Reflected in cellular phenotypes

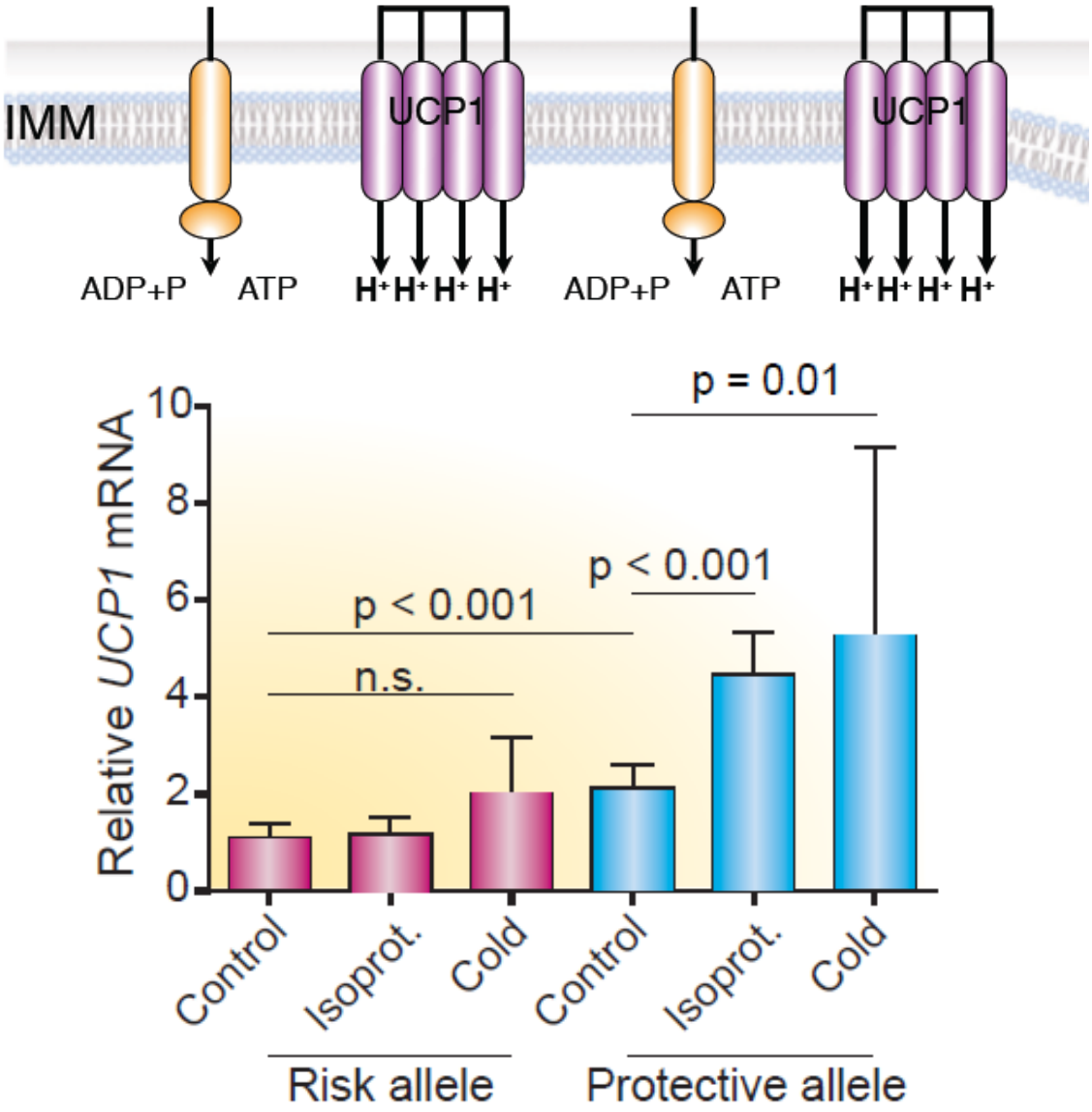


Negative correlation: mitochondria
