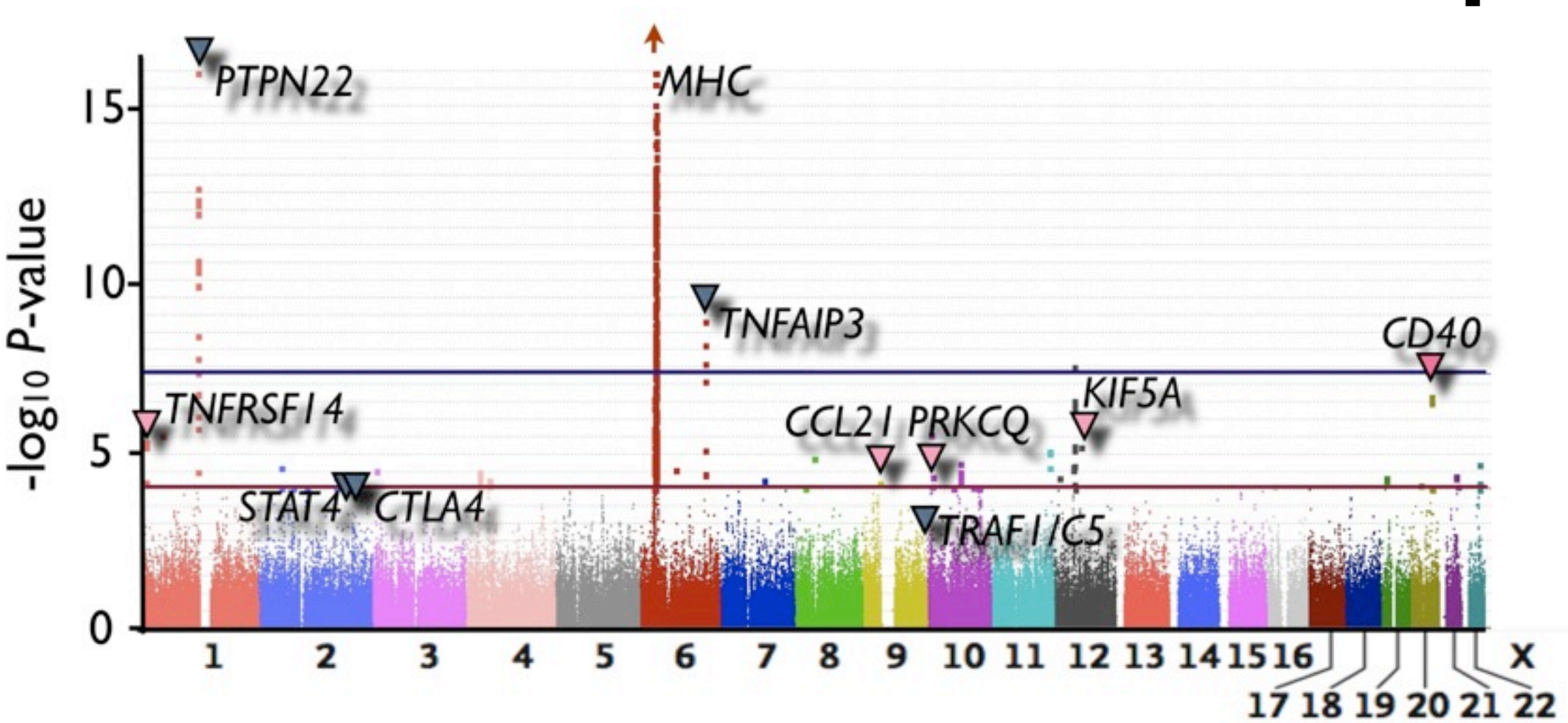
# Fine-Mapping the *CFH* region in Age Related Macular Degeneration

**Soumya Raychaudhuri, MD/PhD**
Postdoctoral Fellow, Daly/Altshuler Laboratory
Medical and Population Genetics Program, Broad Institute

Medical and Population Genetics Program Meeting
*Snapshots of Genome Wide Analysis in Human Disease*

20 May 2010

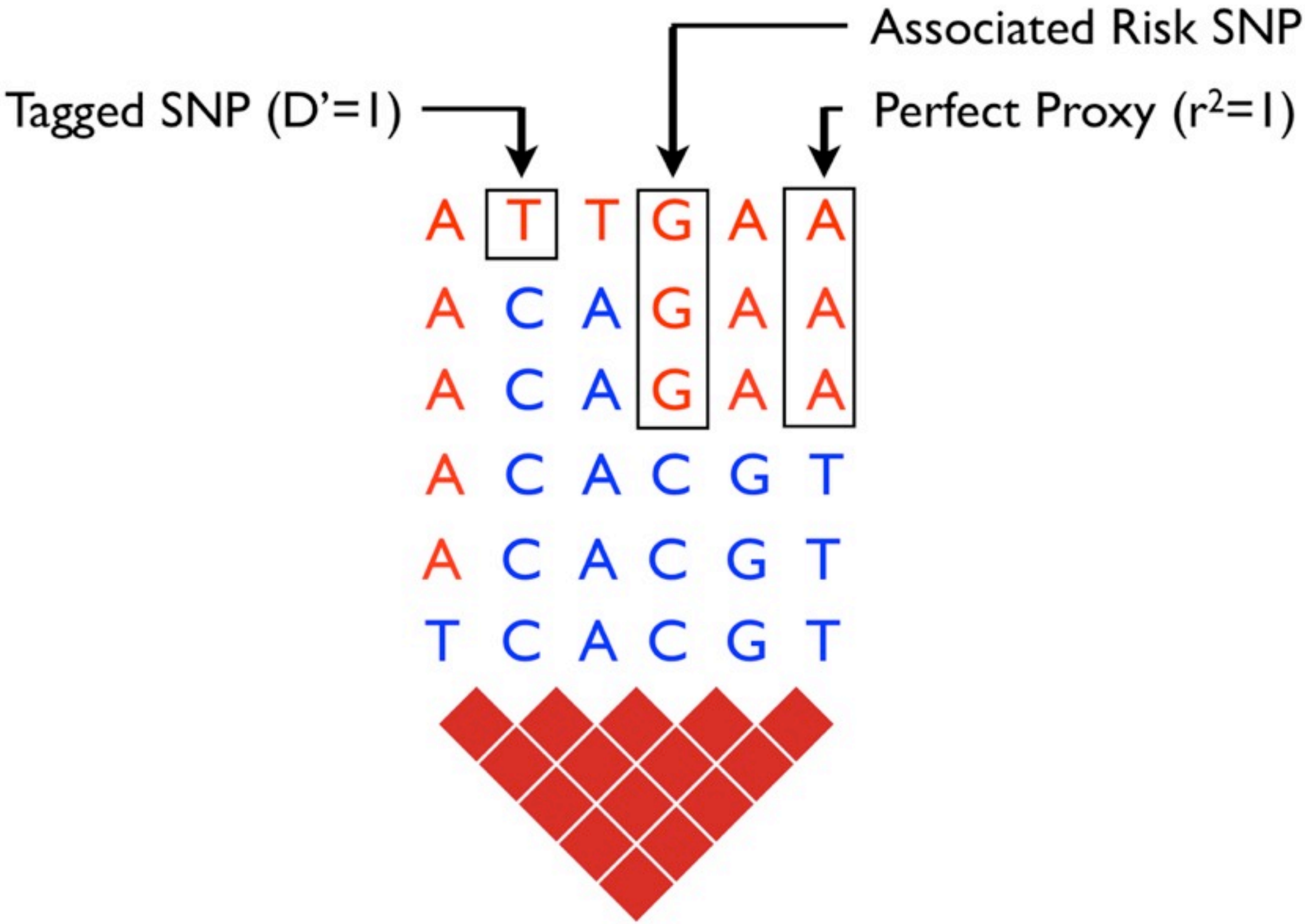
# GWAS is the first step!



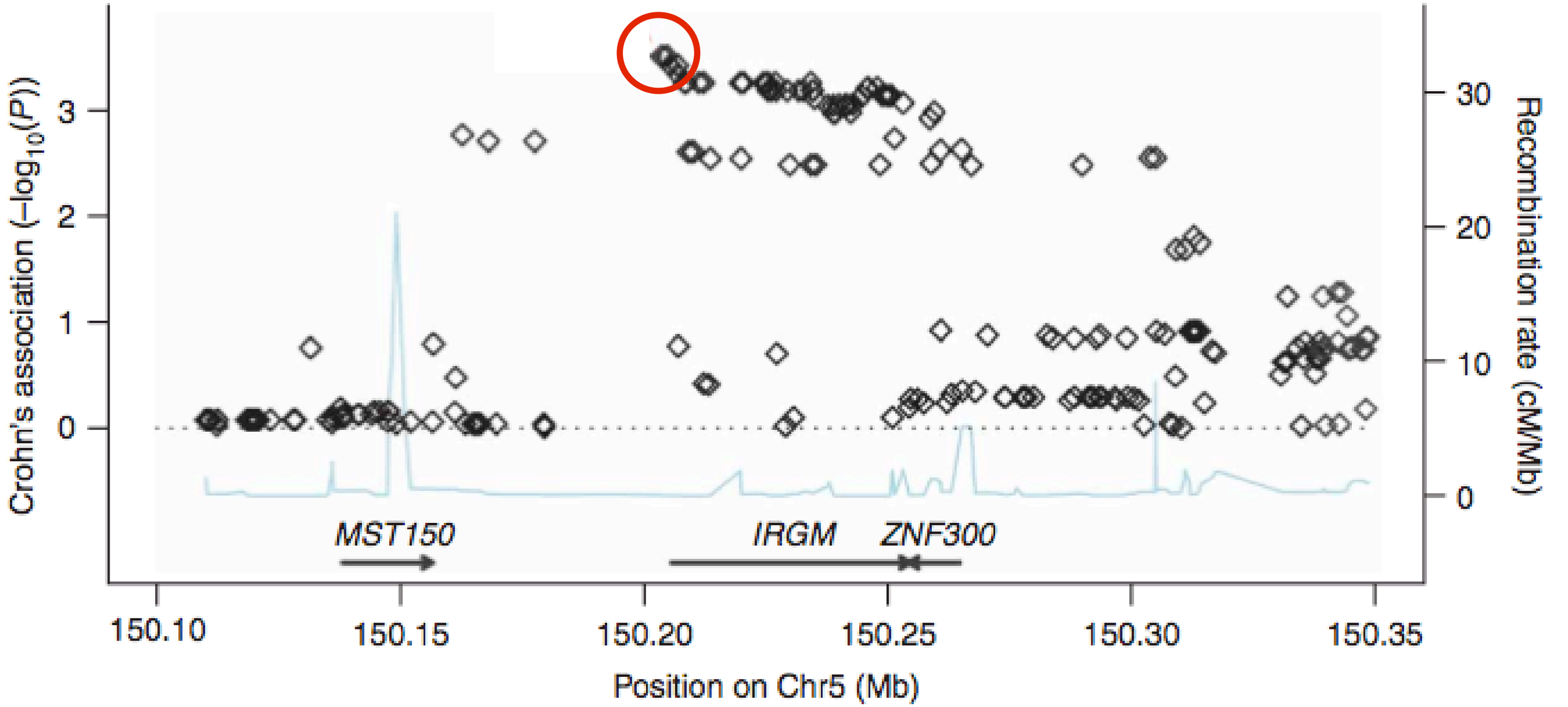

- Highly successful technology to identify key genomic regions
- But ... finding a convincingly associated SNP is the first step!
  - For a given associated SNP, many other nearby SNPs are associated
  - Often not clear whether associated SNP is the 'causal' mutation

# Fine Mapping

- GWAS effectively identifies broad regions of LD associated with disease
- But ...
  - May not implicate specific genes (0-20 genes in LD)
  - May not identify functional variants (intronic/intergenic SNPs)
- The goal of fine mapping is to leverage statistical genetics to find the key 'causal' mutation.


