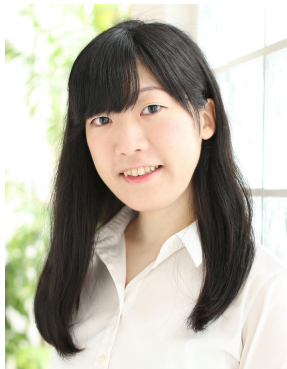Section 5
HLA Imputation

Saori Sakaue
Broad Institute
ssakaue@broadinstitute.org
@saorisakaue

MICHIGAN
IMPUTATIONSERVER

ASHG
American Society of Human Genetics

Learning objectives

Participants will

1. Learn variations within HLA

2. Learn how to impute HLA alleles and amino acid sequences from genotyped SNPs

Genotype Imputation

Genotyped SNPs → Untyped SNPs

statistically impute
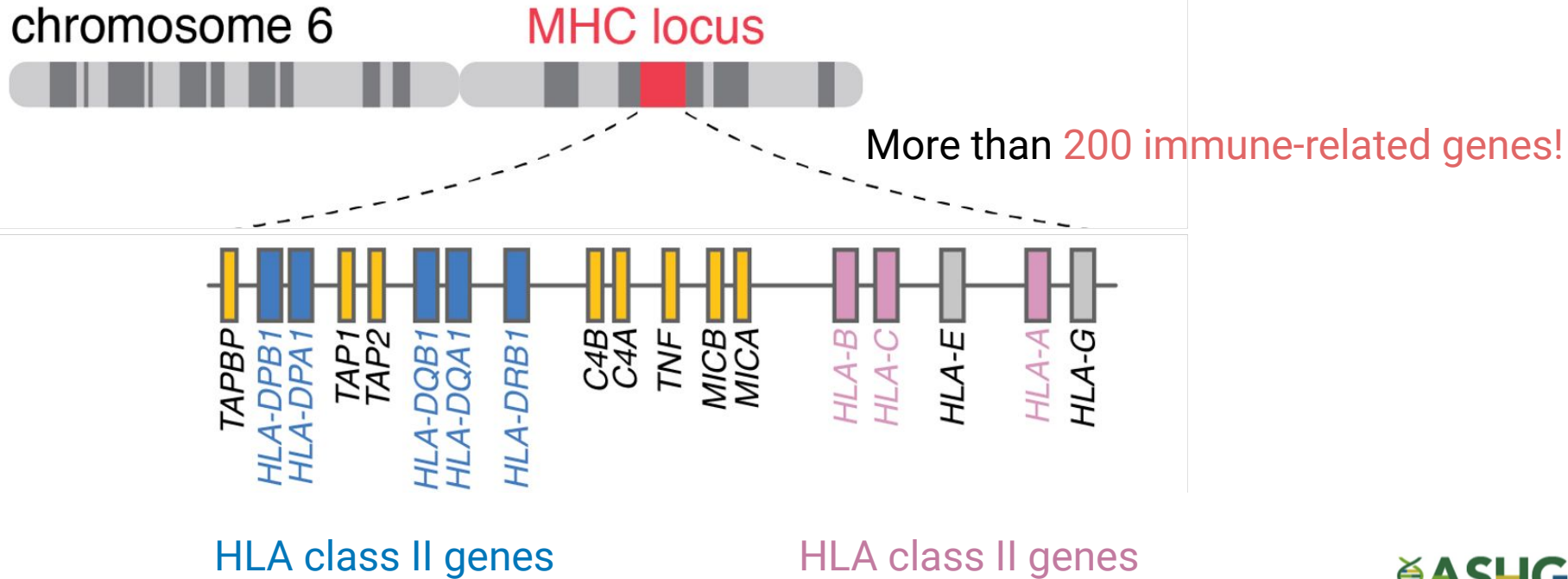
~~Genotype~~ Imputation