Positive correlation: lipid storage

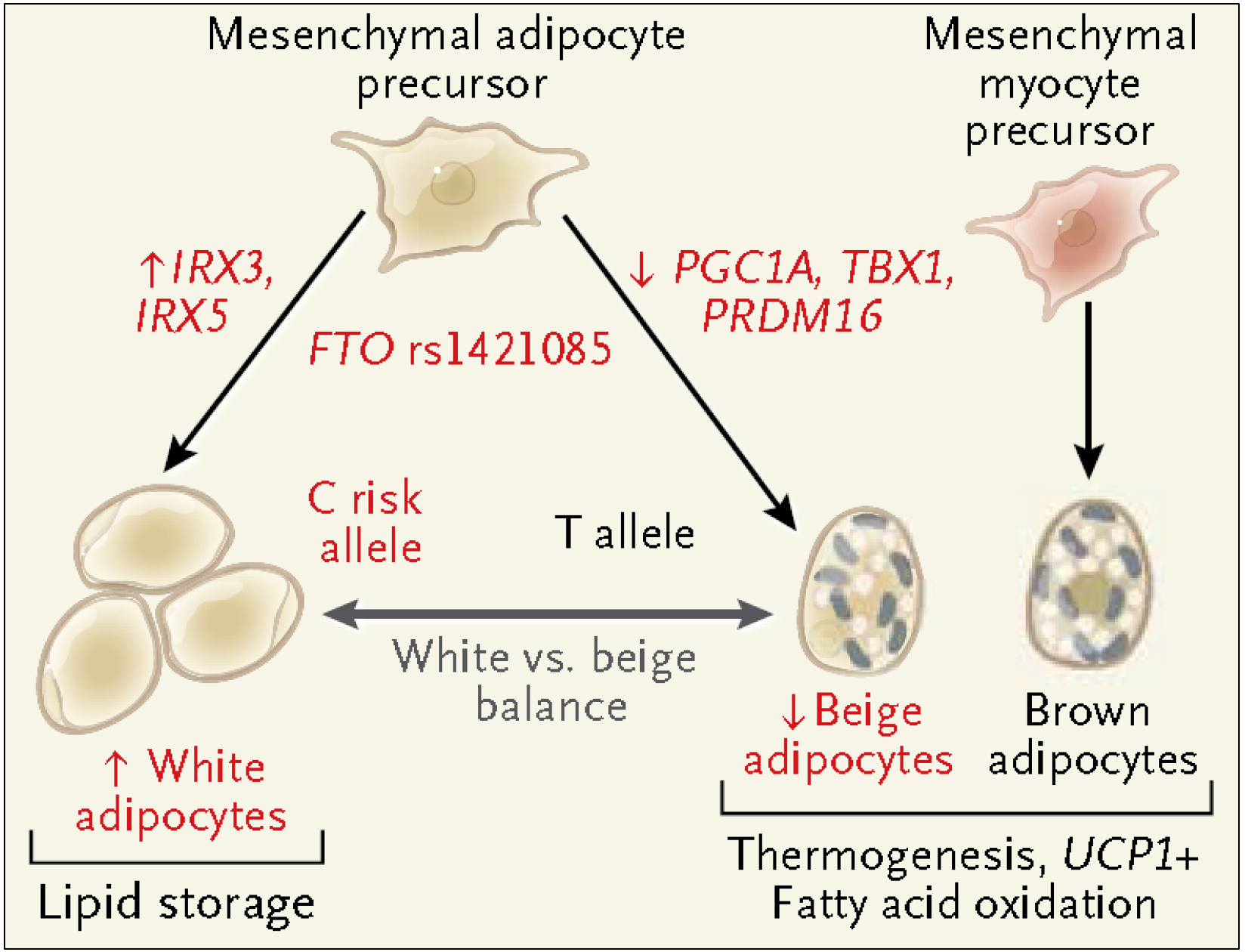
Risk carriers: increased mito
Non-risk: increased adipocytes