Associated Risk SNP
Tagged SNP (D'=1)
Perfect Proxy (r²=1)
A T T G A A
A C A G A A
A C A G A A
A C A C G T
A C A C G T
T C A C G T

# Crohn's Disease Association to *IRGM*



# Crohn's Disease Association to *IRGM*

## Reference Allele

```
AAAATTCCATTTACAGTAGTATCAAAAGAATAAAAACAATTTTATTTAAA  150183300
ATTTAGAAGTAAATTTTTTAAAATTTAGGAATAAATTTAATAATTTAATA  150183350
ATAGAAGATAAAACTTAGAAATAAATCTAACTAAGAGGGTGAAATATCTG  150183400

  .... (20,050 bp) ....

GTTCACAGACATGGTCACAGAAAAAGCCTGGAAAACTATAGGGTTTTCTG  150203500
CTGACCTCCCACACTCACCATGGACAGTCAGTACCCTGCACTGGCCCGTG  150203550
TCGTACCCAAGCAGAGTGTGCTTGAAAATTGGATGTATATTAGTAGACCC  150203600
```

---

```
AAAATTCCATTTACAGTAGTATCAAAAGAATAAAAACAATTTTATTTAAA  150183300
ATTTAGAAGTAAATTTTTTAAAATTTAGGAATAAATTTAATAATTTAATA  150183350
ATA

  AAATAGA

      AGACATGGTCACAGAAAAAGCCTGGAAAACTATAGGGTTTTCTG  150203500
CTGACCTCCCACACTCACCATGGACAGTCAGTACCCTGCACTGGCCCGTG  150203550
TCGTACCCAAGCAGAGTGTGCTTGAAAATCGGATGTATATTAGTAGACCC  150203600
```

**Risk Allele**

Mccarroll et al
*Nature Genetics* 2008

# Crohn's Disease Association to *IRGM*

## Reference Allele

```
AAAATTCCATTTACAGTAGTATCAAAAGAATAAAAACAATTTTATTTAAA  150183300
ATTTAGAAGTAAATTTTTTAAAATTTAGGAATAAATTTAATAATTTAATA  150183350
ATAGAAGATAAAACTTAGAAATAAATCTAACTAAGAGGGTGAAATATCTG  150183400

  .... (20,050 bp) ....

GTTCACAGACATGGTCACAGAAAAAGCCTGGAAAACTATAGGGTTTTCTG  150203500
CTGACCTCCCACACTCACCATGGACAGTCAGTACCCTGCACTGGCCCGTG  150203550
TCGTACCCAAGCAGAGTGTGCTTGAAAATTGGATGTATATTAGTAGACCC  150203600
```

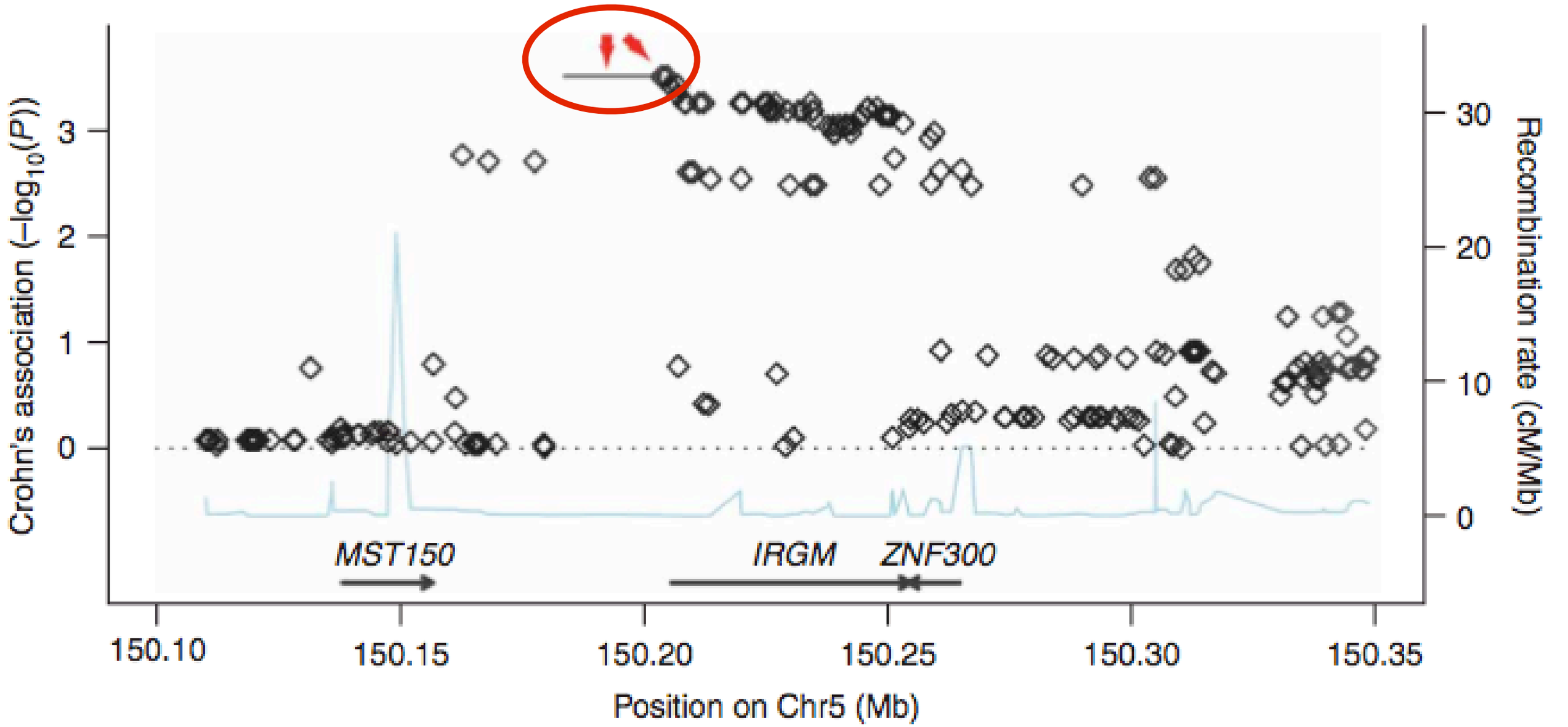
---

```
AAAATTCCATTTACAGTAGTATCAAAAGAATAAAAACAATTTTATTTAAA  150183300
ATTTAGAAGTAAATTTTTTAAAATTTAGGAATAAATTTAATAATTTAATA  150183350
ATA

  AAATAGA

      AGACATGGTCACAGAAAAAGCCTGGAAAACTATAGGGTTTTCTG  150203500
CTGACCTCCCACACTCACCATGGACAGTCAGTACCCTGCACTGGCCCGTG  150203550
TCGTACCCAAGCAGAGTGTGCTTGAAAATCGGATGTATATTAGTAGACCC  150203600
```

## Risk Allele

Mccarroll et al
*Nature Genetics* 2008

# Crohn's Disease Association to *IRGM*



# Fine Mapping with Logistic Regression

- Since multiple variants are involved, a statistical framework is important
- Logistic regression is a key tool,
  - Assumes that each risk allele has an 'additive' effect
  - Observed associations in the region should be obviated when conditioning on the causal variant(s)
  - The statistical evidence of the causal variant might not be obviated when conditioning on other variants

# *CFH* Region in AMD

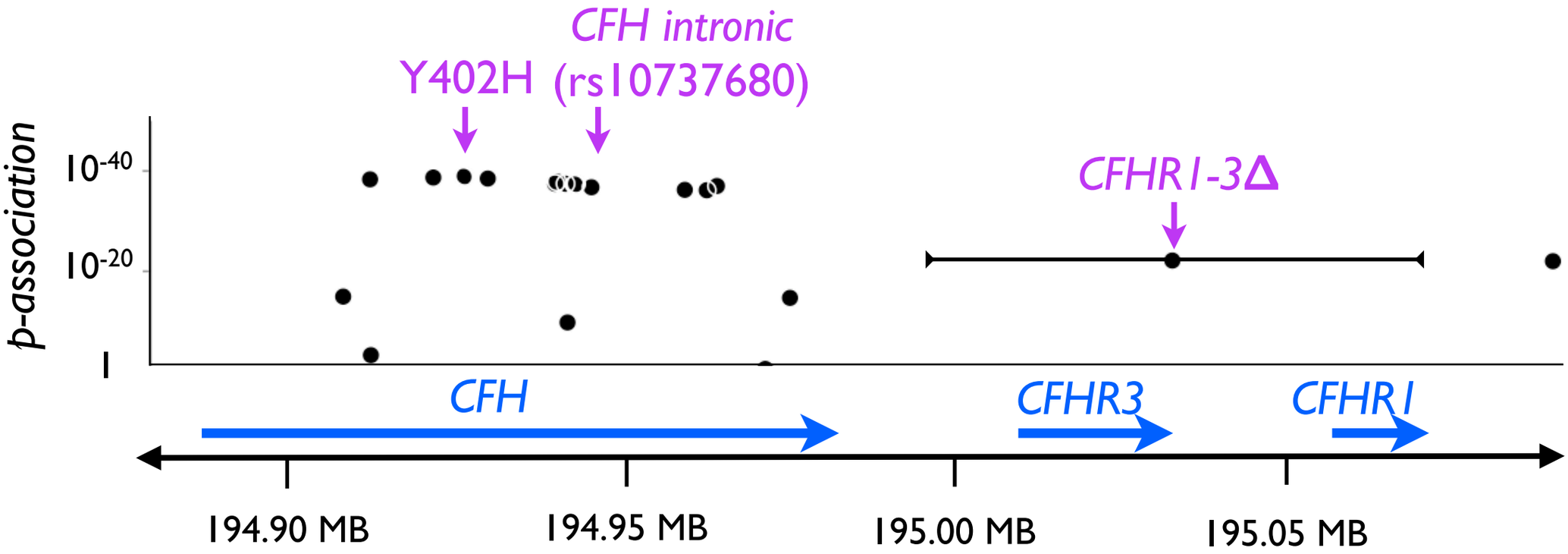
- Multiple alleles have demonstrated strong association to *CFH*
- In many cases strong effect sizes : ORs > 2
- But, much LD across the region
- A nice example for fine mapping

