Using a new bioinformatics tool to impute HLA alleles reveals that three amino acid positions in HLA-DQ and HLA-DR molecules drive Type 1 diabetes risk

Buhm Han, Ph.D.

Department of Convergence Medicine

Asan Medical Center

University of Ulsan College of Medicine

서울아산병원
Asan Medical Center

울산대학교 의과대학
UNIVERSITY OF ULSAN COLLEGE OF MEDICINE

BRIGHAM AND WOMEN'S HOSPITAL
A Teaching Affiliate of Harvard Medical School

HARVARD MEDICAL SCHOOL

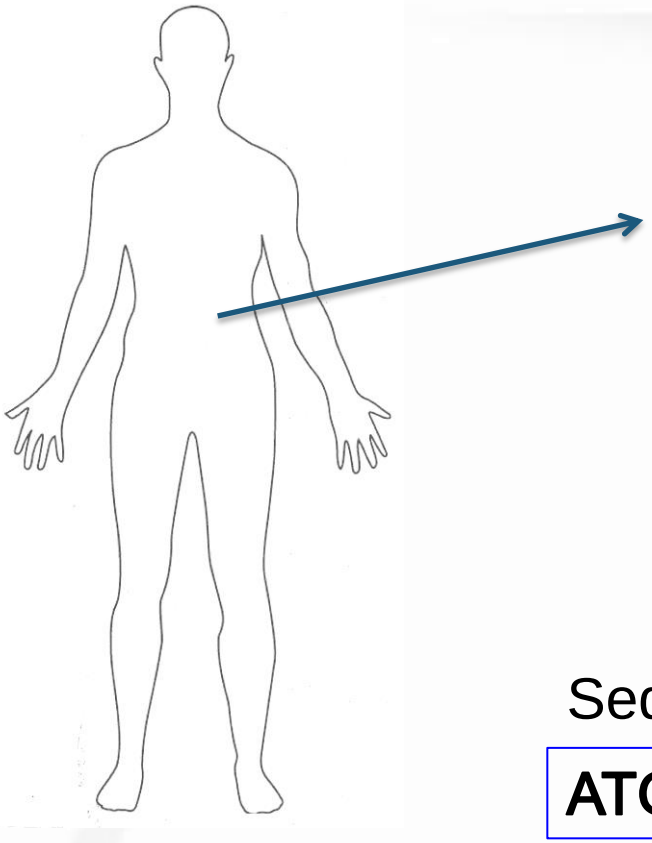
BROAD INSTITUTE

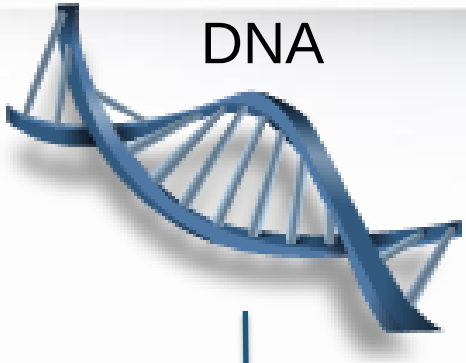
Outline

1. Background (genetic association studies)
2. SNP2HLA
3. T1D MHC fine-mapping

Genetics 101

Human





Sequence

ATGCACATGCAATTCG

Another Sequence

ATGCTCATGAAATTCG

- A series of 3 billion letters where each letter is A, C, T, G
- Humans differ by 0.1% of their DNA
  - Make us all different
  - called "**genetic variants**"
- Majority of differences are **SNPs** (single nucleotide polymorphisms)
  - Single base change
  - ~10 million SNPs in human genome

Genetic variants cause differences in traits

- Humans are different because of genetic variants (also due to environment)
- For example, some SNP may cause people to have different hair

www.23andme.com/gen101/snps

- Some SNPs may cause people to have diseases more easily than others
- How can we find genetic variants (or SNPs) that cause these differences in traits or diseases?
  - Important to uncover the roles of genetics in traits and diseases
- One way is to perform "**association study**"

Association study

cases: people with a disease

controls: people w/o a disease

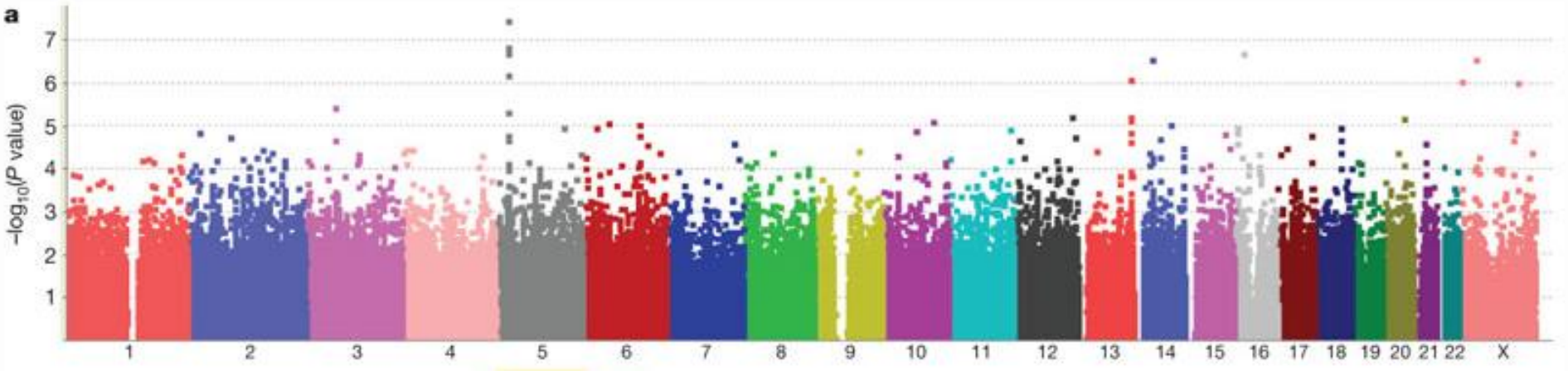
$\hat{p}_X^+ = 0.8$

$\hat{p}_X^- = 0.1$

- We compute correlation (association statistic) between SNP and a disease
  - Association statistic is based on allele freq. difference ($\hat{p}_X^+ - \hat{p}_X^-$)
  - The larger the difference, the higher the correlation
- If correlation is above certain threshold, SNP is associated with a disease
- But, out of many SNPs (10 millions), how do we choose which SNP to test in association study?

Genome-wide Association Studies (GWASs)

- Collect many SNPs (~1 million) over the whole genome
- Compute correlation between each SNP and a disease (perform “association study” on each SNP)
- Find SNPs whose correlations are above the threshold

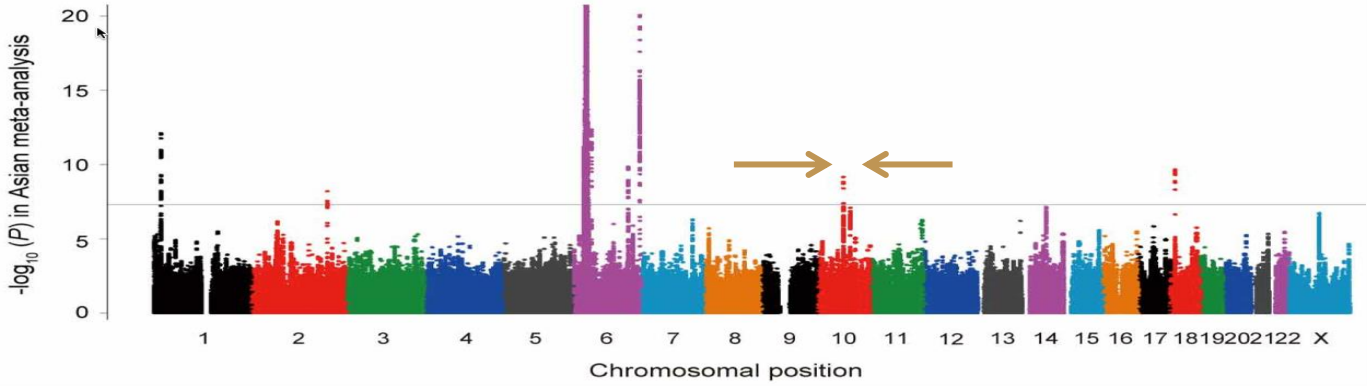


Wang, K., Zhang, H., Ma, D., Bucan, M., et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459, 528-533 (2009).

- A peak in the plot means a strong association between SNP and a disease
- Results of more than 1,600 GWASs have been published

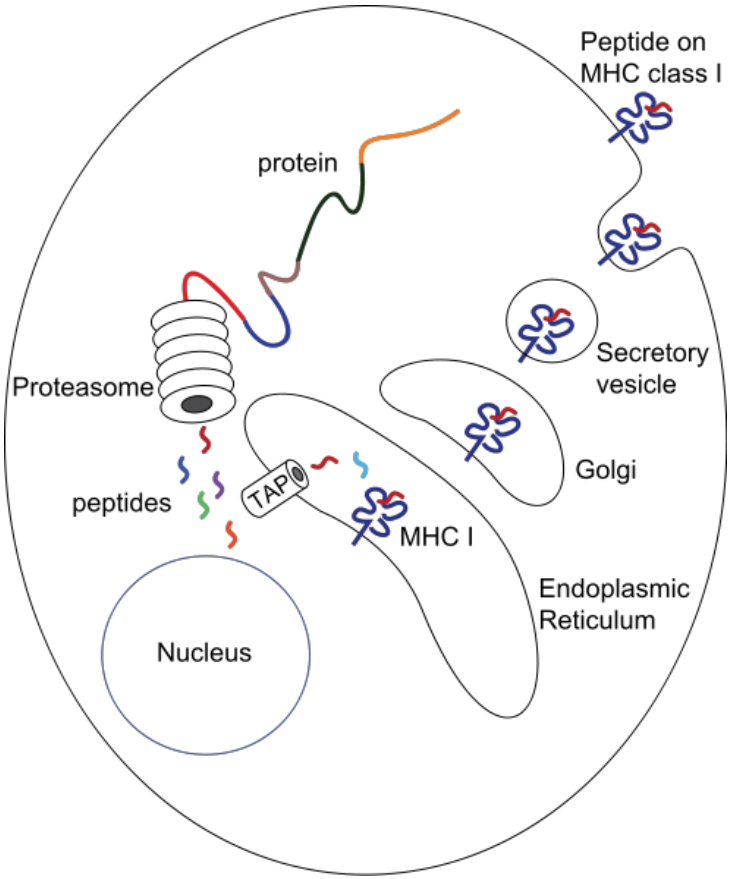
Fine-mapping

- Post-GWAS challenge
- Given an associated region, which gene/variant is actually causal?