Risk allele: shift from dissipation to storage

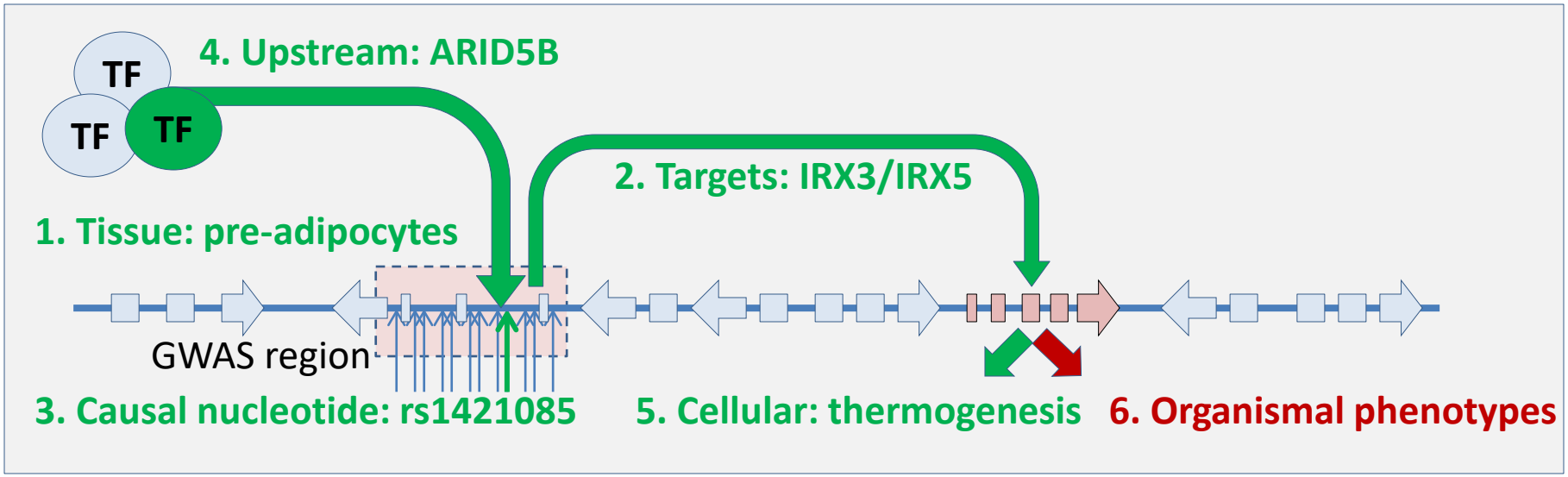
Risk individuals show disrupted thermogenesis



Mechanistic cellular model for FTO obesity locus

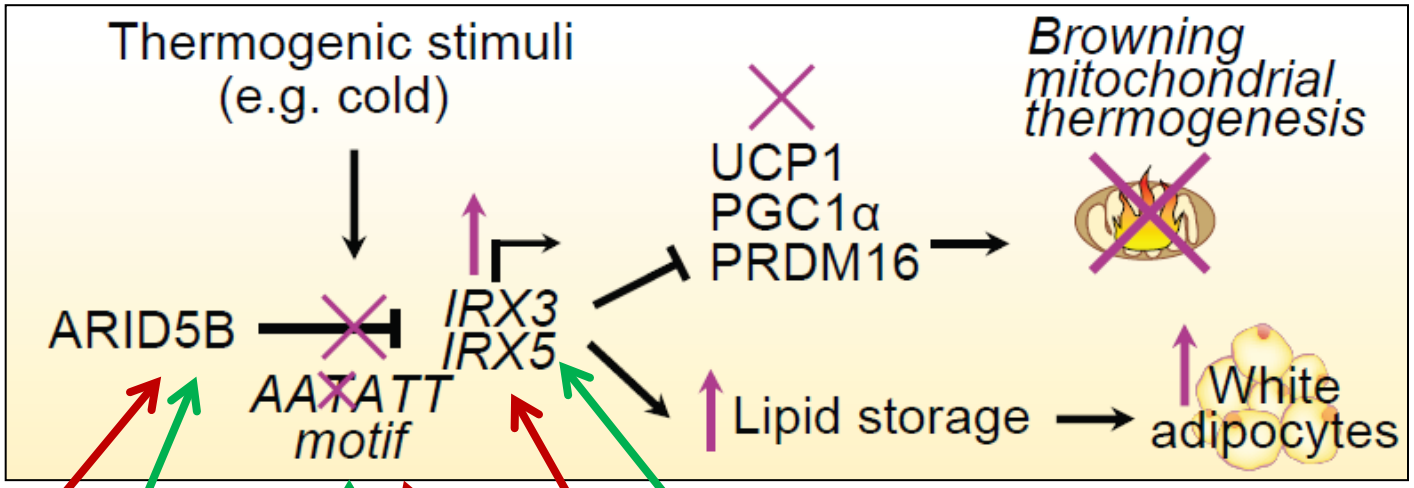


6. Manipulate circuitry to impact organism level



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

Dissected circuitry: entry points for intervention



ARID5B KD (obesity)

ARID5B OE (anti-obesity)