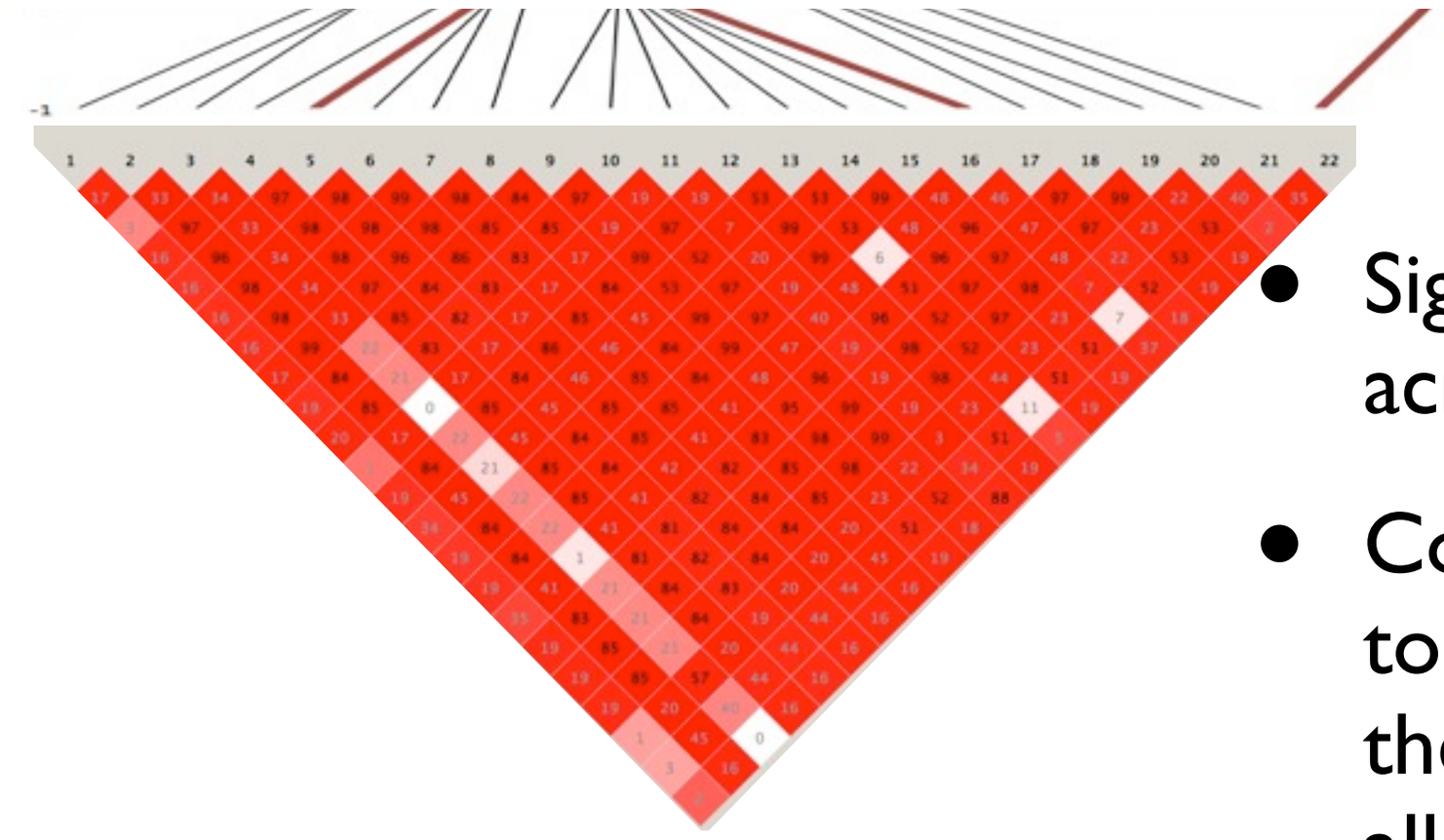
# *CFH* Region in AMD

- Previously associated alleles:

1. **Y402H -** Klein et al (and others) *Science* 2005
2. **rs10737680 (*CFH intronic*)**
    - Maller et al *Nature Genetics* 2006
    - Li et al *Nature Genetics* 2006
3. ***CFHR3-R1Δ*** a common deletion polymorphism
    - Hughes et al *Nature Genetics* 2006

- So ...Which ones are actually "causal" ?



- Significant associations across the entire region!
- Conditional analysis to try to disentangle the effect of these three different alleles

# Conditional Analysis

| Logistic Regression Model | Y402H | | | *CFH* intronic | | | *CFHR3/1*Δ | | |
|---|---|---|---|---|---|---|---|---|---|
| | OR | | *p* | OR | | *p* | OR | | *p* |
| Single Marker | 0.39 | (0.34 -0.46) | 1.2E-35 | 0.38 | (0.33 -0.45) | 1.6E-32 | 0.37 | (0.30 -0.45) | 6.5E-21 |

# Conditional Analysis

| Logistic Regression Model | Y402H | | | *CFH* intronic | | | *CFHR3/1*Δ | | |
|---|---|---|---|---|---|---|---|---|---|
| | OR | | *p* | OR | | *p* | OR | | *p* |
| Single Marker | 0.39 | (0.34 -0.46) | 1.2E-35 | 0.38 | (0.33 -0.45) | 1.6E-32 | 0.37 | (0.30 -0.45) | 6.5E-21 |
| Condition on Y402H | | | | 0.58 | (0.47 -0.71) | 1.1E-07 | 0.58 | (0.46 -0.72) | 2.3E-06 |

# Conditional Analysis

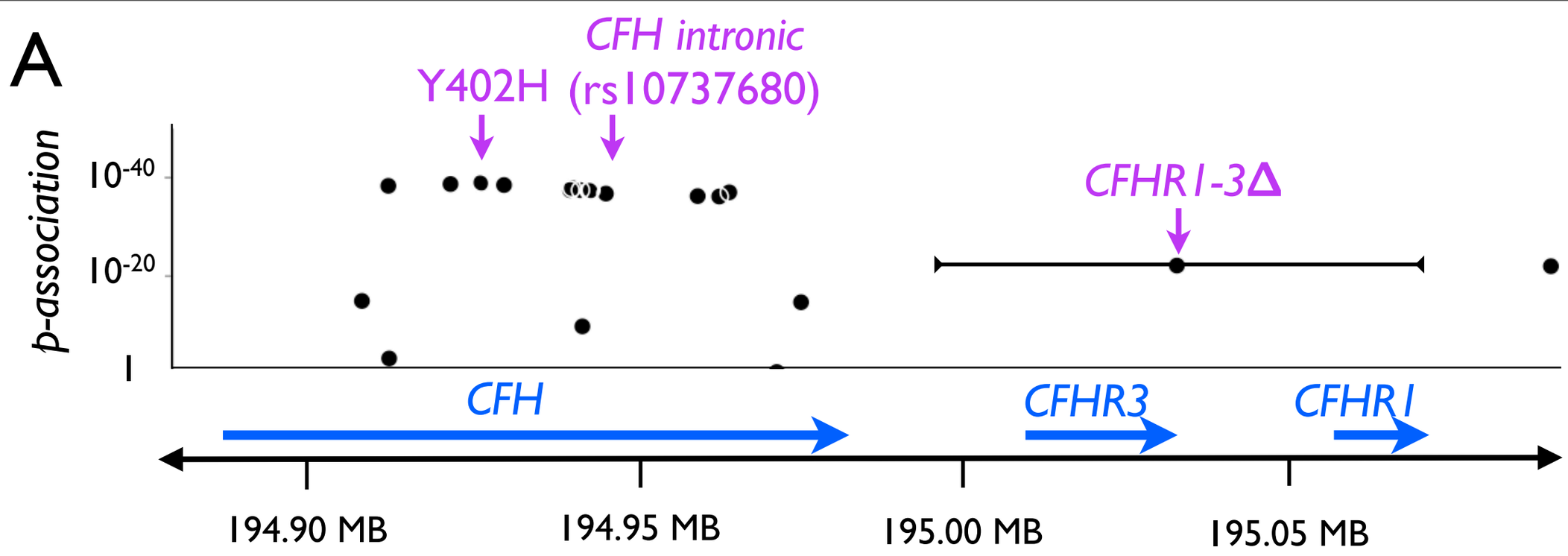
| Logistic Regression Model | Y402H | | | *CFH* intronic | | | *CFHR3/1*Δ | | |
|---|---|---|---|---|---|---|---|---|---|
| | OR | | *p* | OR | | *p* | OR | | *p* |
| Single Marker | 0.39 | (0.34 -0.46) | 1.2E-35 | 0.38 | (0.33 -0.45) | 1.6E-32 | 0.37 | (0.30 -0.45) | 6.5E-21 |
| Condition on Y402H | | | | 0.58 | (0.47 -0.71) | 1.1E-07 | 0.58 | (0.46 -0.72) | 2.3E-06 |
| Condition on CFH intronic | 0.55 | (0.46 -0.66) | 4.5E-10 | | | | 0.72 | (0.55 -0.95) | 0.02 |

# Conditional Analysis

| Logistic Regression Model | Y402H | | | *CFH* intronic | | | *CFHR3/1Δ* | | |
|---|---|---|---|---|---|---|---|---|---|
| | OR | | *p* | OR | | *p* | OR | | *p* |
| Single Marker | 0.39 | (0.34 -0.46) | 1.2E-35 | 0.38 | (0.33 -0.45) | 1.6E-32 | 0.37 | (0.30 -0.45) | 6.5E-21 |
| Condition on Y402H | | | | 0.58 | (0.47 -0.71) | 1.1E-07 | 0.58 | (0.46 -0.72) | 2.3E-06 |
| Condition on CFH intronic | 0.55 | (0.46 -0.66) | 4.5E-10 | | | | 0.72 | (0.55 -0.95) | 0.02 |
| Condition on CFHR3/1Δ | 0.47 | (0.40 -0.55) | 7.7E-21 | 0.45 | (0.37 -0.55) | 1.8E-14 | | | |