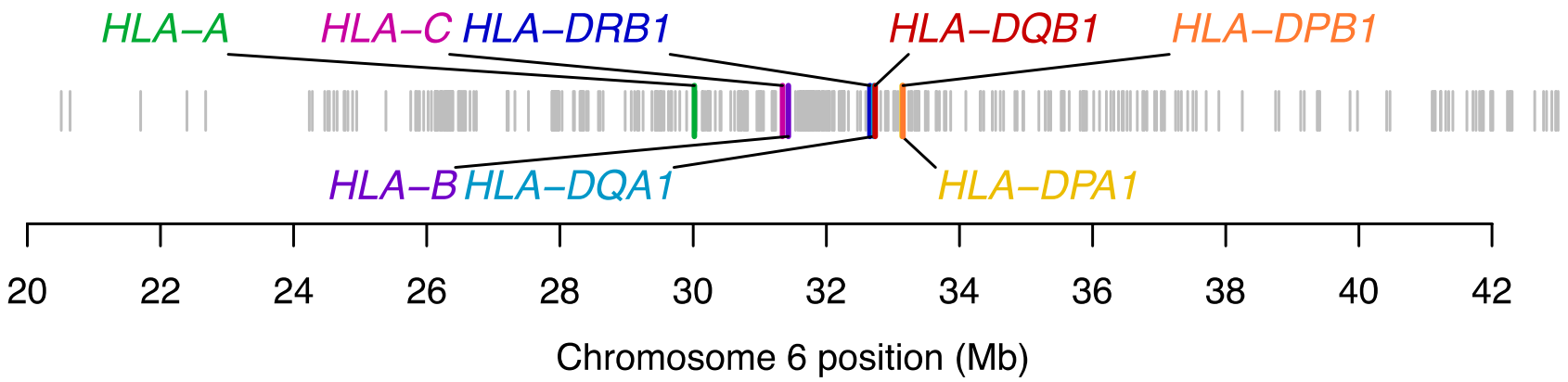


Major Histocompatibility Complex

- Displays antigen peptides to cell surface for T-cells
- Critical role in all immune diseases, including type 1 diabetes



Fine-mapping HLA genes in MHC



- The strongest hit in GWAS for many immune diseases
- 8 classical HLA genes code MHC molecules
  - Which HLA gene is driving the disease?
  - Which amino acid variation is driving?

Fine-mapping

- Association & fine-mapping are difficult
  - Why?

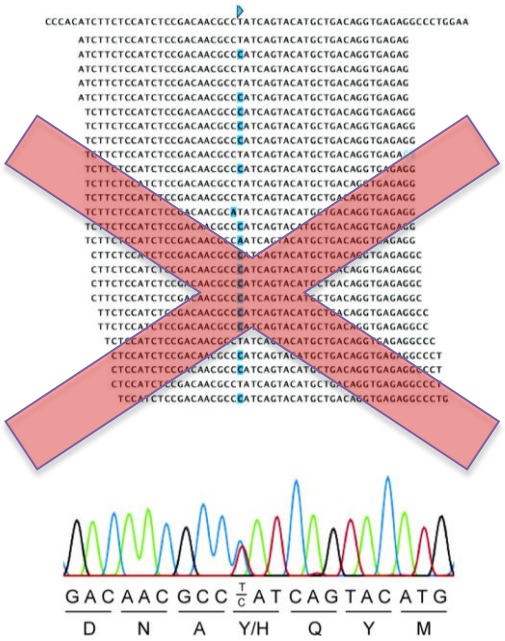
Fine-mapping difficulties in MHC

- HLA genes are highly polymorphic – can't genotype

SNP Microarray

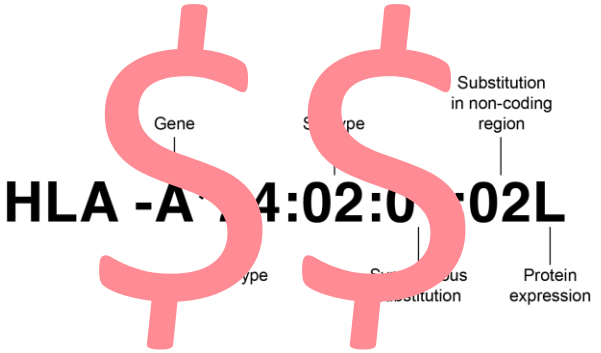
- Flanking sequence doesn't bind
- Only works for intergenic SNPs

Next-gen sequencing



- Doesn't align to reference genome

HLA typing

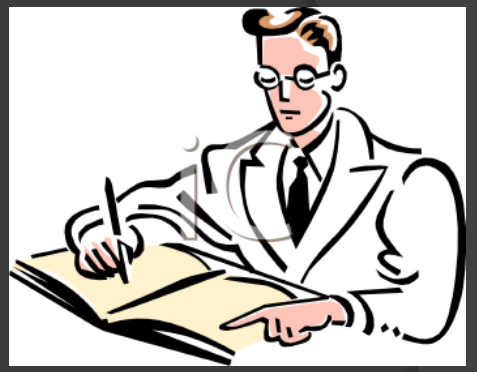


- Expensive
- >$1,000 for 4-digit typing of 8 genes (in Korea)

HLA fine-mapping was practically impossible

Oh, we found the strongest signal at MHC in our GWAS.

This is very interesting.



Well, but we can't figure out which HLA gene is driving the signal and which amino acids are causal.

We can't get the DNA sequence of HLA genes.

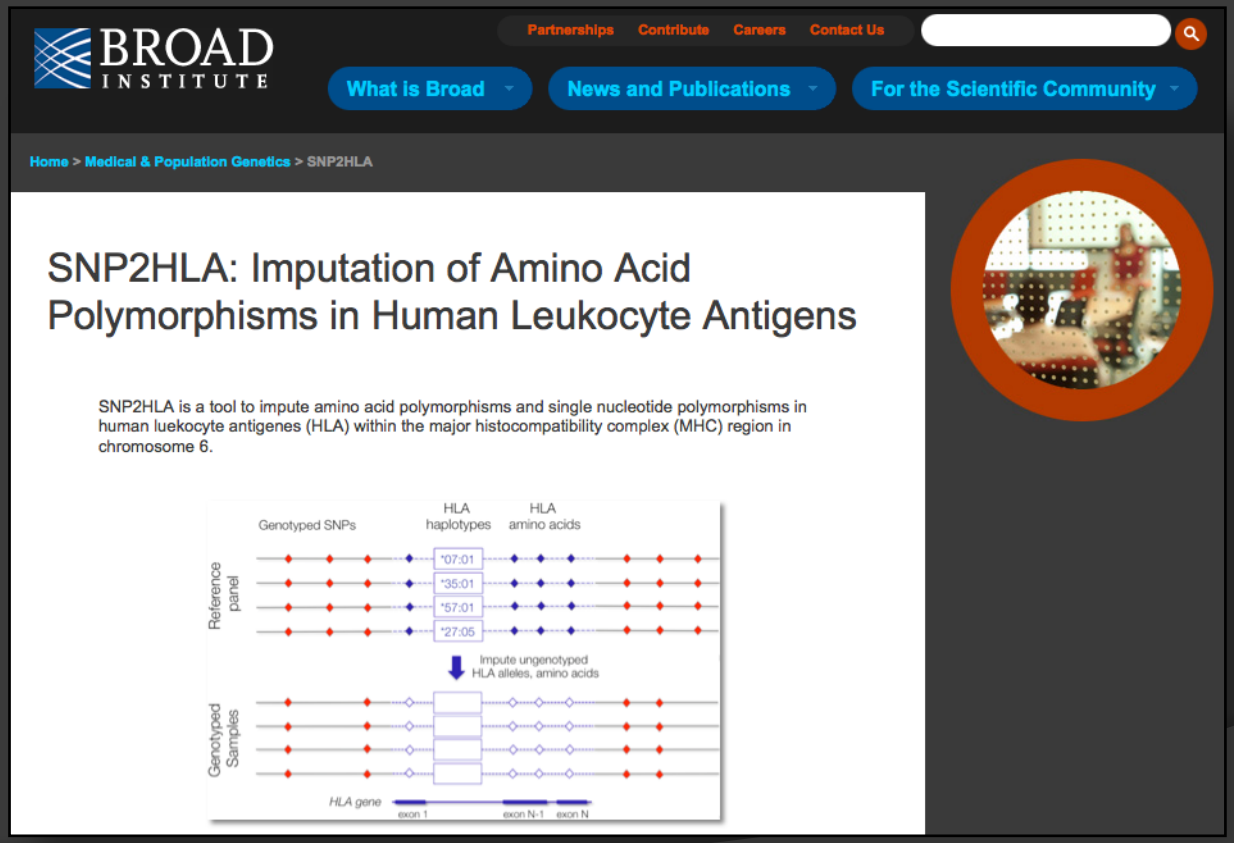
HLA typing will take 10 million dollars.

Outline

1. Background
2. **SNP2HLA**
3. T1D MHC fine-mapping

Our idea: impute HLA genes based on intergenic SNPs!

- SNP2HLA: HLA imputation software
  - 95% accuracy at 4-digit
  - 5,000 European reference panel
  - 900 Asian reference panel