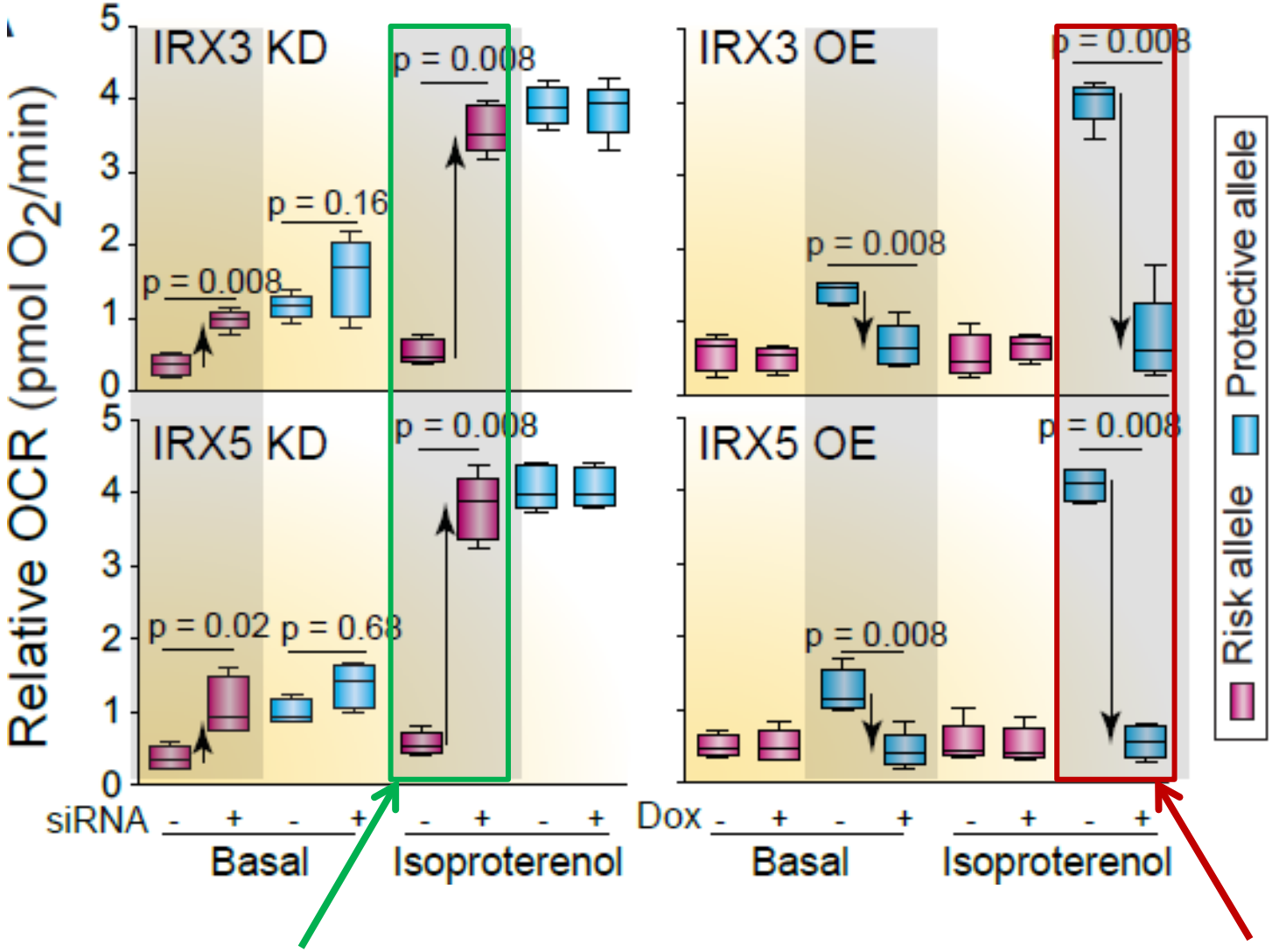
C-to-T motif rescue (anti-obesity phenotypes)

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

IRX3, IRX5 overexpression (pro-obesity phenotypes)