# Summary of Conditional Analysis

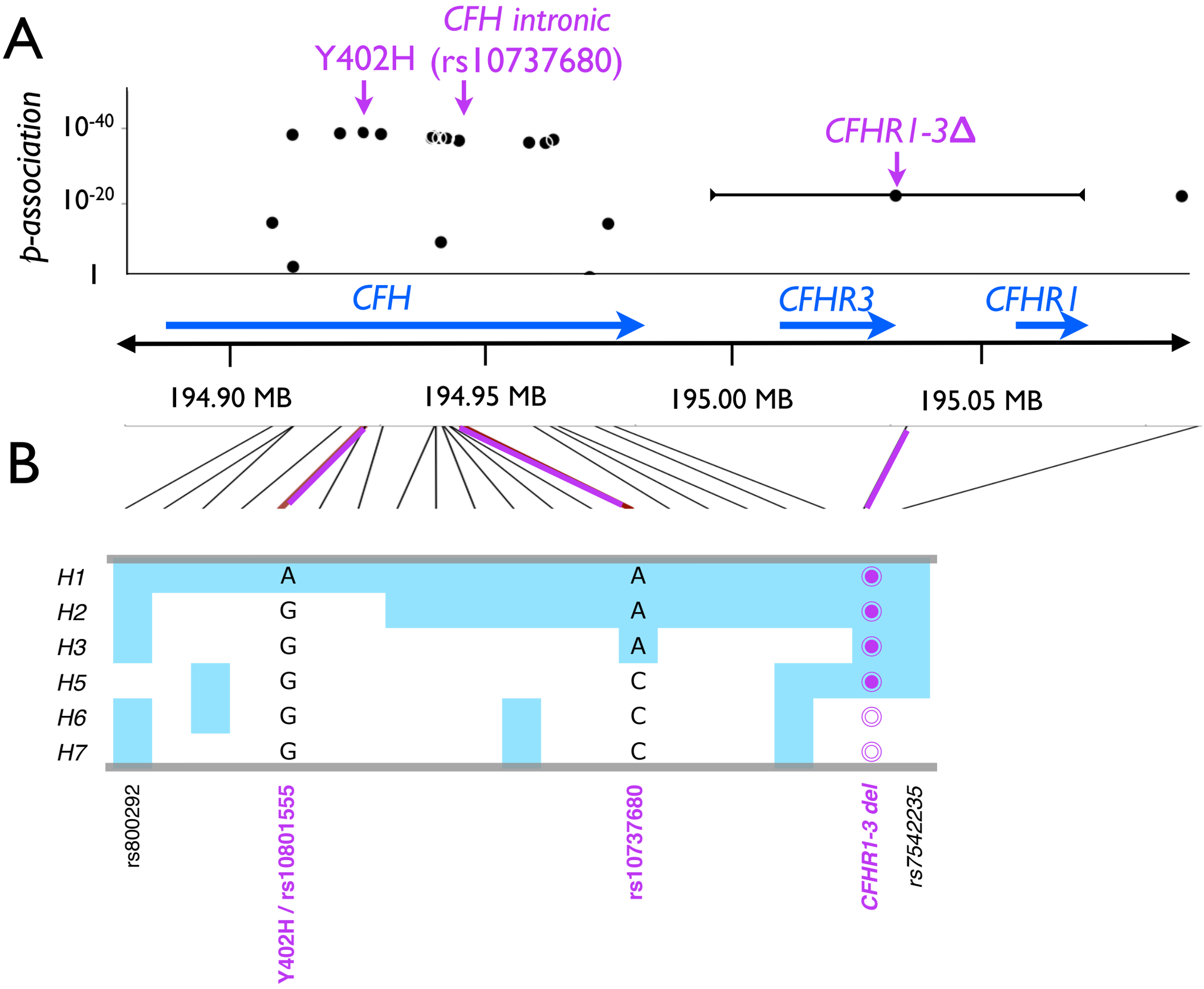
- *Y402H* and the *CFH* intronic SNP are independent
- *Y402H* and *CFHR3-1*Δ are independent
- But the effect of *CFHR3-1*Δ is obviated after conditioning on the *CFH* intronic SNP
- But the effect of the *CFH* intronic SNP is **not** obviated when conditioning on *CFHR3-1*Δ



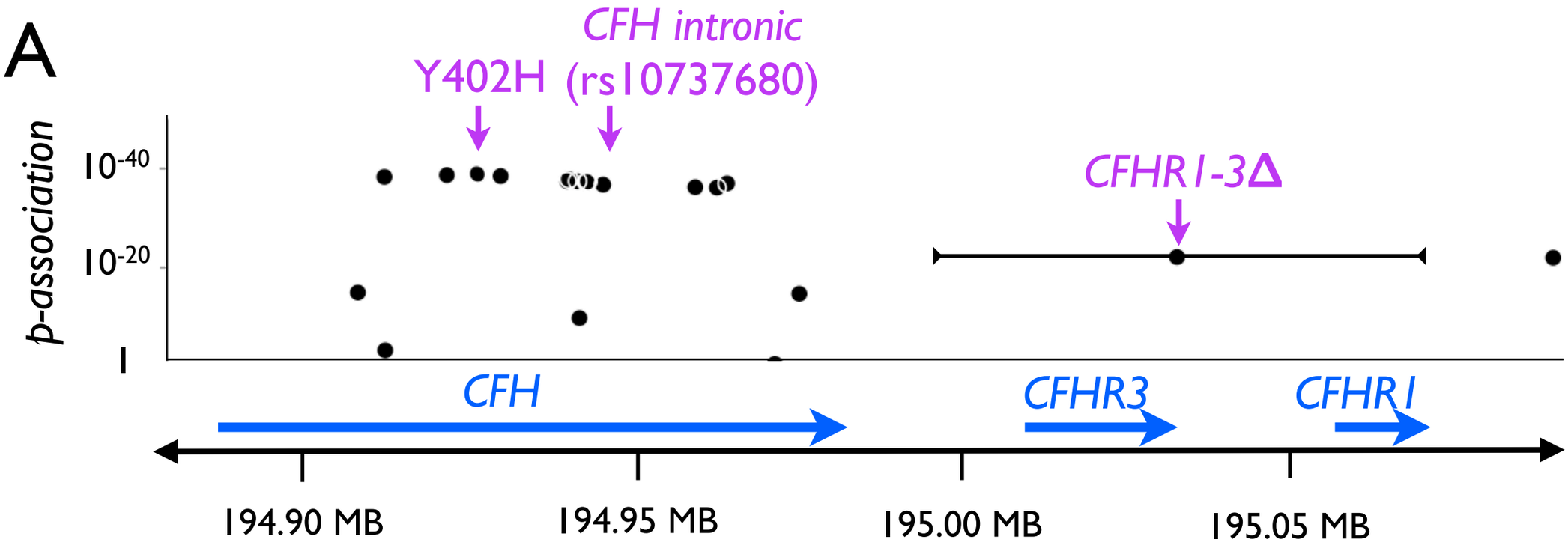

- Haplotype analysis with Beagle & plink

A

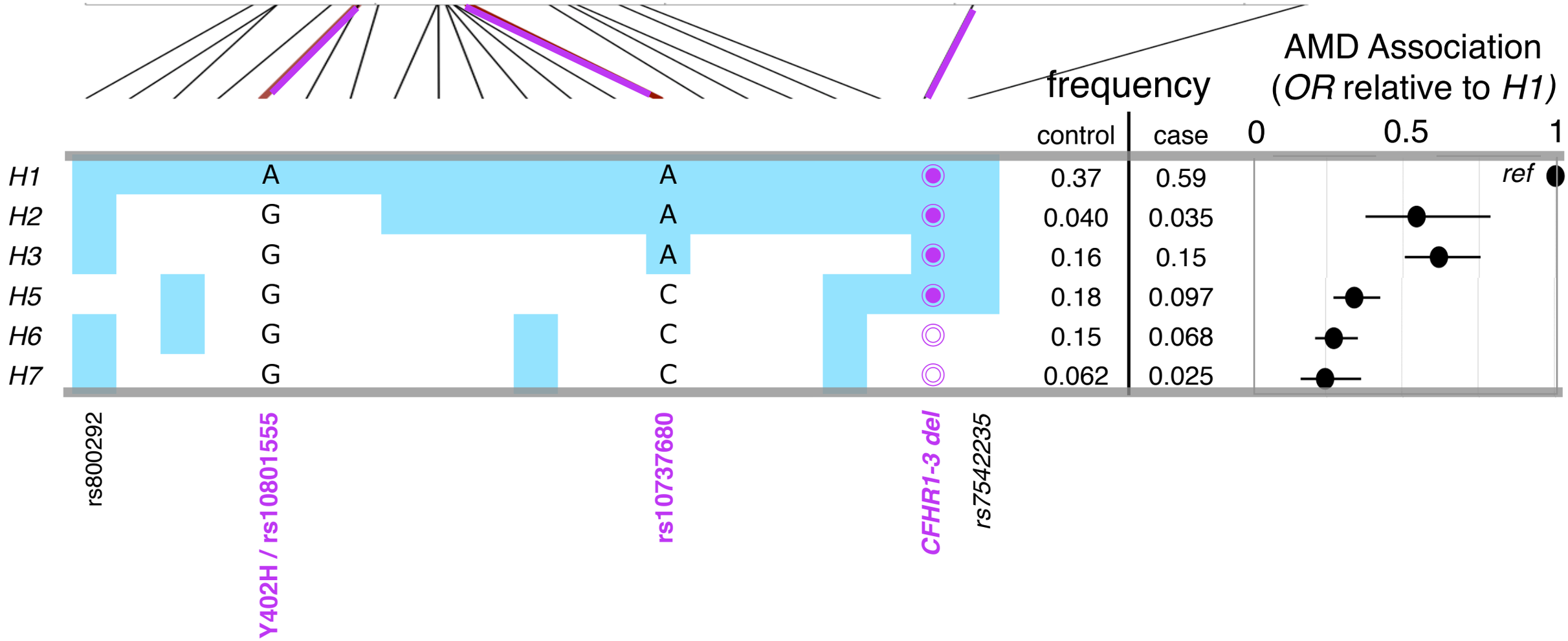
B


CFH intronic
Y402H
(rs10737680)
CFHR1-3Δ
p-association
10-40
10-20
1
CFH
CFHR3
CFHR1
194.90 MB
194.95 MB
195.00 MB
195.05 MB
H1
H2
H3
H5
H6
H7
A
G
G
G
G
G
A
A
A
C
C
C
rs800292
Y402H / rs10801555
rs10737680
CFHR1-3 del
rs7542235

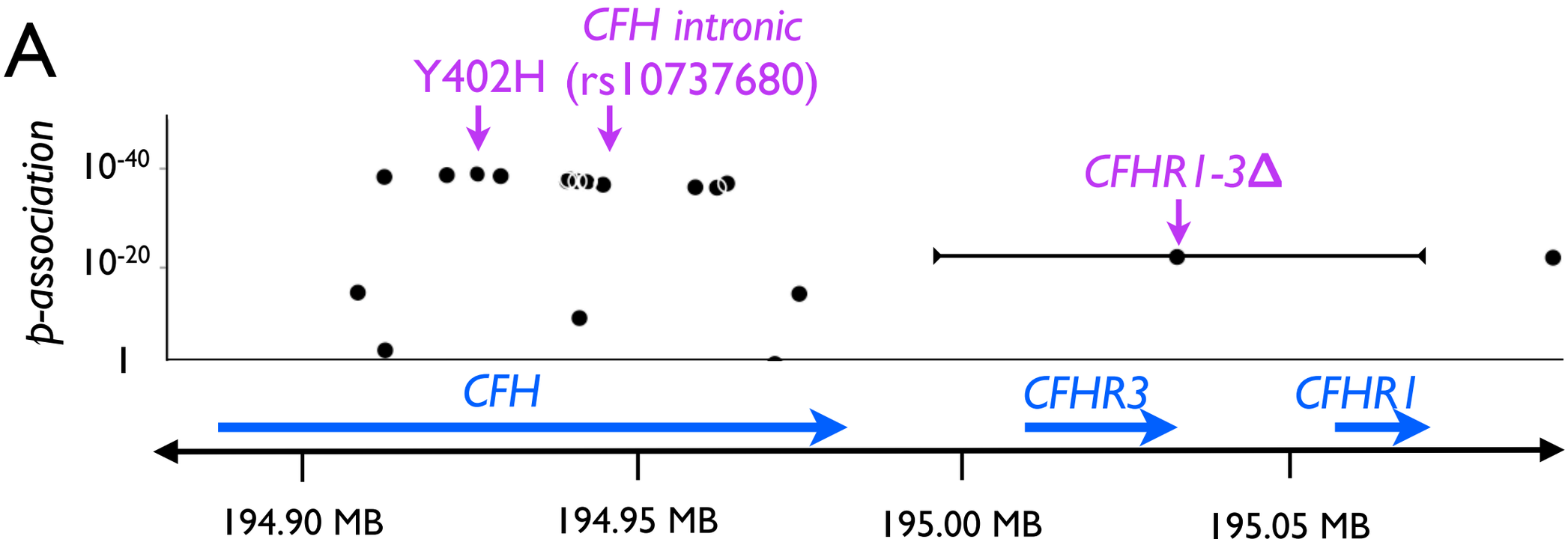
A


CFH intronic
Y402H
(rs10737680)
CFHR1-3Δ
p-association
10-40
10-20
1
CFH
CFHR3
CFHR1
194.90 MB
194.95 MB
195.00 MB
195.05 MB