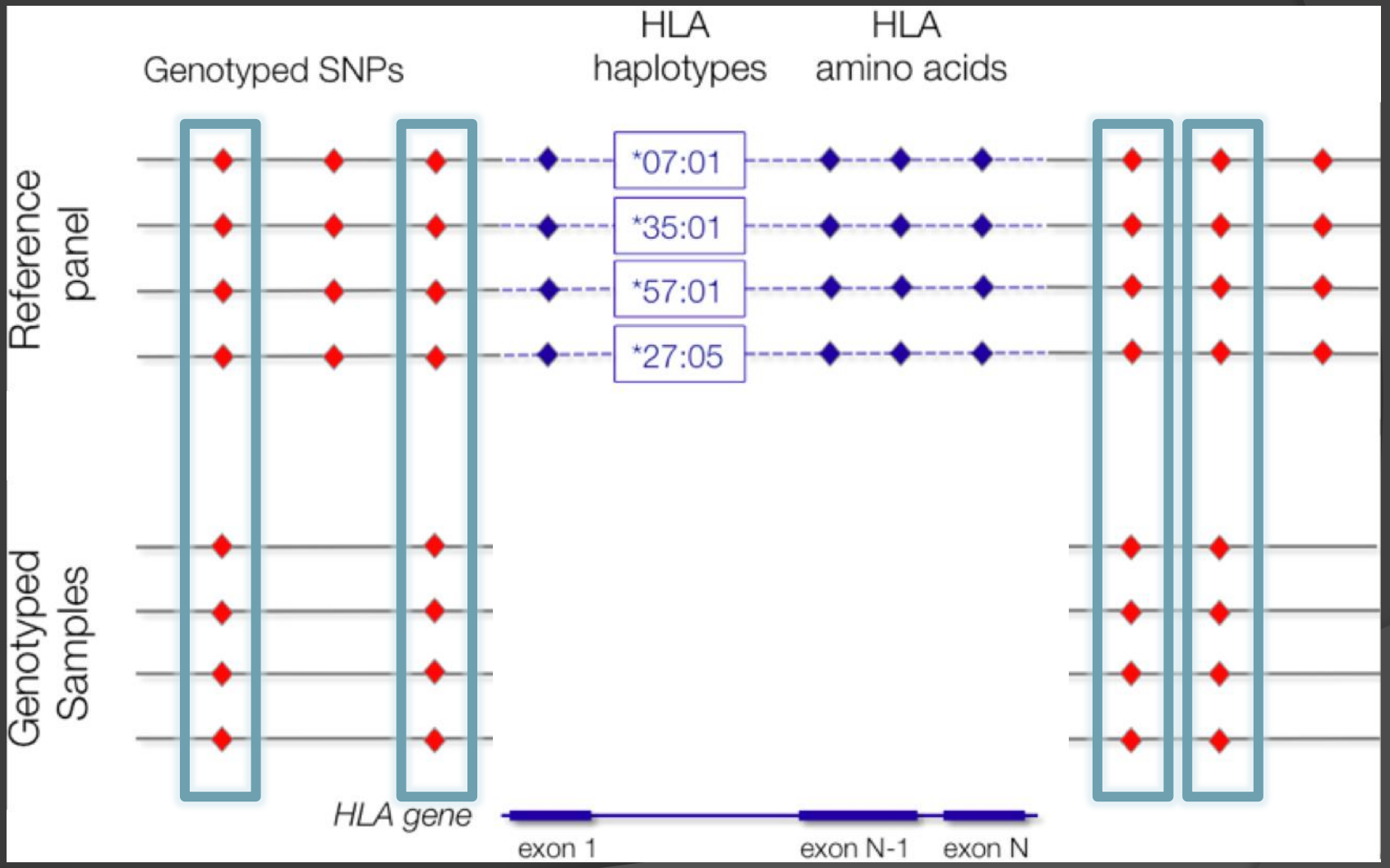


HLA typing

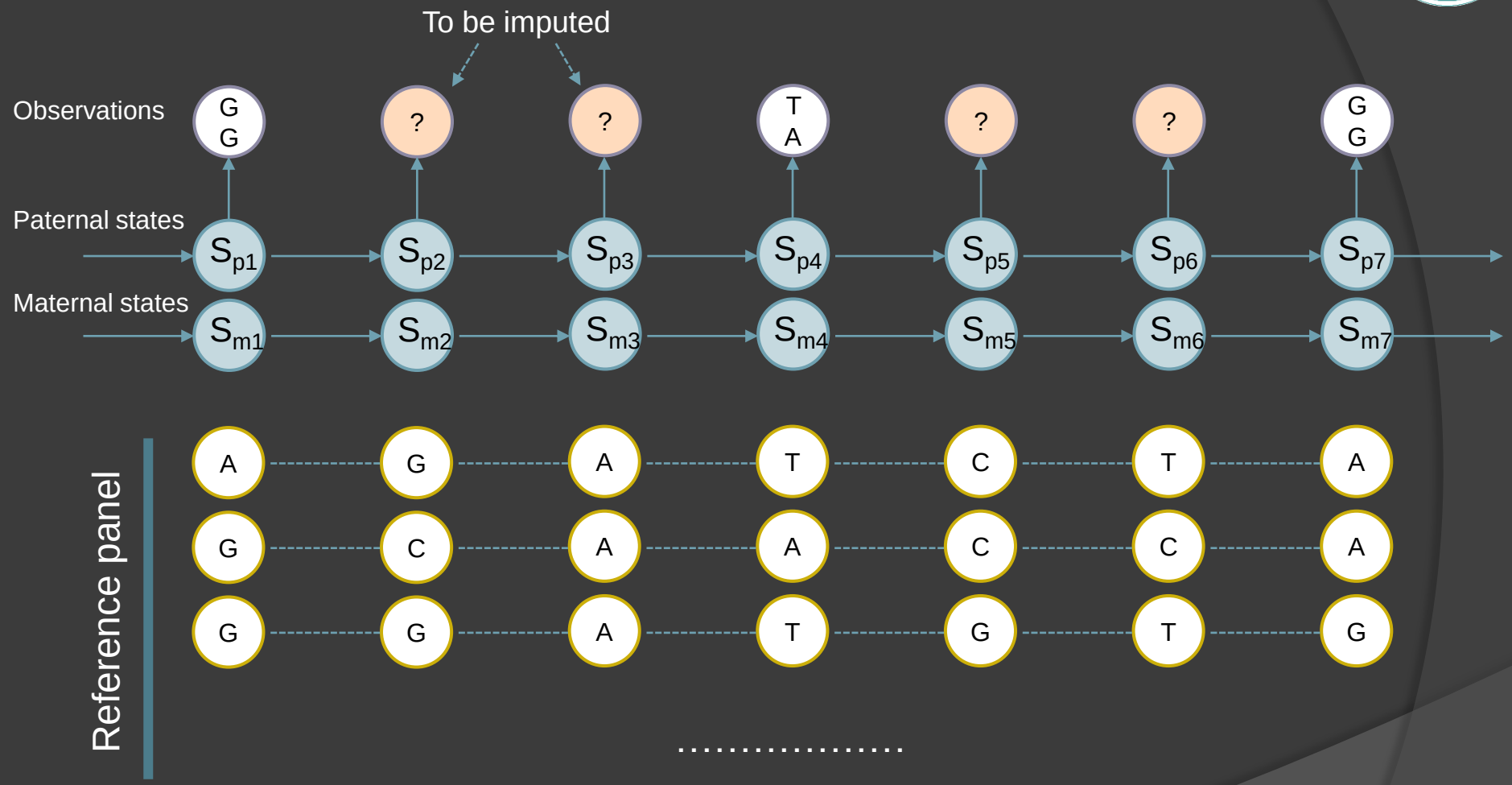
10 million dollars

Method:
Jia* and Han* et al.
PLOS One 2013
Application to RA:
Han et al.
AJHG 2014
Application to PS:
Okada* and Han* et al.
AJHG 2014

SNP2HLA: Overview

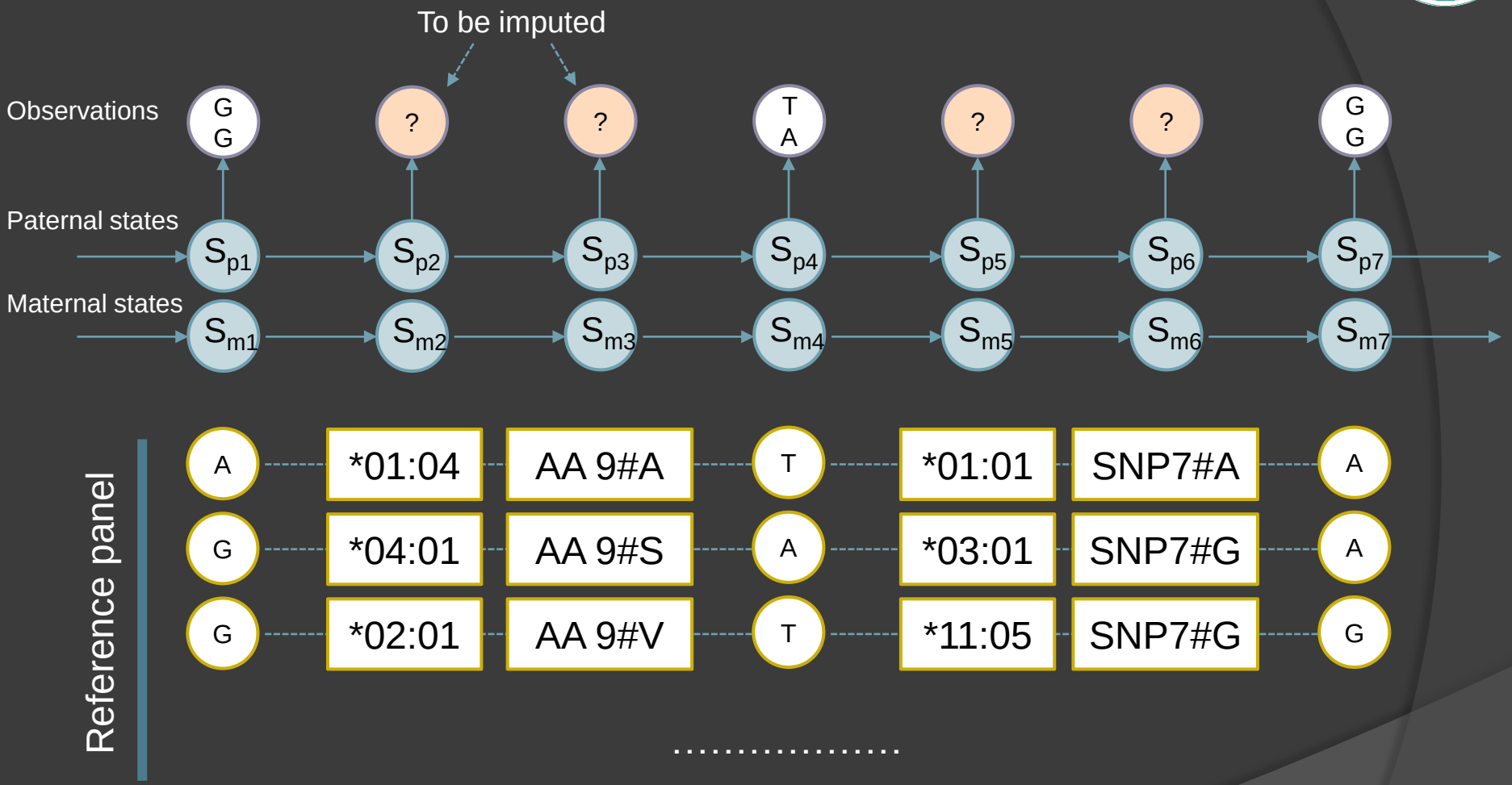


Standard Hidden Markov Model for Imputation



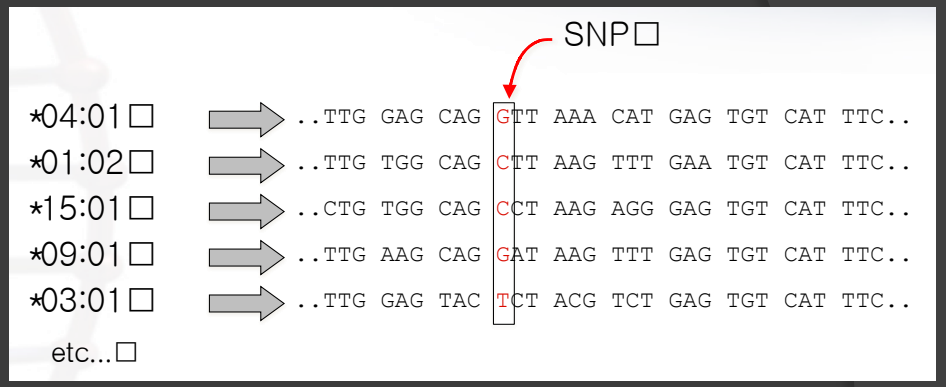
- Transition probability based on recombination rate
- Emission probability based on mutation / error rate

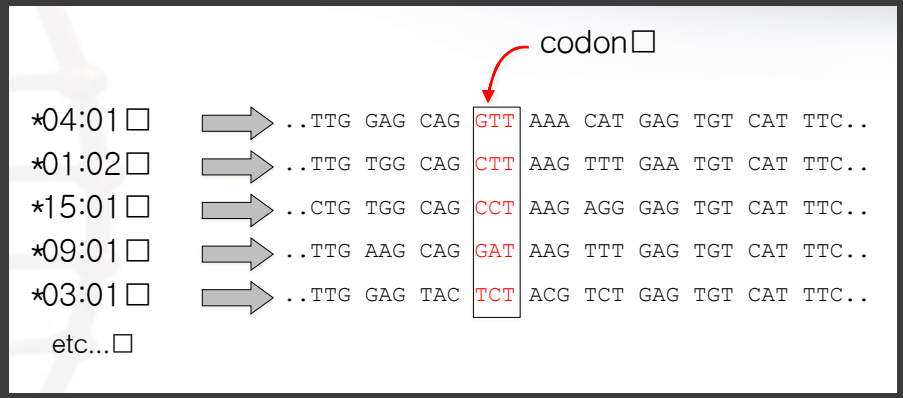
SNP2HLA Hidden Markov Model

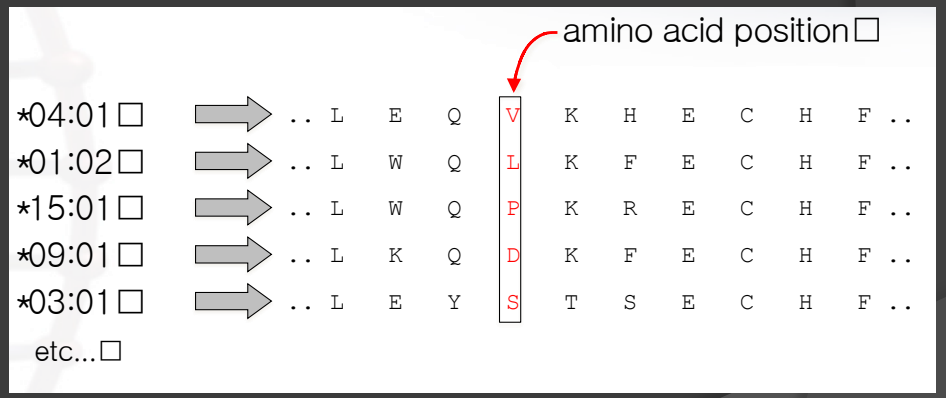


- To account for *k* multi-alleles, define *k* binary markers
- Total >3,000 HLA binary markers