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

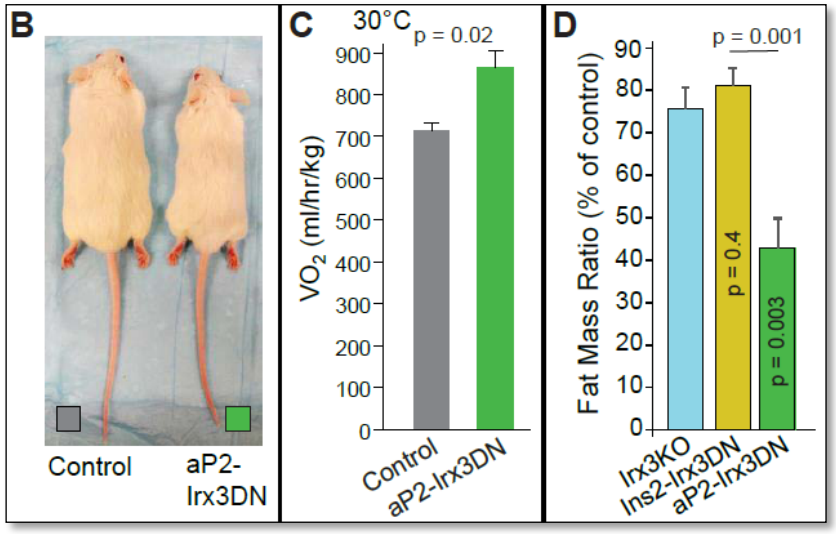
IRX3+IRX5 expression impacts energy utilization



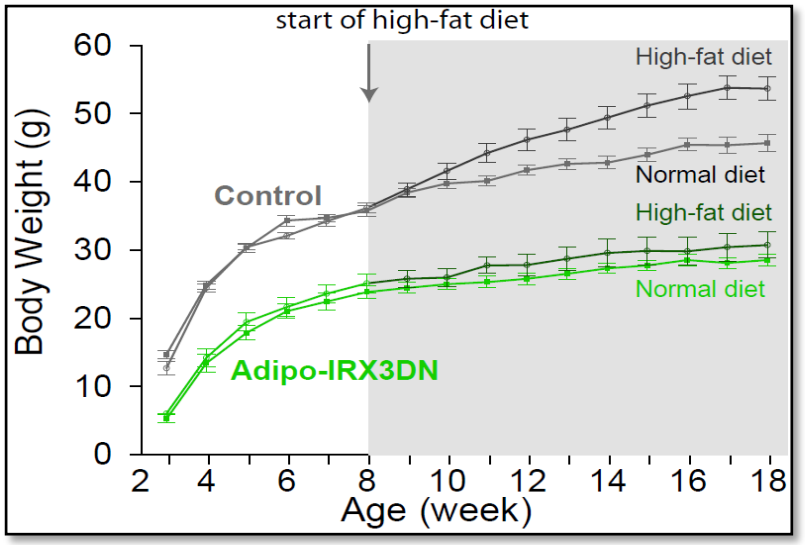
***Risk individuals*: IRX3/5 repression restores respiration,thermogenesis**

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

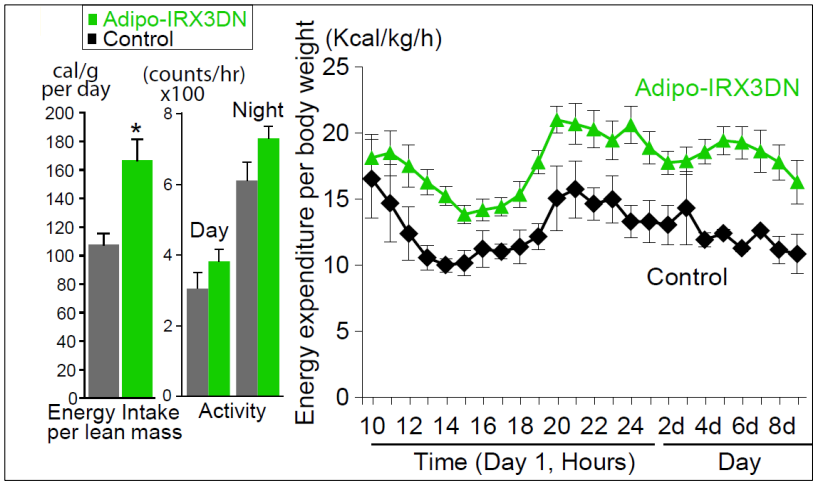
Irx3 adipose repression: anti-obesity phenotypes in mice