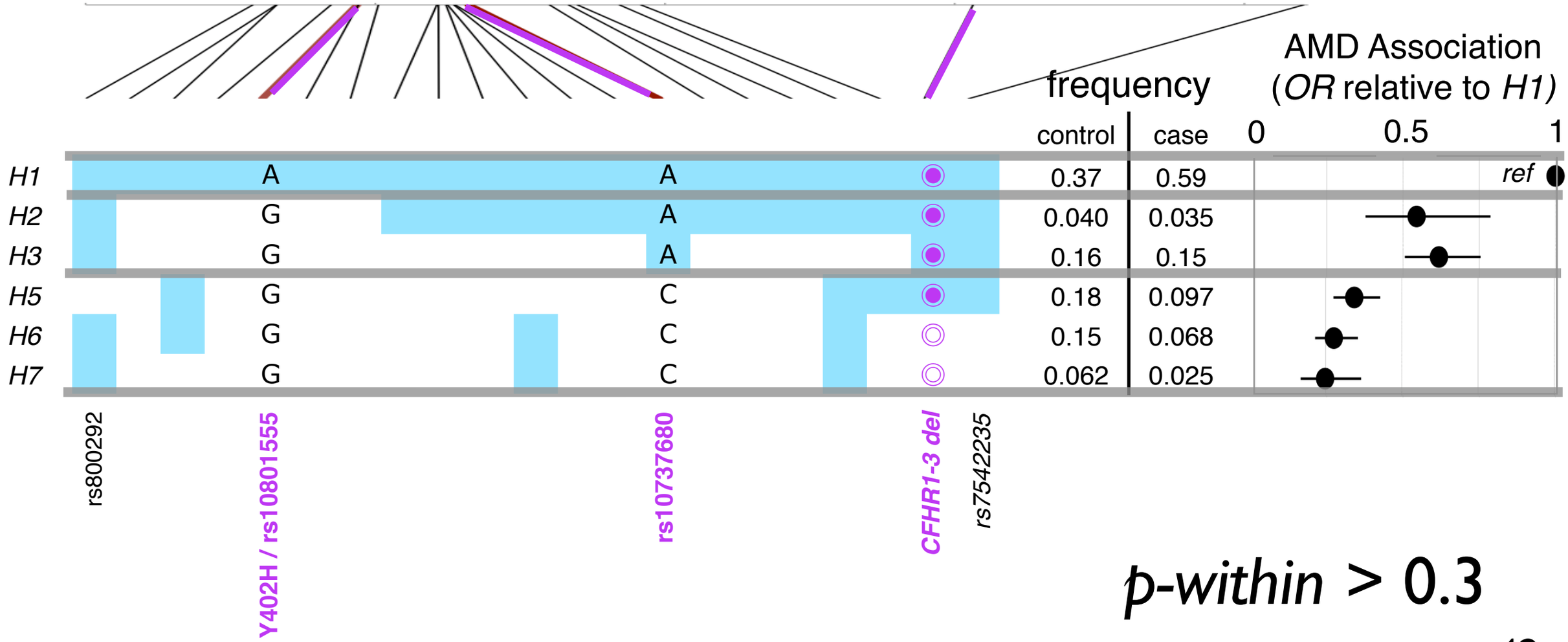
B


frequency
AMD Association
(OR relative to H1)
control
case
0
0.5
1
H1 A A 0.37 0.59 ref
H2 G A 0.040 0.035
H3 G A 0.16 0.15
H5 G C 0.18 0.097
H6 G C 0.15 0.068
H7 G C 0.062 0.025
rs800292
Y402H / rs10801555
rs10737680
CFHR1-3 del
rs7542235

A




B




*p-within* > 0.3

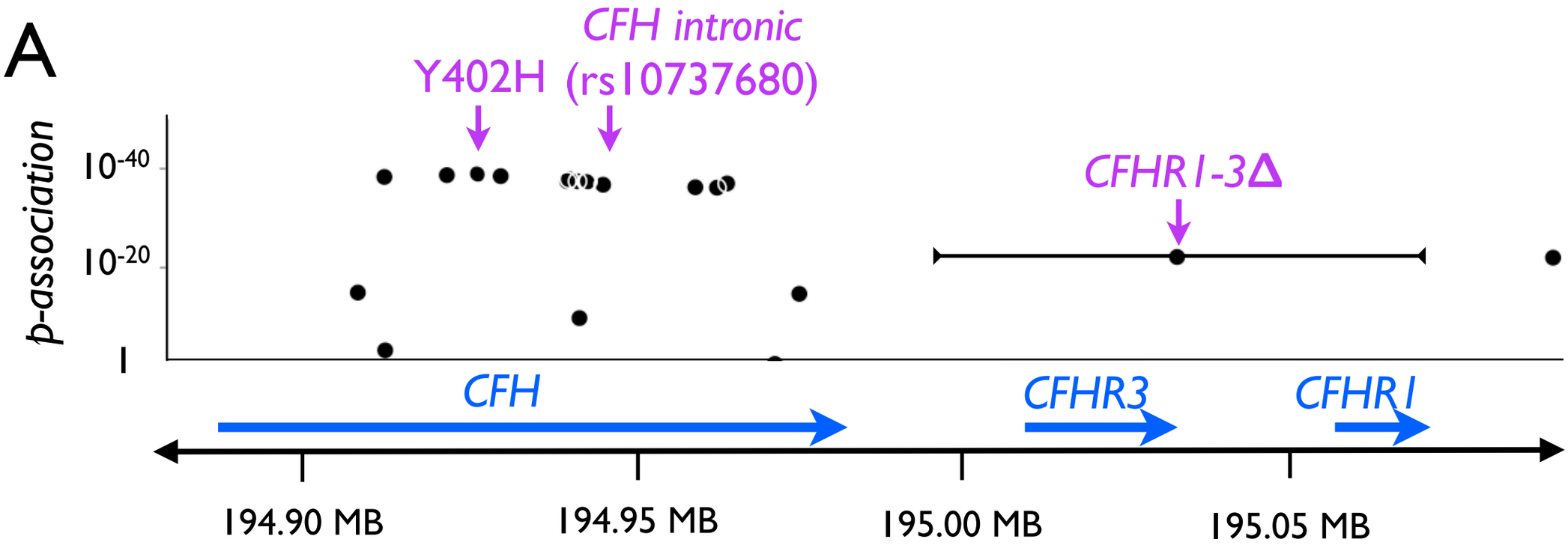
*p-between* < $10^{-42}$

# Summary

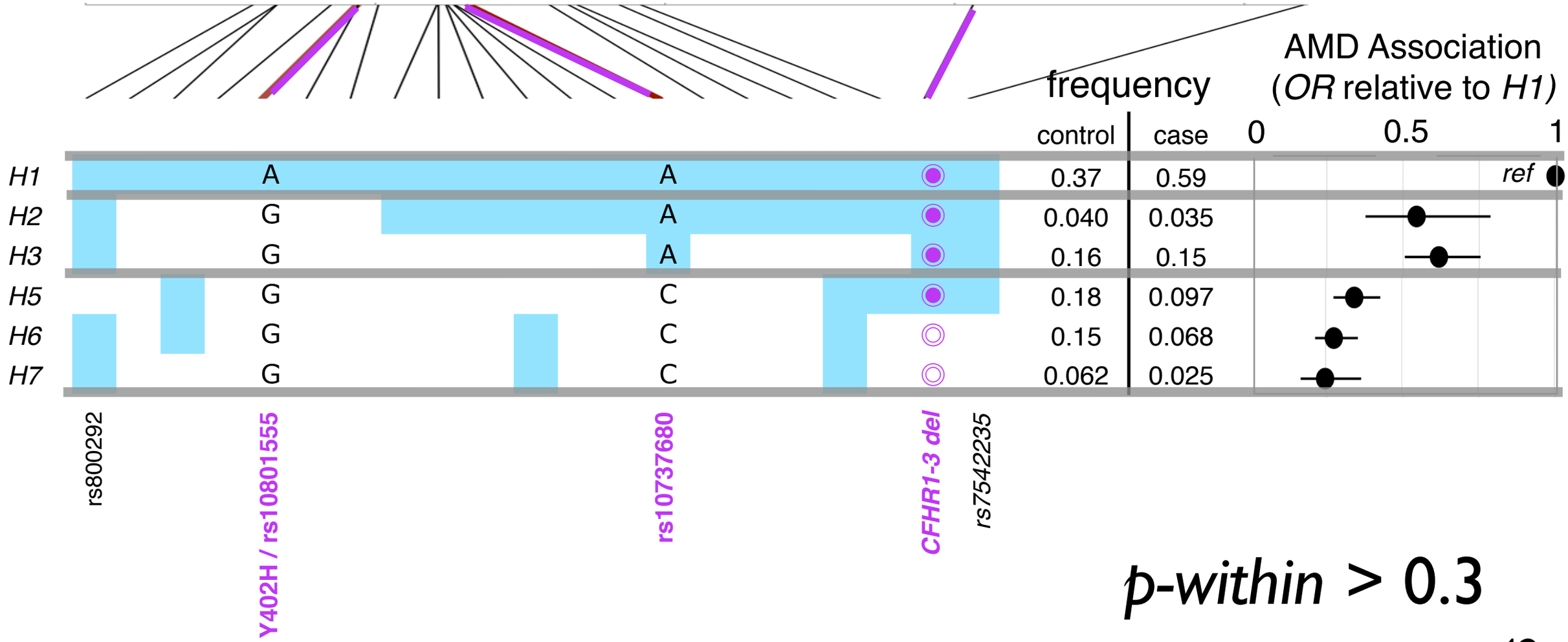
- Fine Mapping of the *CFH* region suggests

1. 3 tiers of risk are evident in the data set
2. *Y402H* is strong and independent of other variants
3. The *CHFR3-1*Δ does not seem to confer independent disease risk
4. At least one other risk conferring variant is present
    - Captured well, by previously published intronic Maller et al *CFH* intronic SNP

A




B




*p-within* > 0.3

*p-between* < $10^{-42}$

A

*CFH intronic*
Y402H (rs10737680)

*CFHR1-3Δ*

*p-association*

$10^{-40}$

$10^{-20}$

1

*CFH*

*CFHR3*

*CFHR1*

194.90 MB

194.95 MB

195.00 MB

195.05 MB

B

| | | | | case | control |
|---|---|---|---|---|---|
| *H1* | A | A | ◉ | 0.59 | 0.37 |
| *H2* | G | A | ◉ | 0.035 | 0.040 |
| *H3* | G | A | ◉ | 0.15 | 0.16 |
| *H4* | G | A | ◉ | 0.0058 | 0.015 |
| *H5* | G | C | ◉ | 0.097 | 0.18 |
| *H6* | G | C | ◎ | 0.068 | 0.15 |
| *H7* | G | C | ◎ | 0.025 | 0.062 |