SNP2HLA output allows testing of many features of HLA







- Unbiased & Simultaneous testing of HLA genes / amino acids / and SNPs

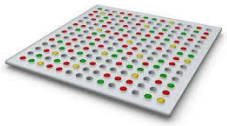
SNP2HLA software (Jia* and Han*, PLOS One 2013)

New tool for imputing HLA genes

HLA Typing

Expensive

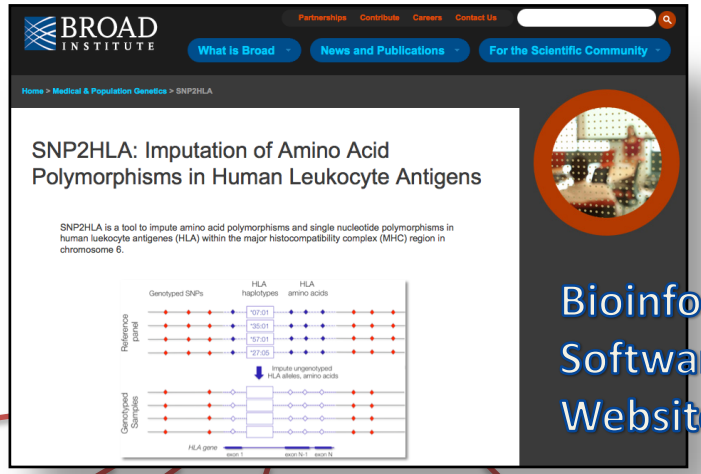
VS



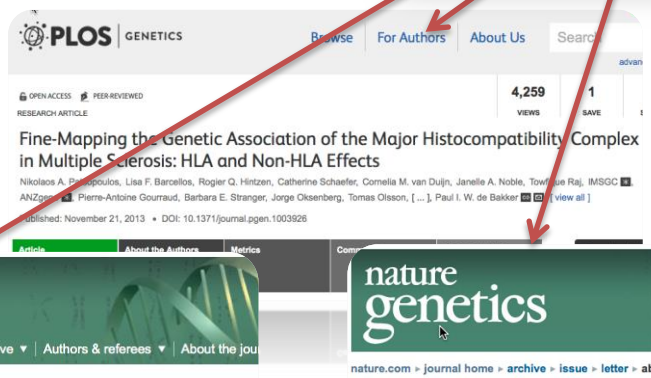
Microarray

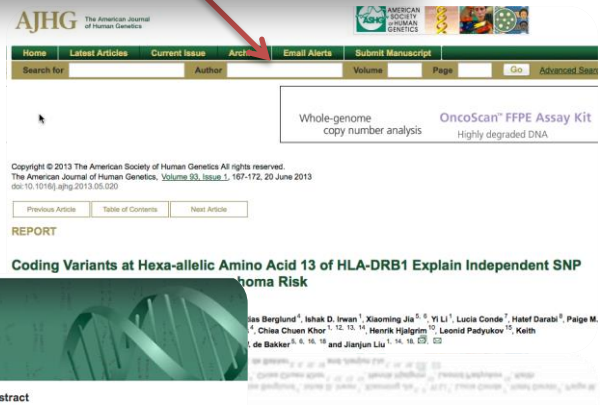
Imputation

Economical



Bioinformatics Software & Website

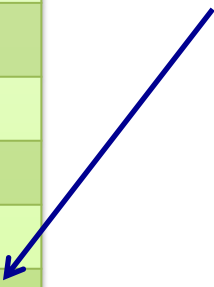




Currently being used by many studies to discover disease-driving HLA alleles

Publications using SNP2HLA since 2014

Trait	Publication	My role
Seronegative RA	**Han** et al. AJHG 2014	Led the analysis
Psoriasis	Okada* and **Han*** et al. AJHG 2014	Led the analysis
Pan-Asian analysis	Hum Mol Gen 2014	Co-author
HCV infection	Gut 2014	
Idiopathic achalasia	Nature Gen 2014	
Seropositive RA (Asian vs European)	Hum Mol Gen 2014	
Pancreatitis induced by thiopurine immunosuppressants	Nature Gen 2014	
Follicular lymphoma	AJHG 2014	
Enteric fever	Nature Gen 2014	Co-author
Systemic lupus erythematosus	Nature Comms 2014	Co-author
Inflammatory bowel disease	Nature Gen 2015	
Narcolepsy protection	AJHG 2015	
Marginal zone lymphoma	Nature Comms 2015	
Alopecia areata	Nature Comms 2015	
Psoriatic arthritis	Nature Comms 2015	
Type 1 diabetes	Nature Genetics 2015	Co-author



nature genetics

Published in 2015 August (2 months ago)

Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk

Xinli Hu[1–6,15], Aaron J Deutsch[1–5,15], Tobias L Lenz[2,7], Suna Onengut-Gumuscu[8], Buhm Han[2,4,9], Wei-Min Chen[8], Joanna M M Howson[10], John A Todd[11], Paul I W de Bakker[12,13], Stephen S Rich[8] & Soumya Raychaudhuri[1–4,14]

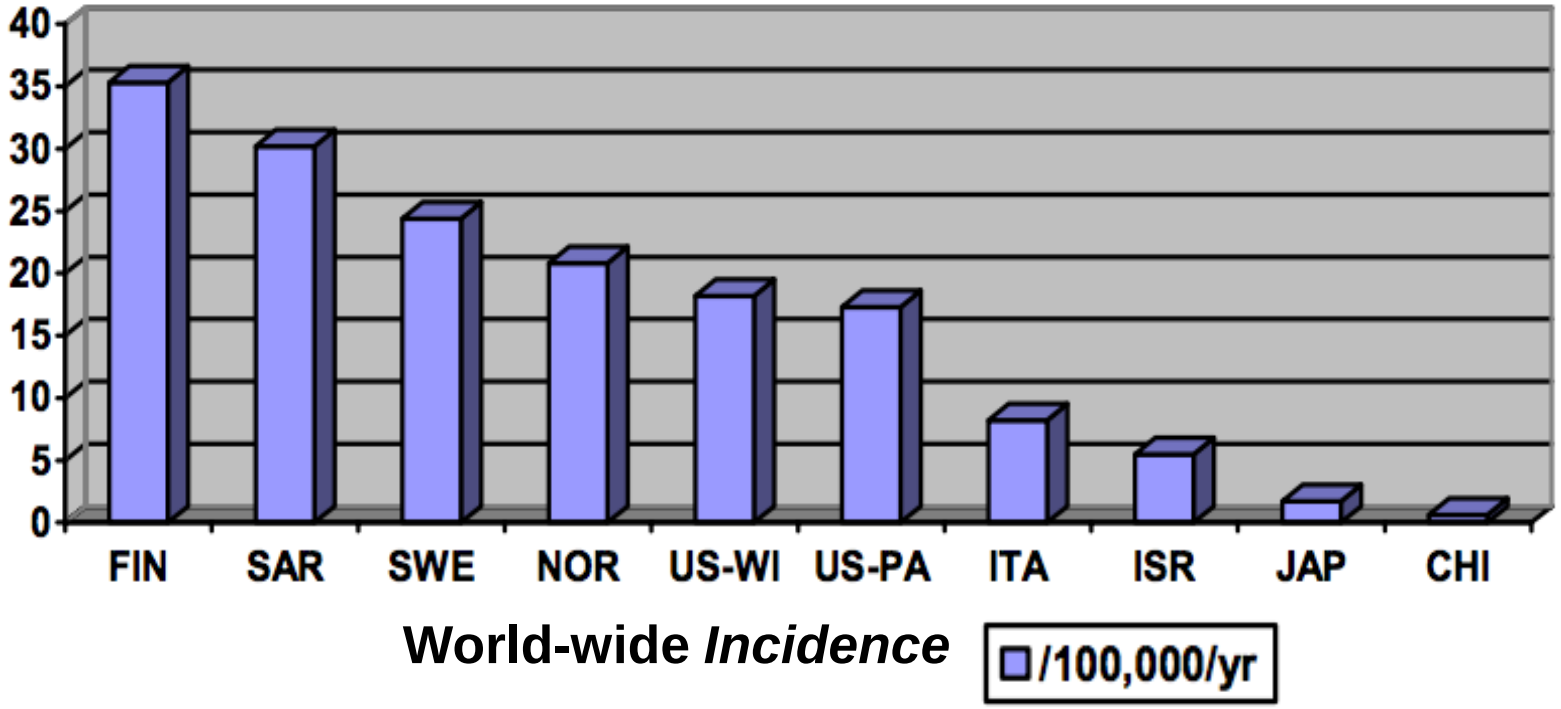
Variation in the human leukocyte antigen (HLA) genes accounts for one-half of the genetic risk in type 1 diabetes (T1D). Amino acid changes in the HLA-DR and HLA-DQ molecules mediate most of the risk, but extensive linkage disequilibrium complicates the localization of independent effects. Using 18,832 case-control samples, we localized the signal to 3 amino acid positions in HLA-DQ and HLA-DR. HLA-DQβ1 position 57 (previously known; $P = 1 \times 10^{-1,355}$) by itself explained 15.2% of the total phenotypic variance. Independent effects at HLA-DRβ1 positions 13 ($P = 1 \times 10^{-721}$) and 71 ($P = 1 \times 10^{-95}$) increased the proportion of variance explained to 26.9%. The three positions together explained 90% of the phenotypic variance in the *HLA-DRB1–HLA-DQA1–HLA-DQB1* locus. Additionally, we observed significant interactions for 11 of 21 pairs of common *HLA-DRB1–HLA-DQA1–HLA-DQB1* haplotypes ($P = 1.6 \times 10^{-64}$). HLA-DRβ1 positions 13 and 71 implicate the P4 pocket in the antigen-binding groove, thus pointing to another critical protein structure for T1D risk, in addition to the HLA-DQ P9 pocket.