54% reduced body weight



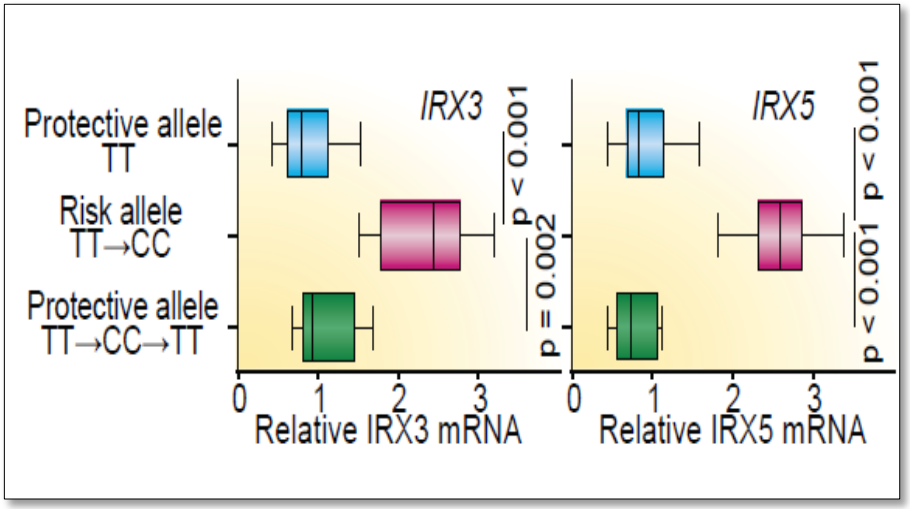
Resistance to high-fat diet



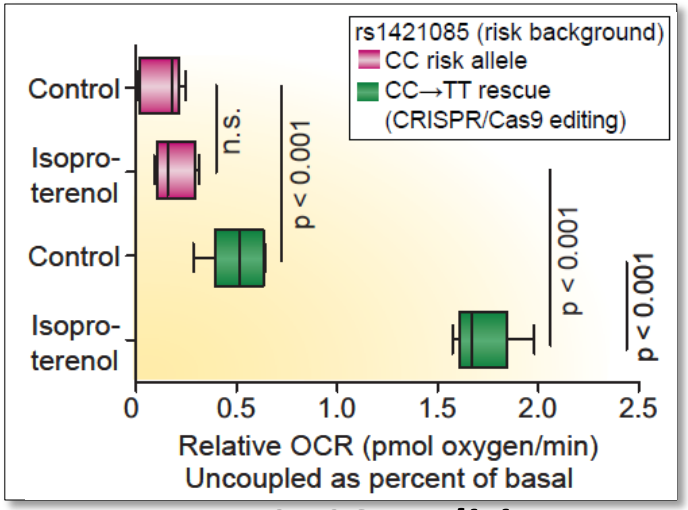
Increased energy dissipation

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

Single-nucleotide editing reverses thermogenesis in humans