AMD Association
(*OR* relative to *H1)*

frequency

0 0.5 1

*ref*

rs800292

Y402H / rs10801555

rs10737680

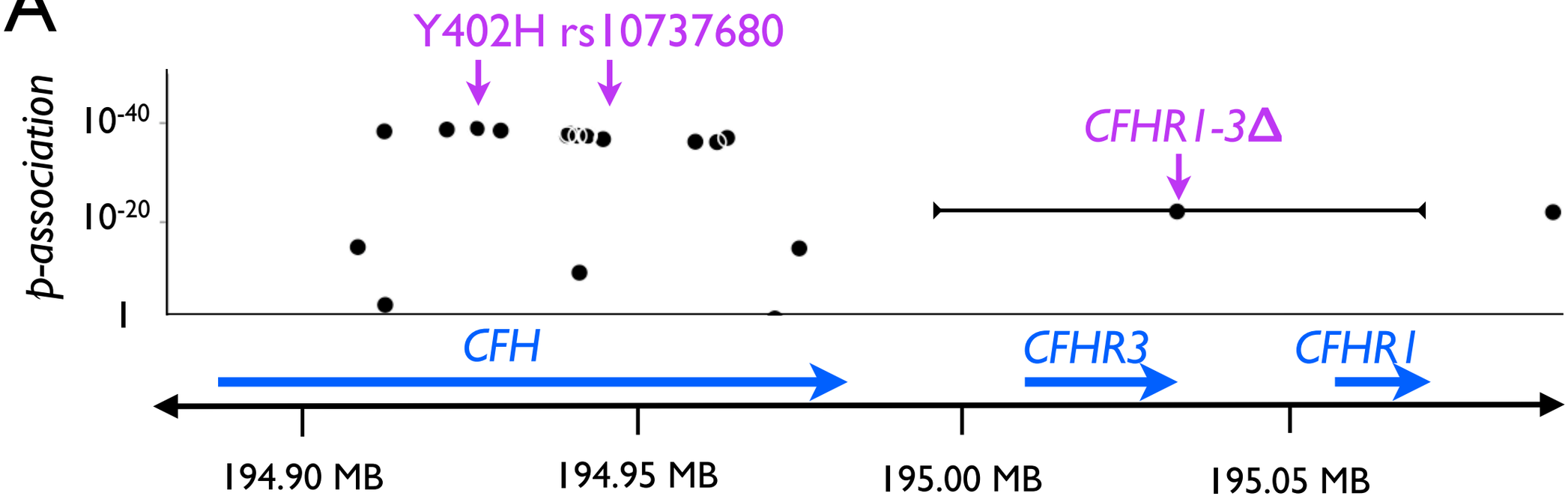
*CFHR1-3 del*

rs7542235

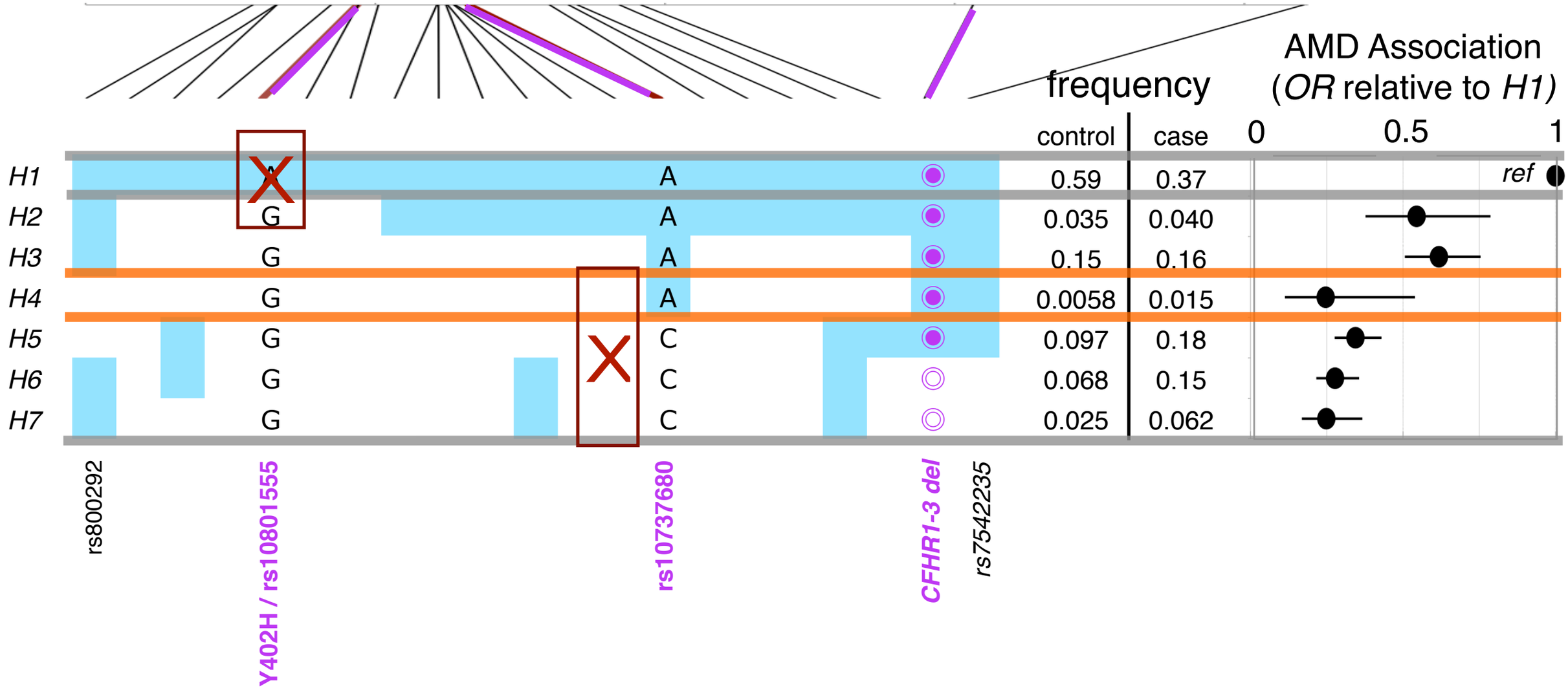
# Two Ways to Understand this Data

- Two Variant Hypothesis
  - *Y402H*
  - Functional variant, strongly correlated with CFH intronic SNP (rs10737680)

A


Y402H rs10737680
CFHR1-3Δ
p-association
10-40
10-20
1
CFH
CFHR3
CFHR1
194.90 MB
194.95 MB
195.00 MB
195.05 MB

B


frequency
AMD Association
(OR relative to H1)
control
case
0
0.5
1
H1
H2
H3
H4
H5
H6
H7
A
A
A
A
C
C
C
G
G
G
G
G
G
G
0.59
0.035
0.15
0.0058
0.097
0.068
0.025
0.37
0.040
0.16
0.015
0.18
0.15
0.062
ref
rs800292
Y402H / rs10801555
rs10737680
CFHR1-3 del
rs7542235

# Two Ways to Understand this Data

- Two Variant Hypothesis
  - *Y402H*
    - Functional variant, strongly correlated with rs10737680,
- Multiple Variant Hypothesis
  - *Y402H*
    - Two other Functional Variants, conferring equivalent effect size.

A

Y402H rs10737680

*p-association*

$10^{-40}$

$10^{-20}$

1

*CFHR1-3Δ*

*CFH*

*CFHR3*

*CFHR1*

194.90 MB

194.95 MB

195.00 MB

195.05 MB

B

| | | | | frequency | | AMD Association (*OR* relative to *H1*) |
|---|---|---|---|---|---|---|
| | rs800292 | Y402H / rs10801555 | rs10737680 | *CFHR1-3 del* | rs7542235 | control | case | 0 0.5 1 |
| *H1* | | T | A | ◉ | | 0.59 | 0.37 | *ref* |
| *H2* | | G | A | ◉ | | 0.035 | 0.040 | |
| *H3* | | G | A | ◉ | | 0.15 | 0.16 | |
| *H4* | | G | A | ◉ | | 0.0058 | 0.015 | |
| *H5* | | G | C | ◉ | | 0.097 | 0.18 | |
| *H6* | | G | C | ◎ | | 0.068 | 0.15 | |
| *H7* | | G | C | ◎ | | 0.025 | 0.062 | |

# Definitively identifying risk variants

- Sequencing the entire *CFH* gene region
  - Exons, introns, promotor, and 5' UTR region
- Sequence ~8 affected/unaffected individuals with each haplotype
  - Total of 9 haplotypes with >0.5% frequency
- Method : long-range PCR & next-generation sequencing

Mark Daly

Johanna Seddon

Stephan Ripke

Ben Neale

Jesen Fagerness

Robyn Reynolds

Lucia Sobrin

David Altshuler

Oleg Iartchouk

Anand Swaroop

Mingyao Li

Goncalo Abecasis


soumya@broadinstitute.org

NIAMS

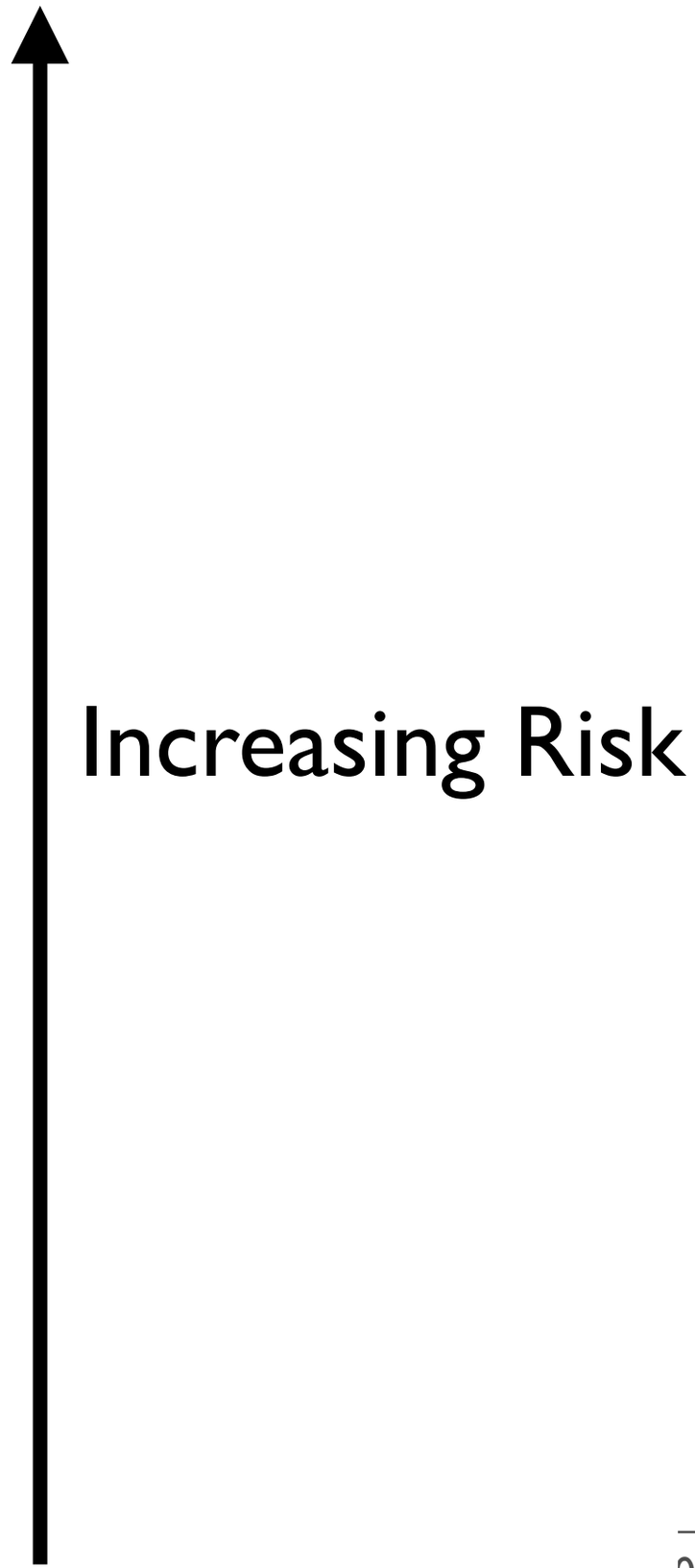
ARRA

# Identifying the Smoking Gun(s)

- Imputation:
  - Imputed SNPs using
    - 1. Hapmap phase III
    - Li/Abecasis reference panel of ~80 CFH SNPs
- No smoking gun SNPs identified fitting the two variant hypothesis