Outline

1. Background
2. SNP2HLA
3. **T1D MHC fine-mapping**

Type 1 diabetes

- World-wide prevalence < 1%; ~ 1 million in US



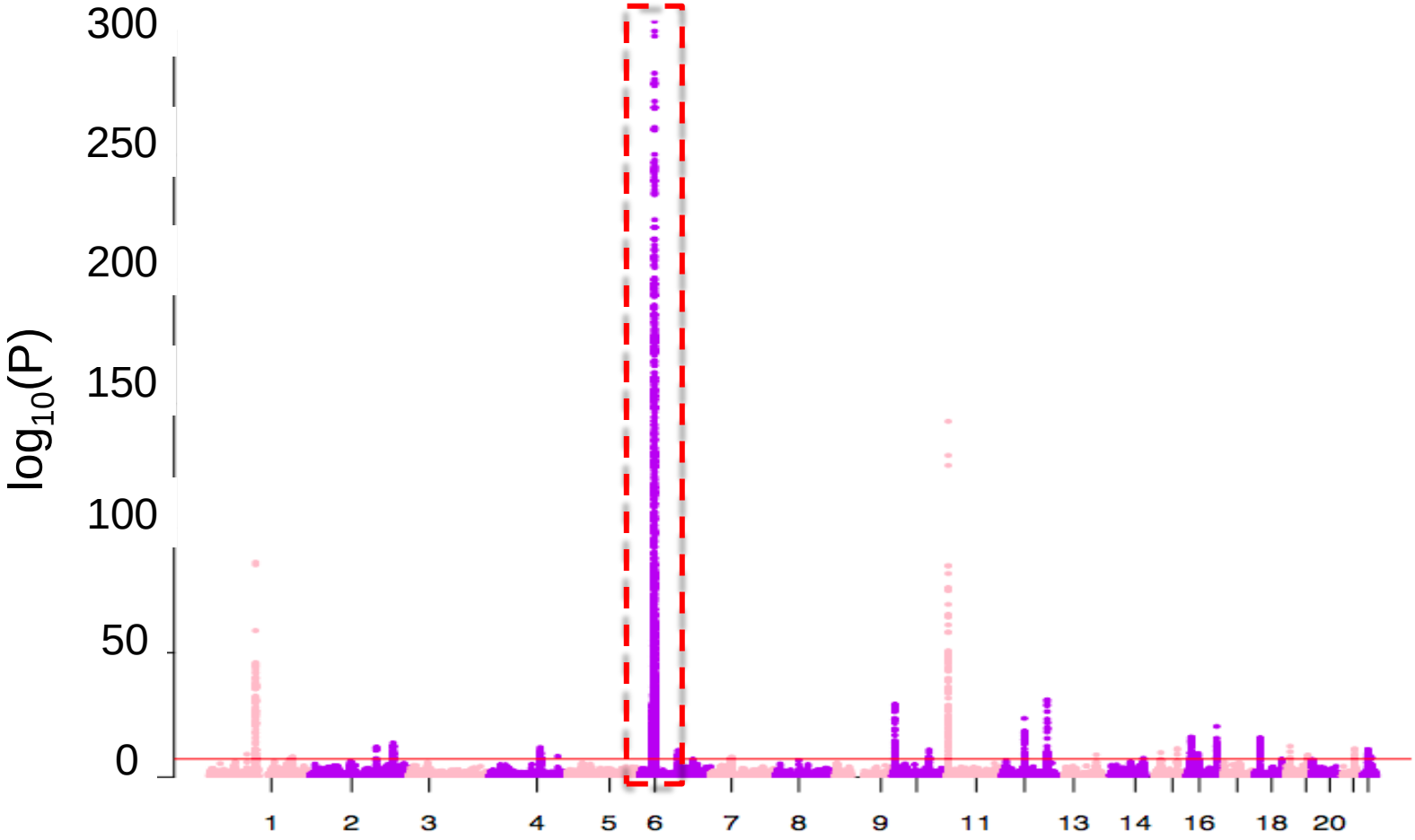
- Early onset; no gender bias
- Treatment: insulin replacement
- ~15 billion$ annual treatment

http://www.who.int/genomics/about/Diabetis-fin.pdf

HLA in T1D

Narrow-sense heritability ~74%

- HLA ~35% (Speed et al. 2012 *AJHG*)
- Main locus: *HLA-DRB1-DQA1-DQB1*



Hypothesis: amino acid polymorphisms drive haplotypic association

Known haplotypic associations

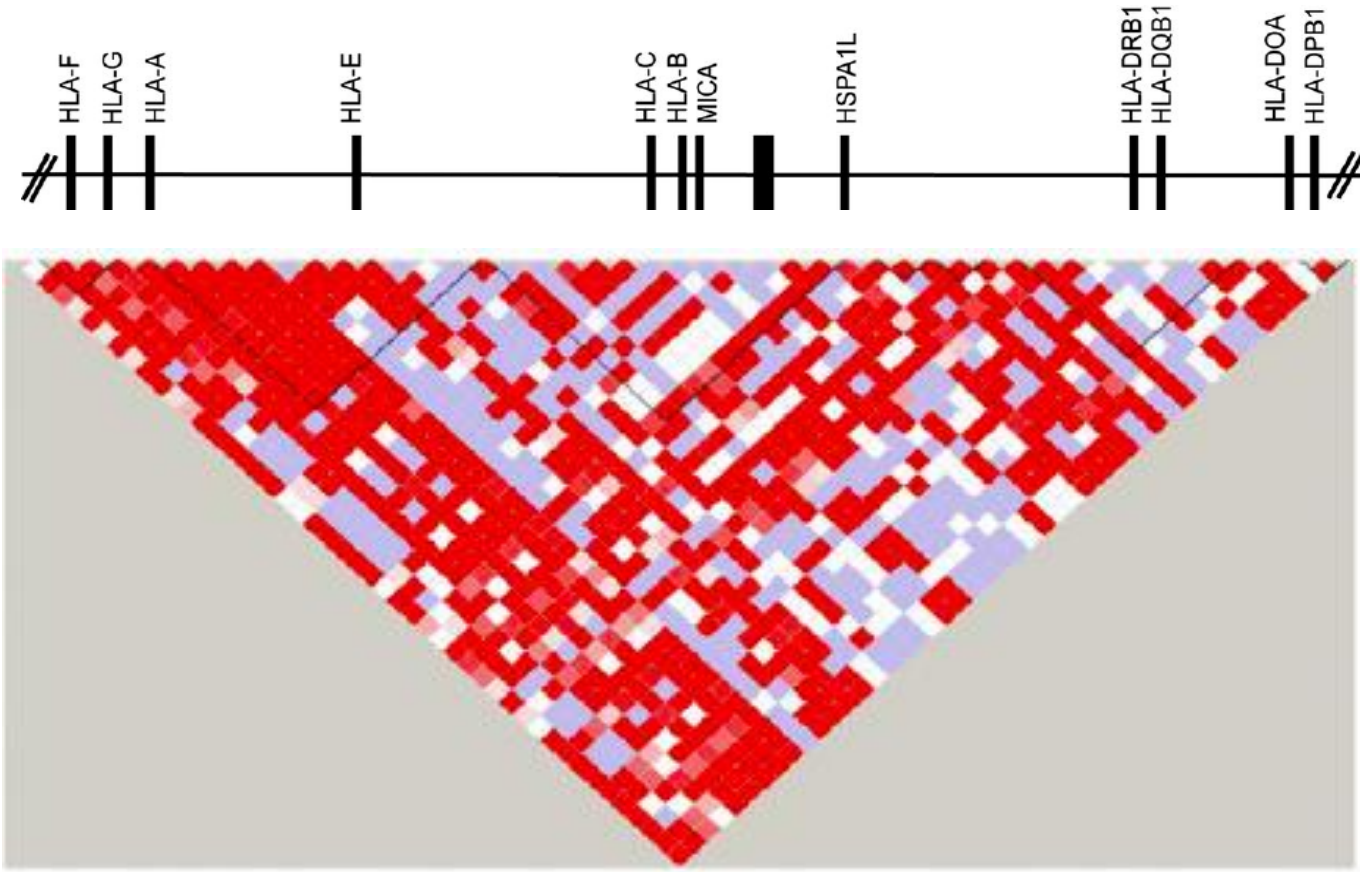
DRB1-DQA1-DQB1	OR	*p*-Value
01:01-01:01-05:01	0.71	0.047
01:03-01:01-05:01	0.15	0.046
03:01-05:01-02:01	3.64	2×10^{-22}
04:01-03:01-03:01	0.35	4×10^{-04}
04:01-03:01-03:02	8.39	6×10^{-36}
04:02-03:01-03:02	3.63	3×10^{-04}
04:03-03:01-03:02	0.27	0.017
04:04-03:01-03:02	1.59	0.049
04:05-03:01-03:02	11.37	4×10^{-05}
04:07-03:01-03:01	0.11	6×10^{-04}
07:01-02:01-02:01	0.32	2×10^{-09}
07:01-02:01-03:03	0.02	4×10^{-12}
08:03-06:01-03:01	0	0.047
11:01-05:01-03:01	0.18	3×10^{-10}
11:03-05:01-03:01	0.25	0.024
11:04-05:01-03:01	0.07	3×10^{-06}
12:01-05:01-03:01	0.29	0.031
13:01-01:03-06:03	0.13	4×10^{-11}
13:02-01:02-06:09	0	0.047
13:03-05:01-03:01	0.08	0.003
14:01-01:01-05:03	0.02	1×10^{-06}
15:01-01:02-06:02	0.03	2×10^{-29}
15:01-01:02-06:03	0	0.047
Overall significance		$\mathbf{5 \times 10^{-124}}$

Noble *et al.* 2007. *Cold Spring Harb Perspect*

- Long-standing hypothesis: amino acids in the peptide-binding grooves alter antigen presentation
- Classical alleles defined by ***combinations*** of amino acid residues
- Best known amino acid: DQβ1#57 (Todd *et al.* 1987. *Nature*); cannot explain all the risk

Problems: polymorphism and LD

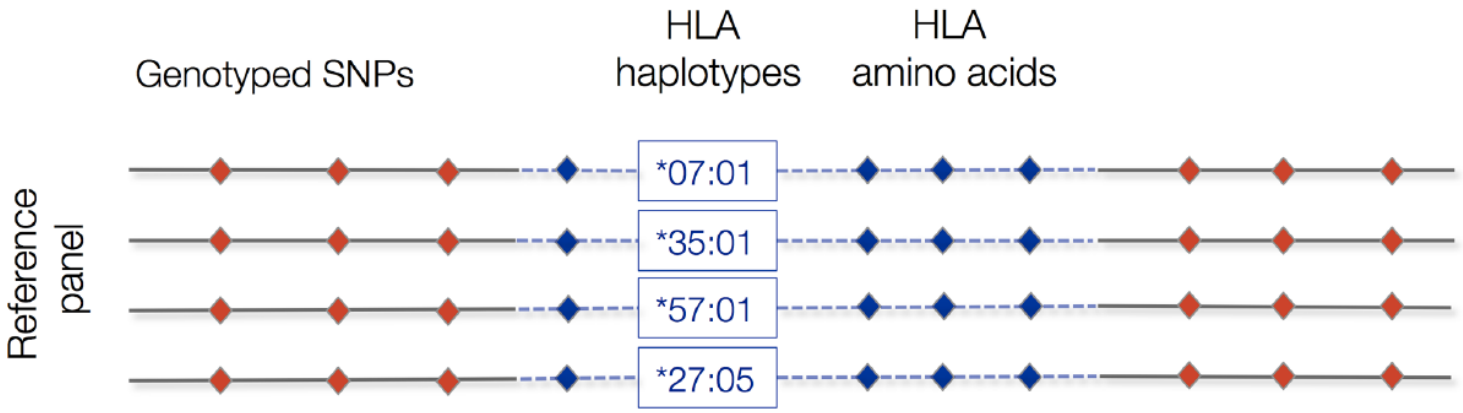
- Difficult to fine-map
  - Highly polymorphic (~12,000 alleles)
  - Extensive linkage disequilibrium



Adapted from Petersdorf. 2013. *Blood*

HLA Imputation

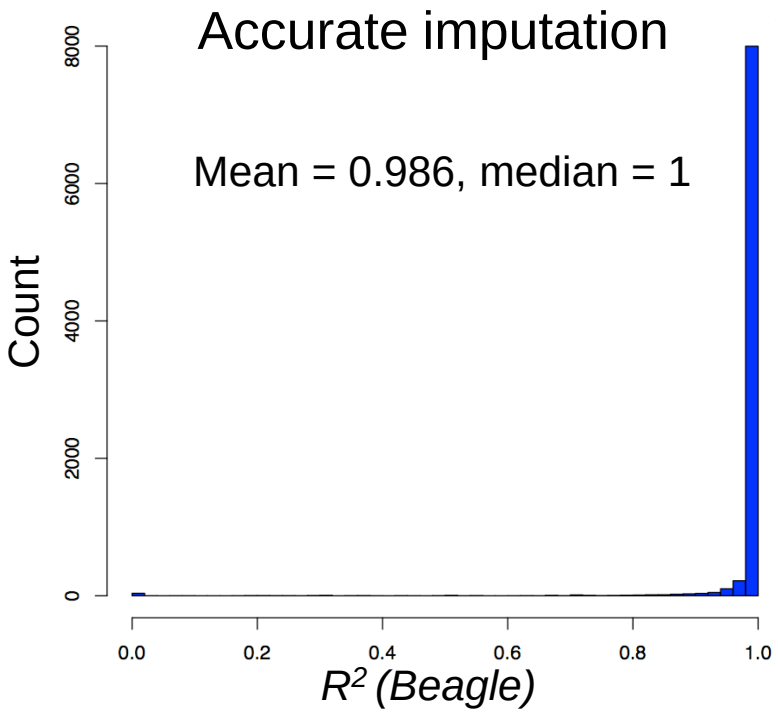
- Impute 2- and 4-digit classical alleles & amino acids
  - *SNP2HLA*
  - T1DGC reference panel (5225 typed European samples)