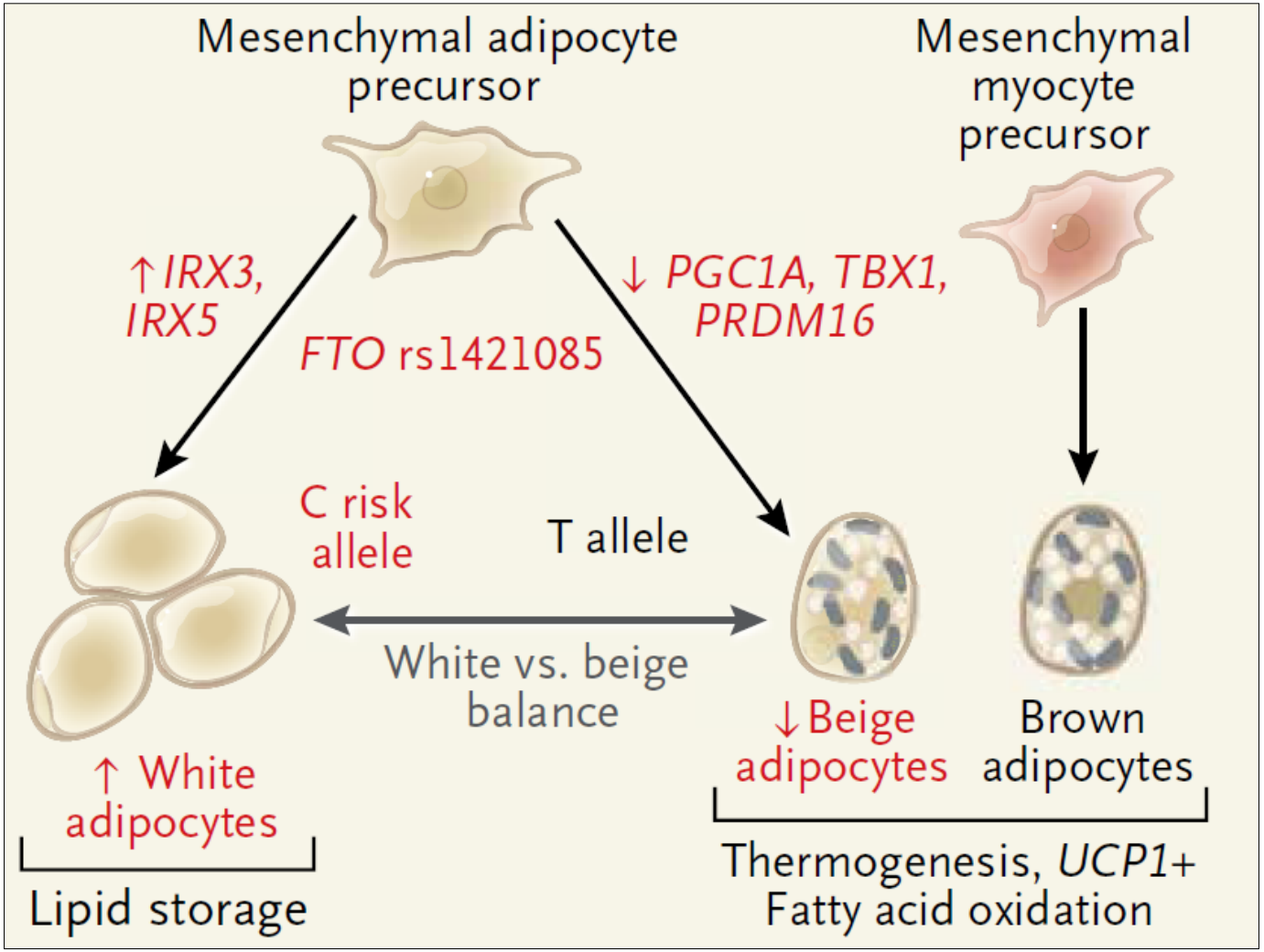
rs1421085 editing alters IRX3+IRX5 expression (500,000 and 1 million nucleotides away!)



rs1421085 editing restores thermogenesis

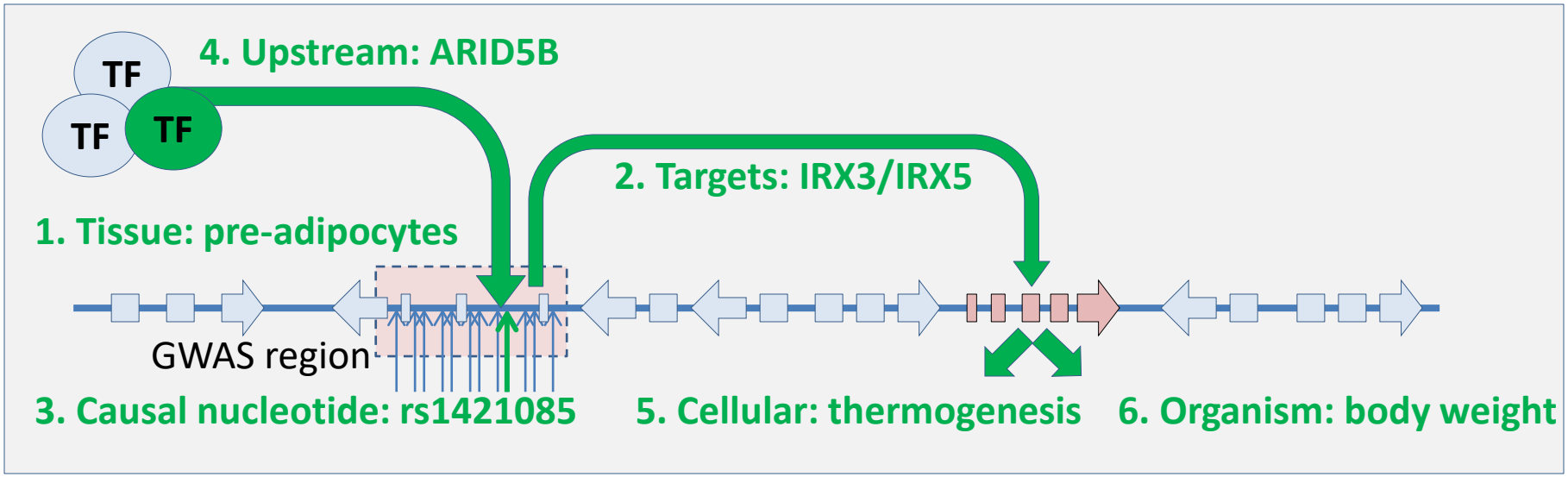
rs1421085 causality: C-to-T editing rescues IRX3/IRX5 expression, thermogenesis