# Available Samples for Sequencing

| | | |
|---|---|---|
| Haplotype | HF | %A |
| *G*GC*C*GGTCTGATA*C*GCGTT*O*A | 0.32% | 0.91 |
| *G*AT*A*AACACAAAC*A*ATATA*O*A | 46.28% | 0.53 |
| *G*GC*C*GGTCTGGTA*A*GCGCT*O*A | 15.64% | 0.40 |
| *G*GC*C*GGTACAAAC*A*ATATA*O*A | 3.78% | 0.36 |
| *A*GT*C*GGTCTGGTA*C*GCGTA*O*A | 14.62% | 0.27 |
| *G*GT*C*GGTCTGGTA*C*GCGTA*O*A | 0.47% | 0.25 |
| *G*GT*C*GGTCTGATA*C*GCGTT*-*G | 12.01% | 0.23 |
| *A*GC*C*GGTCTGGTA*A*GCGCT*O*A | 1.14% | 0.23 |
| *G*GC*C*GGTCTGATA*C*GCGTT*-*G | 4.77% | 0.21 |
| *A*GC*C*GGTCTGGTA*C*GCGTA*O*A | 0.41% | 0.21 |
| *A*GT*C*GGTCTGATA*C*GCGTT*-*G | 0.56% | 0.16 |



# Available Samples for Sequencing

| Haplotype | HF | %A | |
|---|---|---|---|
| *G*GC*C*GGTCTGATA*C*GCGTT*O*A | 0.32% | 0.91 | Rare |
| *G*AT*A*AACACAAAC*A*ATATA*O*A | 46.28% | 0.53 | |
| *G*GC*C*GGTCTGGTA*A*GCGCT*O*A | 15.64% | 0.40 | |
| *G*GC*C*GGTACAAAC*A*ATATA*O*A | 3.78% | 0.36 | |
| *A*GT*C*GGTCTGGTA*C*GCGTA*O*A | 14.62% | 0.27 | |
| *G*GT*C*GGTCTGGTA*C*GCGTA*O*A | 0.47% | 0.25 | Rare |
| *G*GT*C*GGTCTGATA*C*GCGTT*-*G | 12.01% | 0.23 | |
| *A*GC*C*GGTCTGGTA*A*GCGCT*O*A | 1.14% | 0.23 | Rare |
| *G*GC*C*GGTCTGATA*C*GCGTT*-*G | 4.77% | 0.21 | |
| *A*GC*C*GGTCTGGTA*C*GCGTA*O*A | 0.41% | 0.21 | Rare |
| *A*GT*C*GGTCTGATA*C*GCGTT*-*G | 0.56% | 0.16 | Rare |

Increasing Risk

# Available Samples for Sequencing

| Haplotype | HF | %A | |
|---|---|---|---|
| *G*AT*A*AACACAAAC*A*ATATA*O*A | 46.28% | 0.53 | Rare |
| *G*GC*C*GGTCTGGTA*A*GCGCT*O*A | 15.64% | 0.40 | |
| *G*GC*C*GGTACAAAC*A*ATATA*O*A | 3.78% | 0.36 | |
| *A*GT*C*GGTCTGGTA*C*GCGTA*O*A | 14.62% | 0.27 | Rare |
| *G*GT*C*GGTCTGATA*C*GCGTT*-*G | 12.01% | 0.23 | Rare |
| *G*GC*C*GGTCTGATA*C*GCGTT*-*G | 4.77% | 0.21 | Rare |
| | | | Rare |

Increasing Risk

# Available Samples for Sequencing

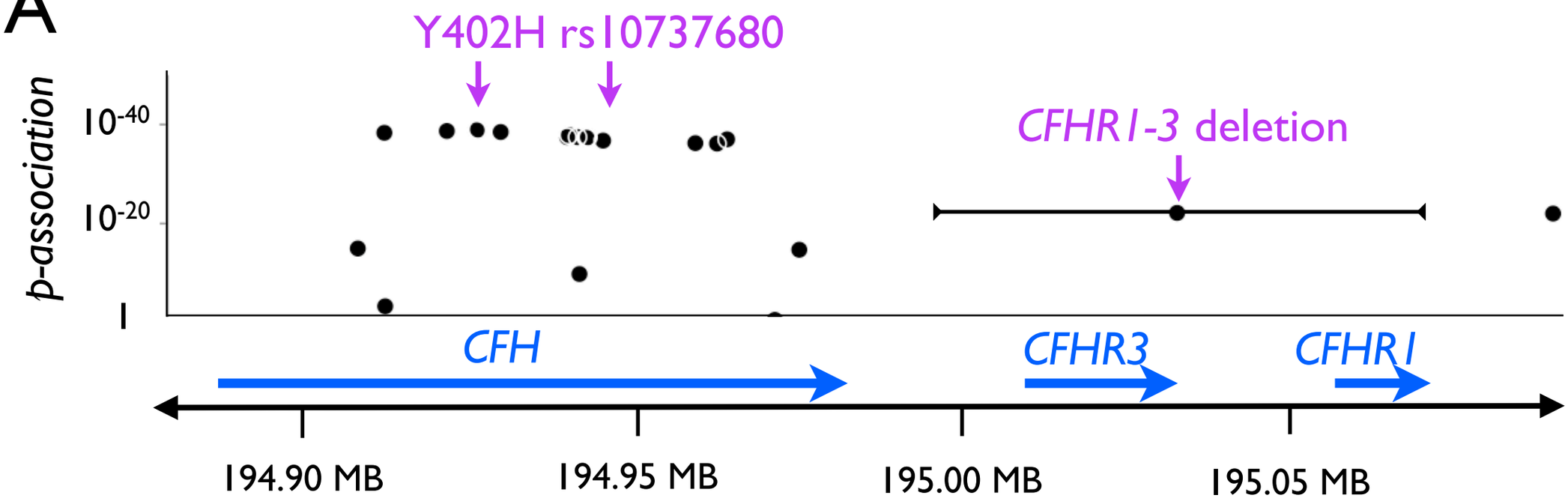
| | | |
|---|---|---|
| Haplotype | HF | %A |
| *G*GC*C*GGTCTGATA*C*GCGTT*O*A | 0.32% | 0.91 |
| *G*AT*A*AACACAAAC*A*ATATA*O*A | 46.28% | 0.53 |
| *G*GC*C*GGTCTGGTA*A*GCGCT*O*A | 15.64% | 0.40 |
| *G*GC*C*GGTACAAAC*A*ATATA*O*A | 3.78% | 0.36 |
| *A*GT*C*GGTCTGGTA*C*GCGTA*O*A | 14.62% | 0.27 |
| *G*GT*C*GGTCTGGTA*C*GCGTA*O*A | 0.47% | 0.25 |
| *G*GT*C*GGTCTGATA*C*GCGTT*-*G | 12.01% | 0.23 |
| *A*GC*C*GGTCTGGTA*A*GCGCT*O*A | 1.14% | 0.23 |
| *G*GC*C*GGTCTGATA*C*GCGTT*-*G | 4.77% | 0.21 |
| *A*GC*C*GGTCTGGTA*C*GCGTA*O*A | 0.41% | 0.21 |
| *A*GT*C*GGTCTGATA*C*GCGTT*-*G | 0.56% | 0.16 |