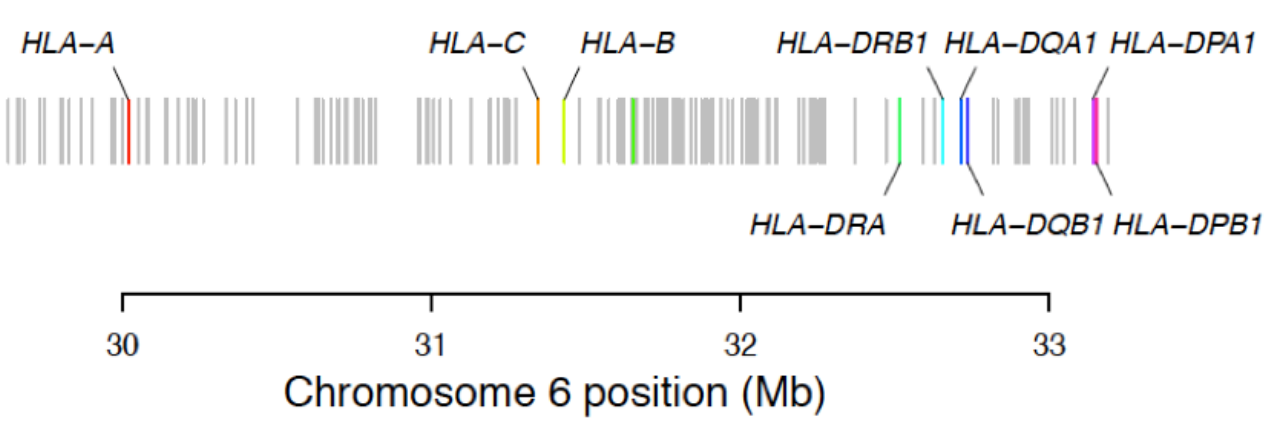


Jia & Han *et al.* 2013. *PLoS ONE*

Dataset (T1DGC)

- 18,832 samples from UK
  - 8,095 cases
  - 10,737 controls
  - 13 geographical regions
- Eight HLA genes
- 8617 binary variants (>0.05%)
- 260 classical alleles
- 399 amino acid positions





Statistical framework

- Logistic regression

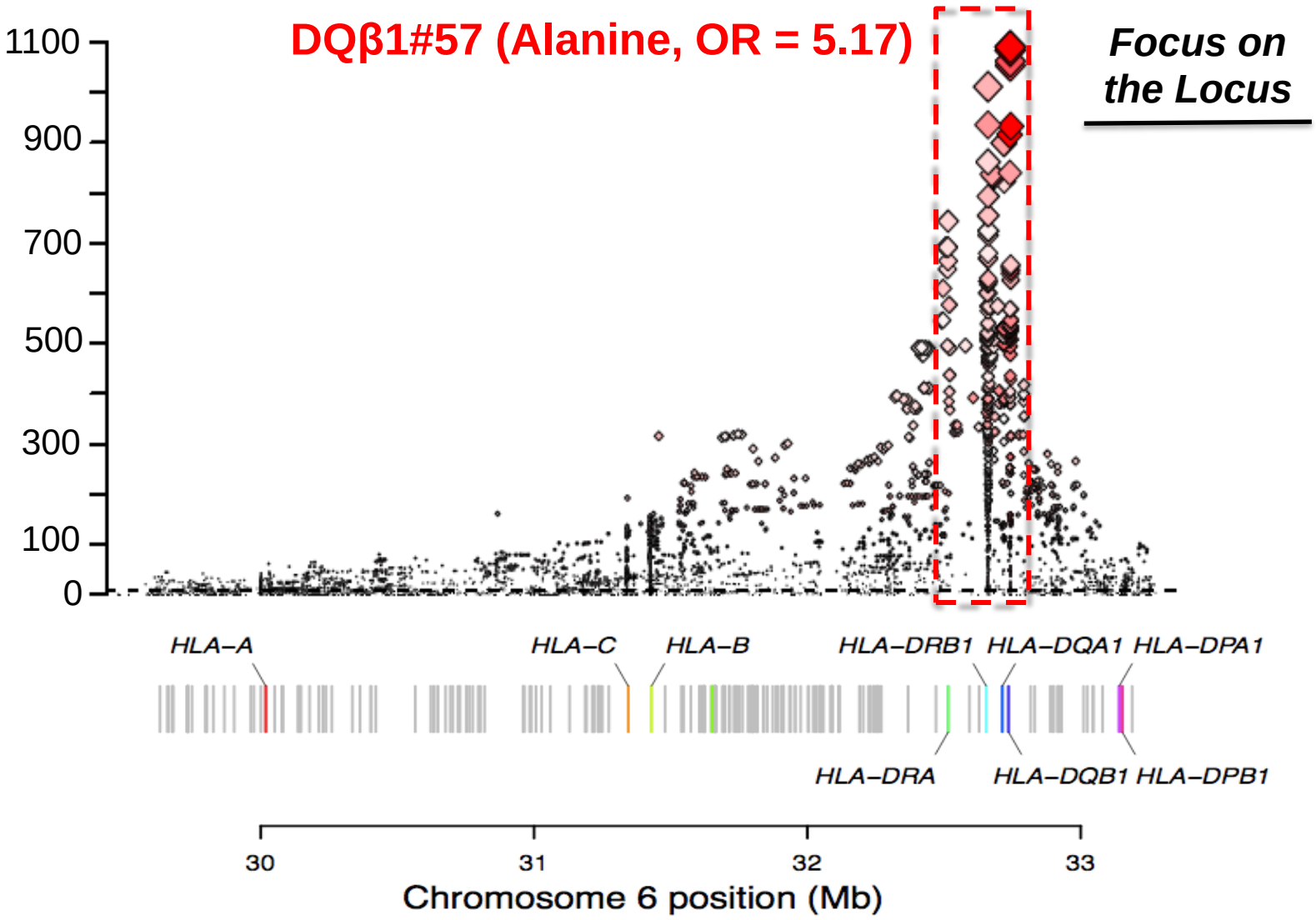
null

$$\ln(odds_j) \sim b_o + \sum_{j=1}^{m-1} b_{1,j} x_j$$

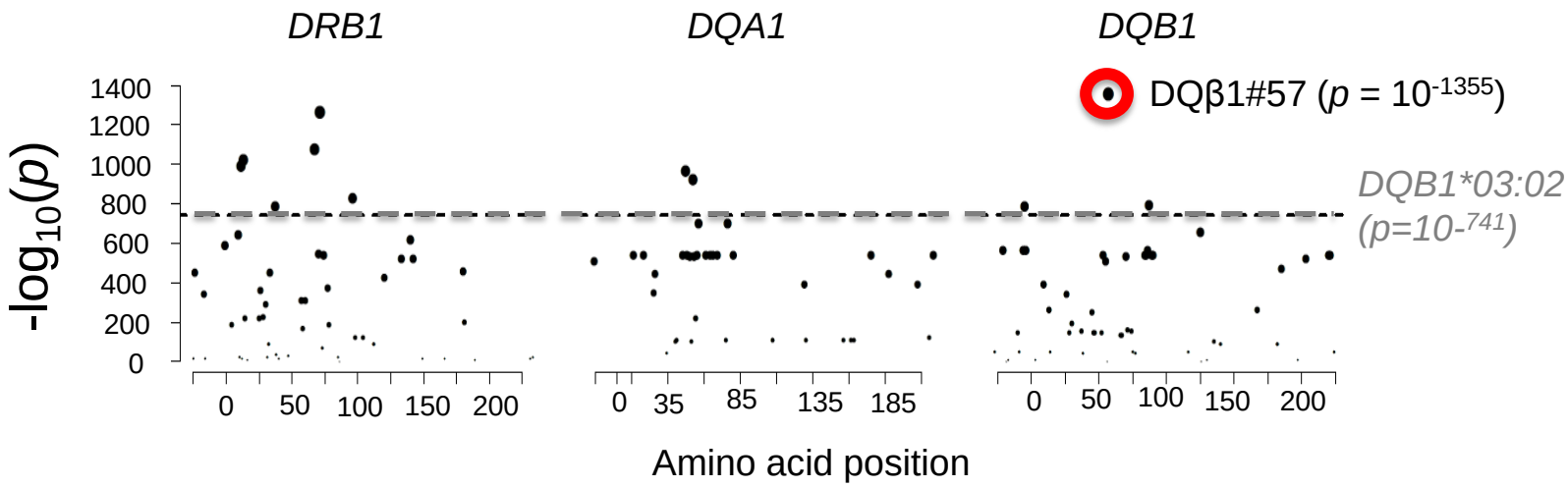
x = genotype/dosage

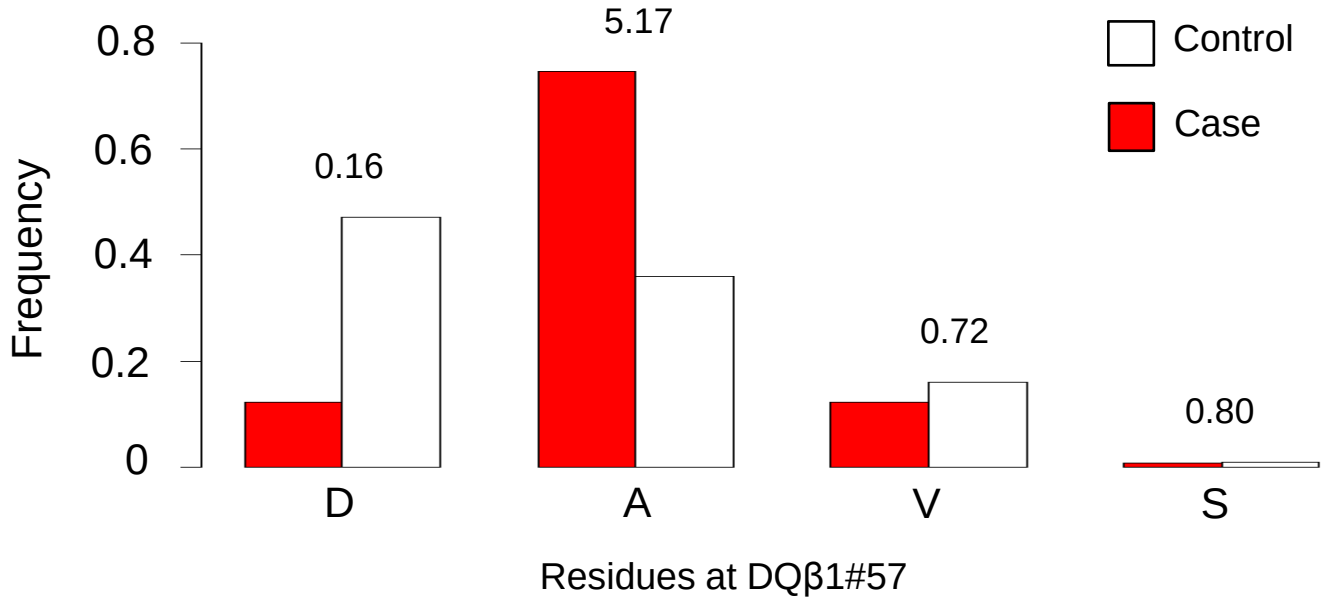
$$\Delta Deviance_{alt-null} = -2\ln(likelihood_{alternative} / likelihood_{null})$$