Model: beige ⇔ white adipocyte development



Expand therapeutic focus from brain to adipocytes

FTO obesity locus as a model for dissecting GWAS



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**

Scaling up dissection efforts to 100s of variants

	LeadSNP	NumSNPs	NumExonS	MAF	pval	OddRatio	Study
	rs9268839	1	0	45%	1.E-250	2.28	RheumatoidArthritis_24390342
	rs1260326	8	1	39%	2.E-239	8.70	Cholesterol_24097068
	rs12143842	7	0	24%	1.E-213	3.50	QT_24952745
	rs1532085	9	0	40%	1.E-188	9.35	Cholesterol_24097068
	rs1367117	3	2	32%	1.E-182	8.40	Cholesterol_24097068
	rs629301	11	5	24%	2.E-170	7.46	Cholesterol_24097068
	rs2981579	11	0	40%	2.E-170	1.27	BreastCancer_23535729
	rs2476601	2	1	9%	9.E-170	1.80	RheumatoidArthritis_24390342
	rs11209026	27	1	7%	8.E-161	2.01	CrohnIBDUC_23128233
	rs12678919	84	5	13%	1.E-149	6.45	Cholesterol_24097068
	rs4420638	6	0	19%	1.E-149	5.08	Cholesterol_24097068
	rs6927022	1	1	47%	5.E-133	1.44	CrohnIBDUC_23128233
	rs3934467	27	0	22%	3.E-129	2.74	QT_24952745
#14	rs1558902	89	0	42%	5.E-120	2.56	BMI_20935630
	rs3803662	19	0	26%	2.E-114	1.24	BreastCancer_23535729
	rs7759938	31	0	32%	8.E-110	8.33	Menarche_25231870
	rs2954029	22	0	47%	1.E-107	13.16	Cholesterol_24097068
	rs11742570	53	0	40%	2.E-82	1.20	CrohnIBDUC_23128233
	rs2131925	254	9	34%	3.E-74	15.15	Cholesterol_24097068
	rs12916	19	2	40%	5.E-74	1.47	Cholesterol_24097068
	rs4299376	9	0	31%	3.E-73	12.66	Cholesterol_24097068
	rs12994997	72	7	48%	4.E-70	1.23	CrohnIBDUC_23128233
	rs10401969	10	2	9%	1.E-69	8.26	Cholesterol_24097068
	rs6426833	3	0	46%	2.E-68	1.27	CrohnIBDUC_23128233
	rs9533090	6	0	49%	5.E-68	10.00	BoneMineralDensity_22504420
	rs11153730	20	0	50%	2.E-67	1.65	QT_24952745
	rs10453225	81	0	32%	6.E-66	11.11	Menarche_25231870
	rs1883025	3	0	25%	2.E-65	14.29	Cholesterol_24097068
	rs614367	2	0	15%	2.E-63	1.21	BreastCancer_23535729
	rs1366594	5	0	46%	4.E-61	12.50	BoneMineralDensity_22504420
	rs16857031	1	0	13%	7.E-61	2.37	QT_24952745
	rs2153127	14	0	48%	6.E-59	12.50	Menarche_25231870

Top 895 SNPs

Avg:	26.8	1.2	29%	1.E-08	11.41
Median	14	0	30%	2.E-11	1.78
Stdev	35.7	2.8	13%	8.E-08	13.60

Across 11 well-powered association studies:

- BMI
- Bone Mineral Density
- Bipolar
- BreastCancer
- Cholesterol
- CrohnIBDUC
- Height
- Menarche
- QT
- RheumatoidArthritis
- Schizophrenia

895 associated loci

572 (64%) have no protein-coding variants

Genetics, Variation, GWAS, PRS, Mechanism

1. Genetics, Variation, GWAS

2. Polygenic scores (PGS)

3. From GWAS to biological insights

4. From Region to Mechanism / Circuitry