Increasing Risk

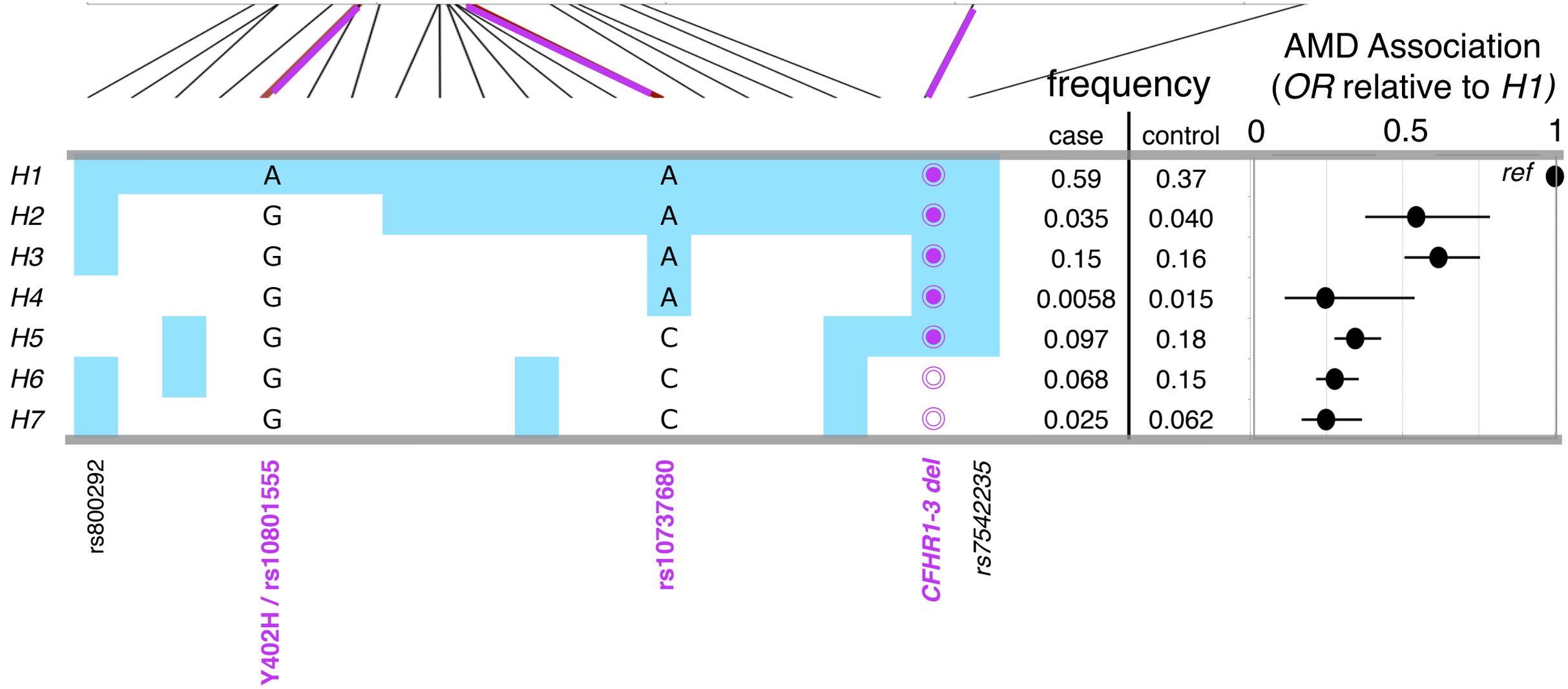
# Available Samples for Sequencing

| | | | Homozygotes | |
|---|---|---|---|---|
| Haplotype | HF | %A | A | U |
| *G*GC*C*GGTCTGATA*C*GCGTT*O*A | 0.32% | 0.91 | | |
| *G*AT*A*AACACAAAC*A*ATATA*O*A | 46.28% | 0.53 | 236 | 38 |
| *G*GC*C*GGTCTGGTA*A*GCGCT*O*A | 15.64% | 0.40 | 12 | 9 |
| *G*GC*C*GGTACAAAC*A*ATATA*O*A | 3.78% | 0.36 | 1 | |
| *A*GT*C*GGTCTGGTA*C*GCGTA*O*A | 14.62% | 0.27 | 11 | 8 |
| *G*GT*C*GGTCTGGTA*C*GCGTA*O*A | 0.47% | 0.25 | | |
| *G*GT*C*GGTCTGATA*C*GCGTT*-*G | 12.01% | 0.23 | 2 | 6 |
| *A*GC*C*GGTCTGGTA*A*GCGCT*O*A | 1.14% | 0.23 | | |
| *G*GC*C*GGTCTGATA*C*GCGTT*-*G | 4.77% | 0.21 | | 3 |
| *A*GC*C*GGTCTGGTA*C*GCGTA*O*A | 0.41% | 0.21 | | |
| *A*GT*C*GGTCTGATA*C*GCGTT*-*G | 0.56% | 0.16 | | |

A


Y402H rs10737680
CFHR1-3 deletion
p-association
10-40
10-20
1
CFH
CFHR3
CFHR1
194.90 MB
194.95 MB
195.00 MB
195.05 MB

B


AMD Association
(OR relative to H1)
frequency
case
control
0
0.5
1
ref
H1 A A 0.59 0.37
H2 G A 0.035 0.040
H3 G A 0.15 0.16
H4 G A 0.0058 0.015
H5 G C 0.097 0.18
H6 G C 0.068 0.15
H7 G C 0.025 0.062
rs800292
Y402H / rs10801555
rs10737680
CFHR1-3 del
rs7542235

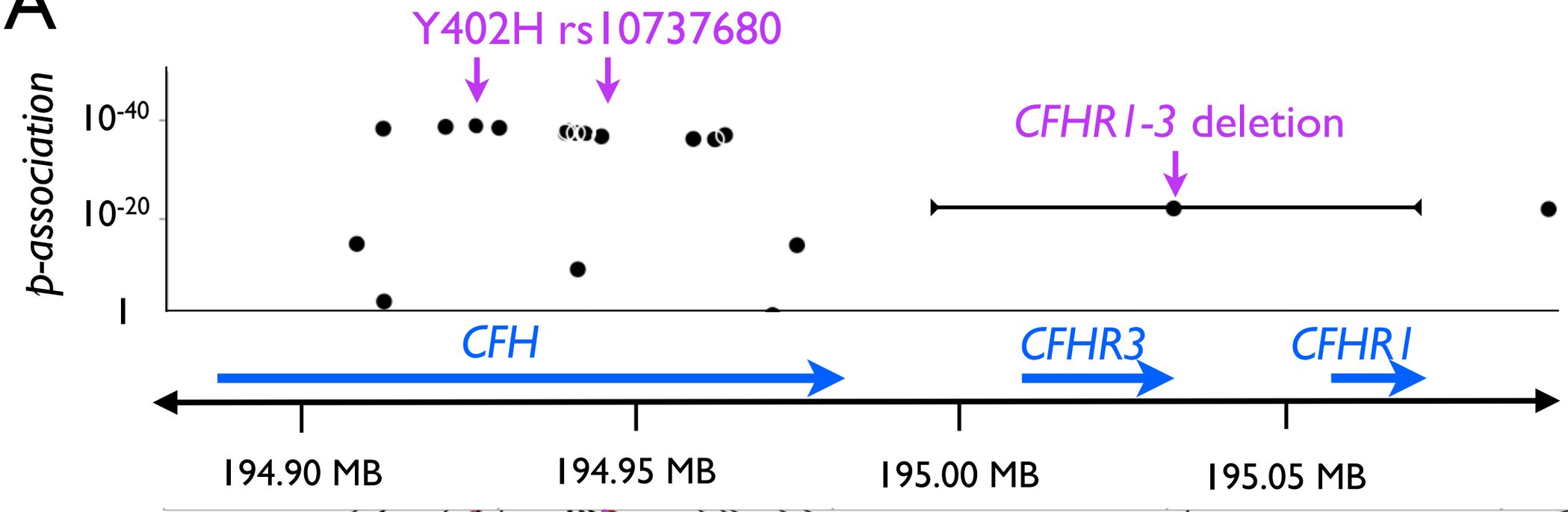
# Available Samples for Sequencing

| | | | Homozygotes | | Heterozygotes | |
|---|---|---|---|---|---|---|
| Haplotype | HF | %A | A | U | A | U |
| *G*GC*C*GGTCTGATA*C*GCGTT*O*A | 0.32% | 0.91 | | | 10 | |
| *G*AT*A*AACACAAAC*A*ATATA*O*A | 46.28% | 0.53 | 236 | 38 | 360 | 134 |
| *G*GC*C*GGTCTGGTA*A*GCGCT*O*A | 15.64% | 0.40 | 12 | 9 | 192 | 72 |
| *G*GC*C*GGTACAAAC*A*ATATA*O*A | 3.78% | 0.36 | 1 | | 45 | 23 |
| *A*GT*C*GGTCTGGTA*C*GCGTA*O*A | 14.62% | 0.27 | 11 | 8 | 114 | 96 |
| *G*GT*C*GGTCTGGTA*C*GCGTA*O*A | 0.47% | 0.25 | | | 4 | 3 |
| *G*GT*C*GGTCTGATA*C*GCGTT*-*G | 12.01% | 0.23 | 2 | 6 | 92 | 86 |
| *A*GC*C*GGTCTGGTA*A*GCGCT*O*A | 1.14% | 0.23 | | | 9 | 6 |
| *G*GC*C*GGTCTGATA*C*GCGTT*-*G | 4.77% | 0.21 | | 3 | 35 | 34 |
| *A*GC*C*GGTCTGGTA*C*GCGTA*O*A | 0.41% | 0.21 | | | 3 | |
| *A*GT*C*GGTCTGATA*C*GCGTT*-*G | 0.56% | 0.16 | | | 3 | 3 |

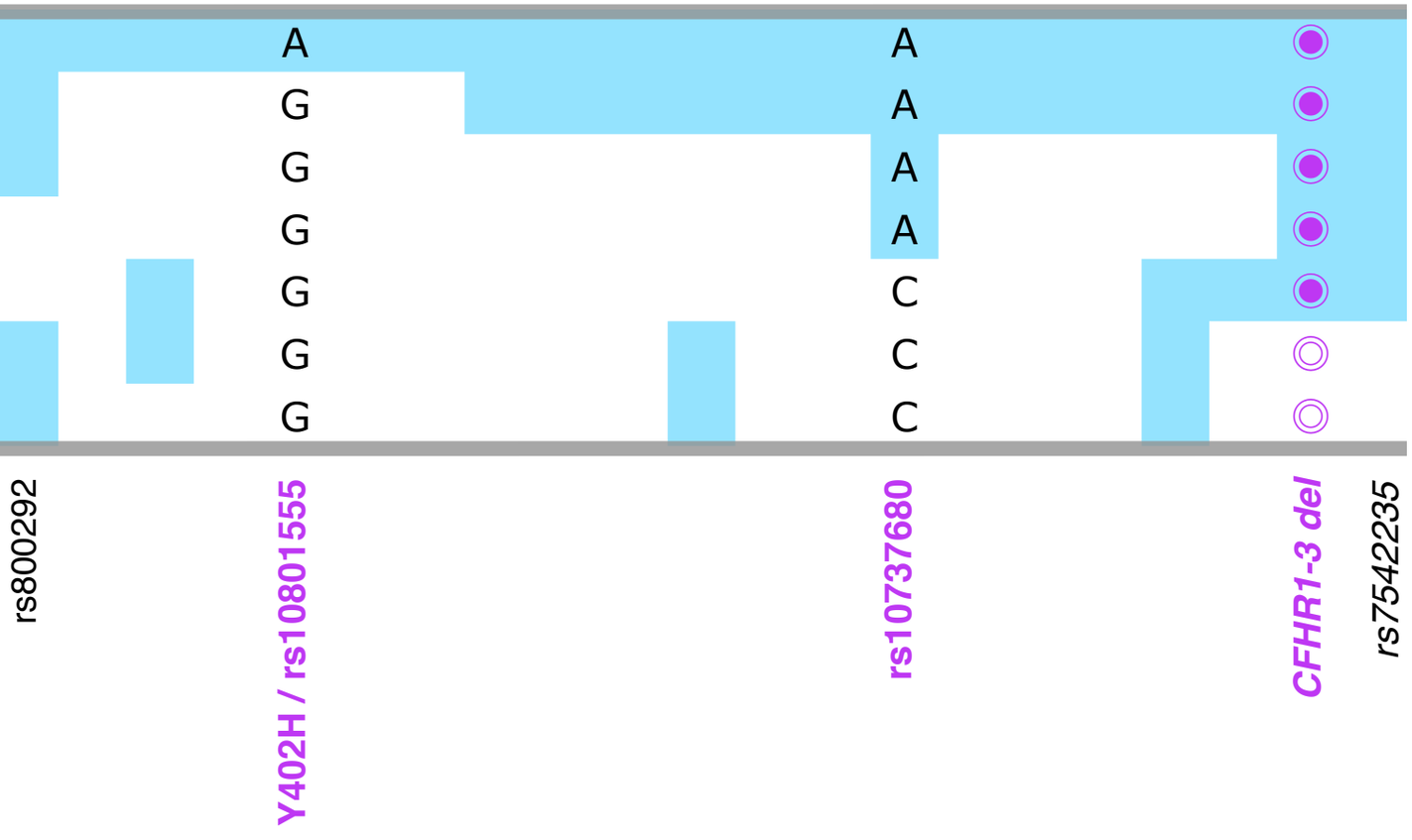
# Future Directions

- Sanger sequencing focused on exons
- Sequence 4 Affected and 4 Unaffected Homozygotes with each common haplotypes (32)
- Sequence available homzygotes with rare haplotypes (4)
- Sequence up to 10 affected heterozygotes with rare haplotypes (46)
- Goal - To identify rare functional variants that explain risk on rare haplotypes, and also more common functional variants that explain common variation.

A


Y402H rs10737680
CFHR1-3 deletion
p-association
10-40
10-20
1
CFH
CFHR3
CFHR1
194.90 MB
194.95 MB
195.00 MB
195.05 MB

B


H1
H2
H3
H4
H5
H6
H7
A
G
G
G
G
G
G
A
A
A
A
C
C
C
rs800292
Y402H / rs10801555
rs10737680
CFHR1-3 del
rs7542235