Top signal - DQβ1#57 (best-known)

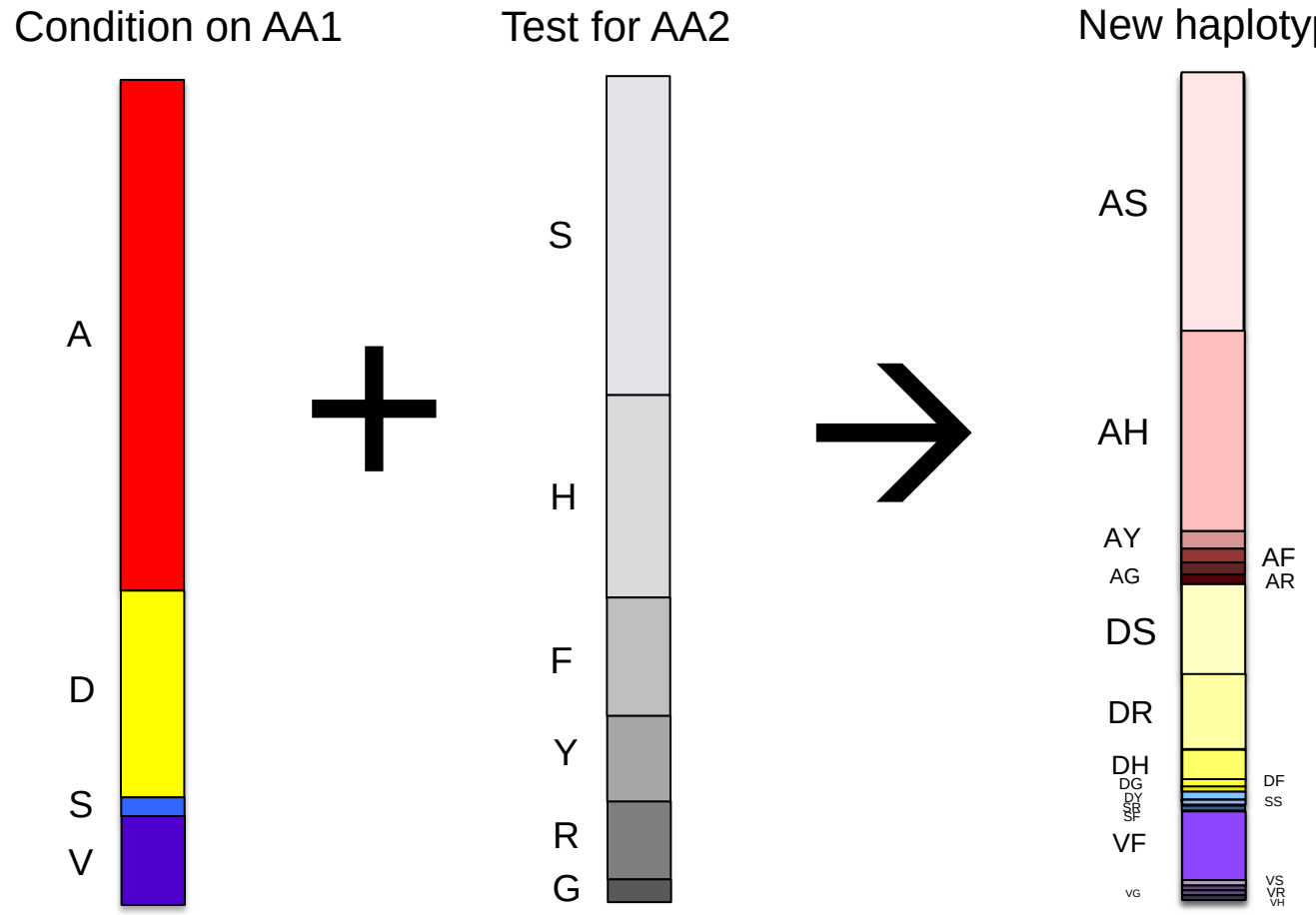


Amino acid positions (omnibus test)





Conditional analysis by forward-search

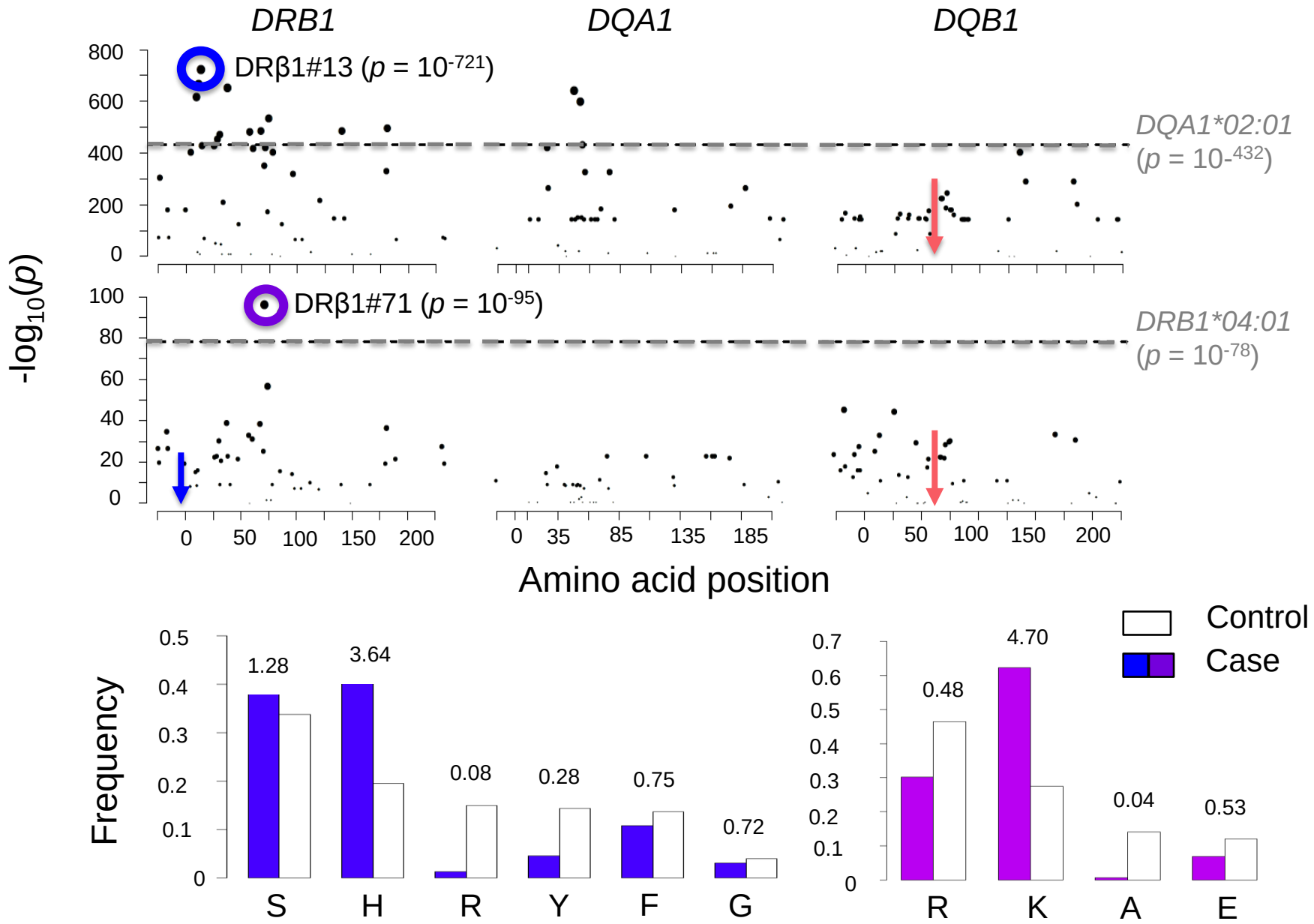


Null model, df = 3

Alternative model, df = 19

$$\Delta Deviance_{alt-null} = -2\ln(likelihood_{alternative} / likelihood_{null})$$ *Δdf = 16*

DRβ1#13 & #71


DRB1
DQA1
DQB1
DRβ1#13 (p = 10-721)
DQA1*02:01
(p = 10-432)
DRβ1#71 (p = 10-95)
DRB1*04:01
(p = 10-78)
-log10(p)
Amino acid position
Frequency
Control
Case
1.28
3.64
0.08
0.28
0.75
0.72
S
H
R
Y
F
G
0.48
4.70
0.04
0.53
R
K
A
E

To confirm the associations

- Genotyping errors/imputation uncertainty could introduce noise as signal strength decreases

- Forward-search may converge on local minima
  - Solution: exhaustively test all combinations

Exhaustive testing

DQβ1#57+DRβ1#13: Best of 9,870 pairs

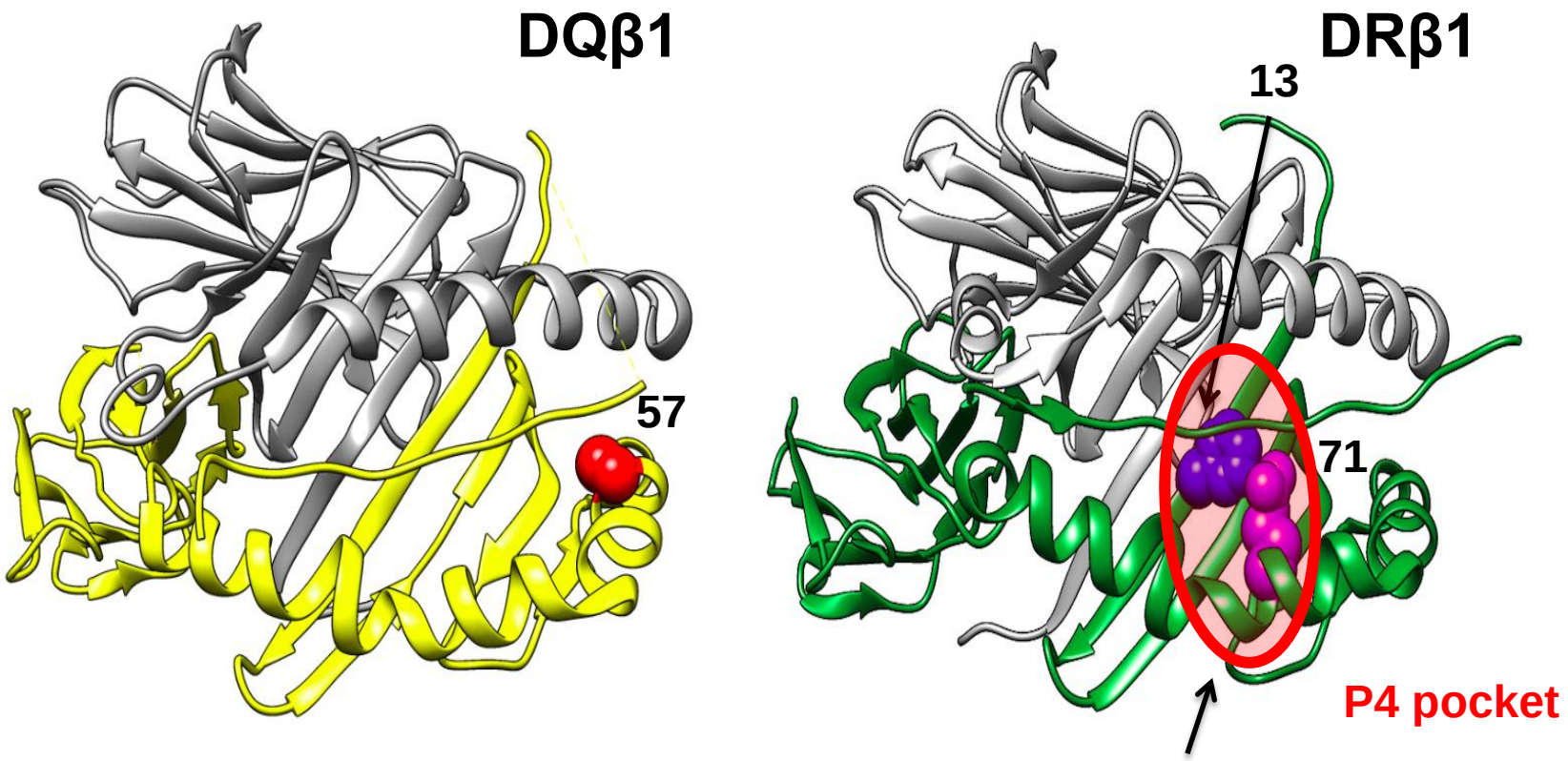
AA1	AA2	DeltaDeviance	df	log10(p)
AA_DQB1_57	AA_DRB1_13	9661.011751	19	-2071.62
AA_DQB1_57	AA_DRB1_11	9404.827094	18	-2017.46
AA_DQB1_57	AA_DRB1_37	9324.046636	15	-2004.12
AA_DQA1_47	AA_DQB1_57	9245.520421	10	-1994.36
AA_DQB1_57	AA_DRB1_9	9111.50804	9	-1966.80
AA_DQA1_52	AA_DQB1_57	9032.194409	8	-1951.13
AA_DQB1_57	AA_DRB1_74	8768.560717	15	-1883.67
AA_DQB1_57	AA_DRB1_181	8564.448428	8	-1849.63
AA_DQB1_57	AA_DRB1_67	8533.010207	11	-1838.30
AA_DQB1_57	AA_DRB1_140	8501.312677	7	-1837.49

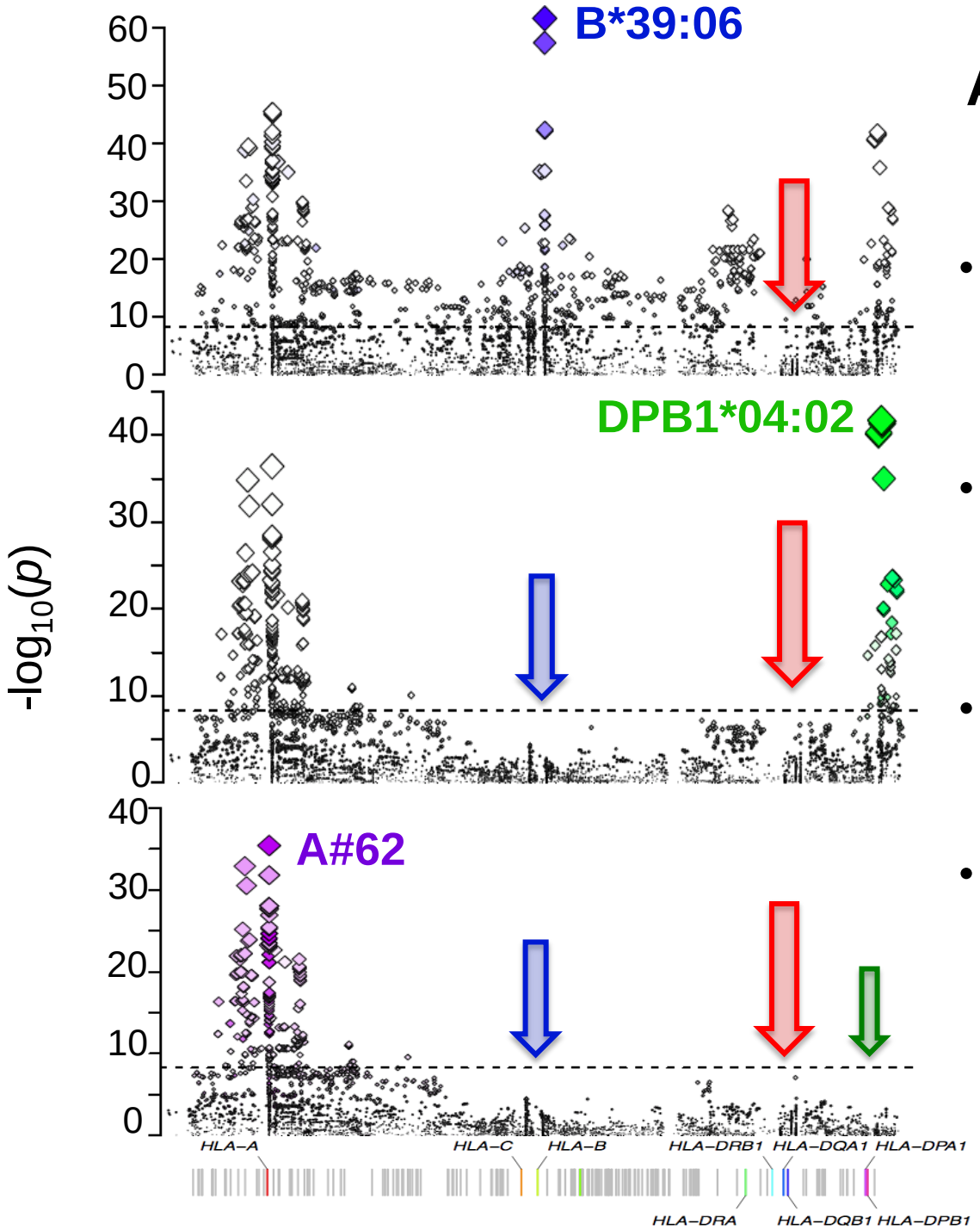
DQβ1#57+DRβ1#13+DRβ1#71: Best of 457,450 trios

AA1	AA2	AA3	DeltaDeviance	df	log10(p)
AA_DQB1_57	AA_DRB1_71	AA_DRB1_13	10148.52492	31.00	-2161.52
AA_DQB1_57	AA_DRB1_86	AA_DRB1_13	10124.62638	29.00	-2158.89
AA_DQB1_-18	AA_DRB1_71	AA_DRB1_37	10045.15659	25.00	-2146.85
AA_DQB1_57	AA_DRB1_74	AA_DRB1_11	9987.049638	31.00	-2126.56
AA_DQB1_75	AA_DQB1_-18	AA_DRB1_13	9938.438515	26.00	-2122.43
AA_DQB1_74	AA_DQB1_-18	AA_DRB1_13	9943.814444	27.00	-2122.30
AA_DQB1_26	AA_DQB1_-10	AA_DRB1_13	9937.416871	26.00	-2122.21
AA_DQB1_26	AA_DQB1_-18	AA_DRB1_13	9937.416871	26.00	-2122.21
AA_DQB1_57	AA_DRB1_86	AA_DRB1_37	9941.233586	27.00	-2121.74
AA_DQB1_57	AA_DRB1_74	AA_DRB1_13	9962.368443	31.00	-2121.21

4th independent signal: DQβ1#-18 ($p = 10^{-40}$, signal peptide):
Many better combinations in exhaustive test

Amino acids in peptide-binding groove

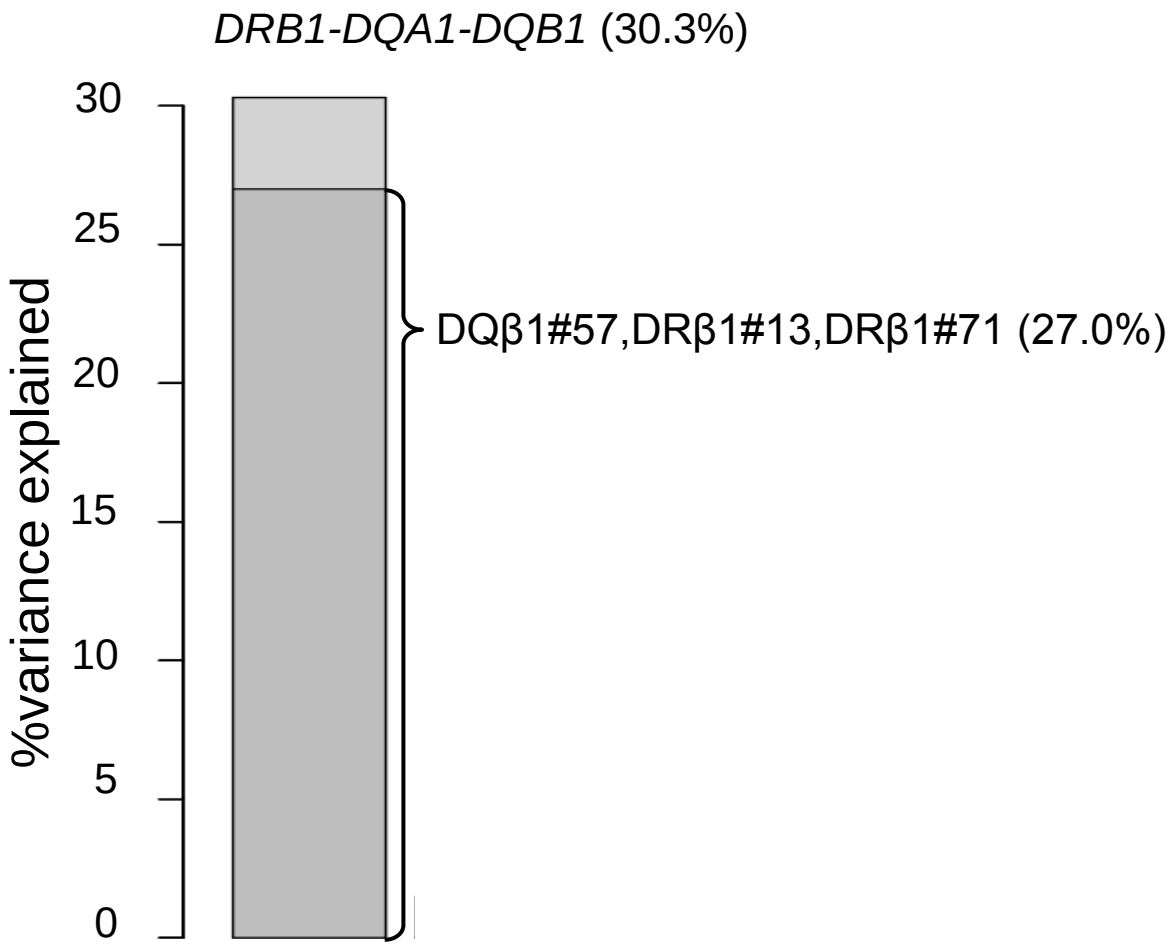




Associations outside of *DRB1-DQA1-DQB1*

- ***HLA-B:***
  *B*39:06, B*50:01, B*18:01*, etc
- ***HLA-DPB1:***
  *DPB1*04:02, DPB1*01:01*, etc
- ***HLA-A:***
  *#62, A*03, A*24:02*, etc
- No independent signal in *HLA-C* or *HLA-DPA1*

Phenotypic variance explained



Conclusion

- We developed HLA imputation tool, SNP2HLA.
- When applied this tool to T1D data, we identified that three amino acid positions are driving the traditional DRB1-DQA1-DQB1 allelic associations.

Acknowledgment


Soumya Raychaudhuri


Xinli Hu


Steve Rich


Suna Onengut


Paul de Bakker


Aaron Deutsch


Yukinori Okada


Tobias Lenz


Joanna Howson


John Todd

HST Harvard–MIT Health Sciences & Technology

HARVARD MEDICAL SCHOOL

BRIGHAM AND WOMEN'S HOSPITAL

BROAD INSTITUTE

Type 1 Diabetes Genetics Consortium (T1DGC)