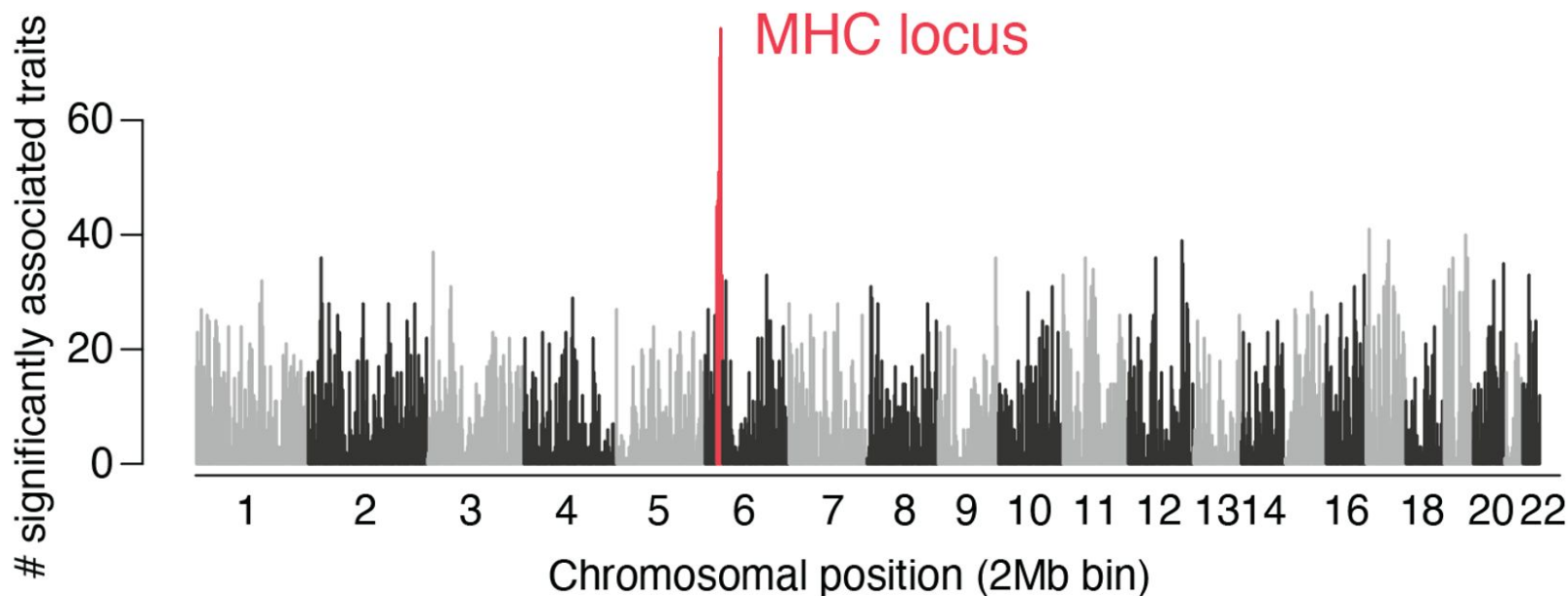## HLA (human leukocyte antigen)

Genotyped SNPs → HLA variations

statistically impute

HLA genes are within MHC (major histocompatibility complex) locus on chromosome 6
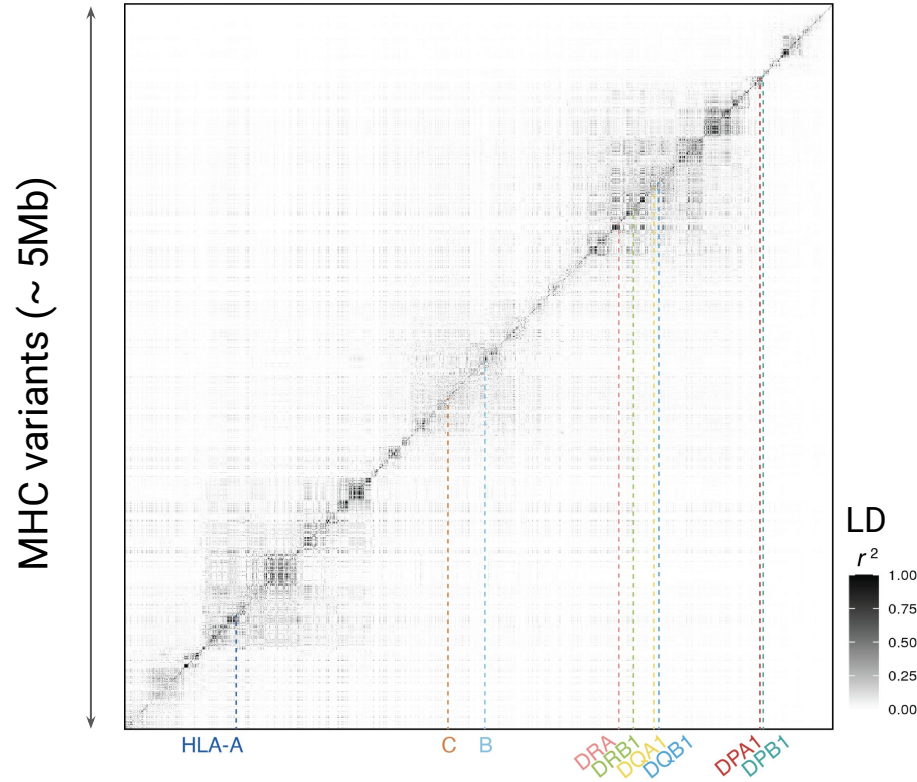

chromosome 6   MHC locus

More than 200 immune-related genes!

TAPBP HLA-DPB1 HLA-DPA1 TAP1 TAP2 HLA-DQB1 HLA-DQA1 HLA-DRB1 C4B C4A TNF MICB MICA HLA-B HLA-C HLA-E HLA-A HLA-G

HLA class II genes   HLA class II genes

MHC locus confers the largest number of associations of any locus genome-wide

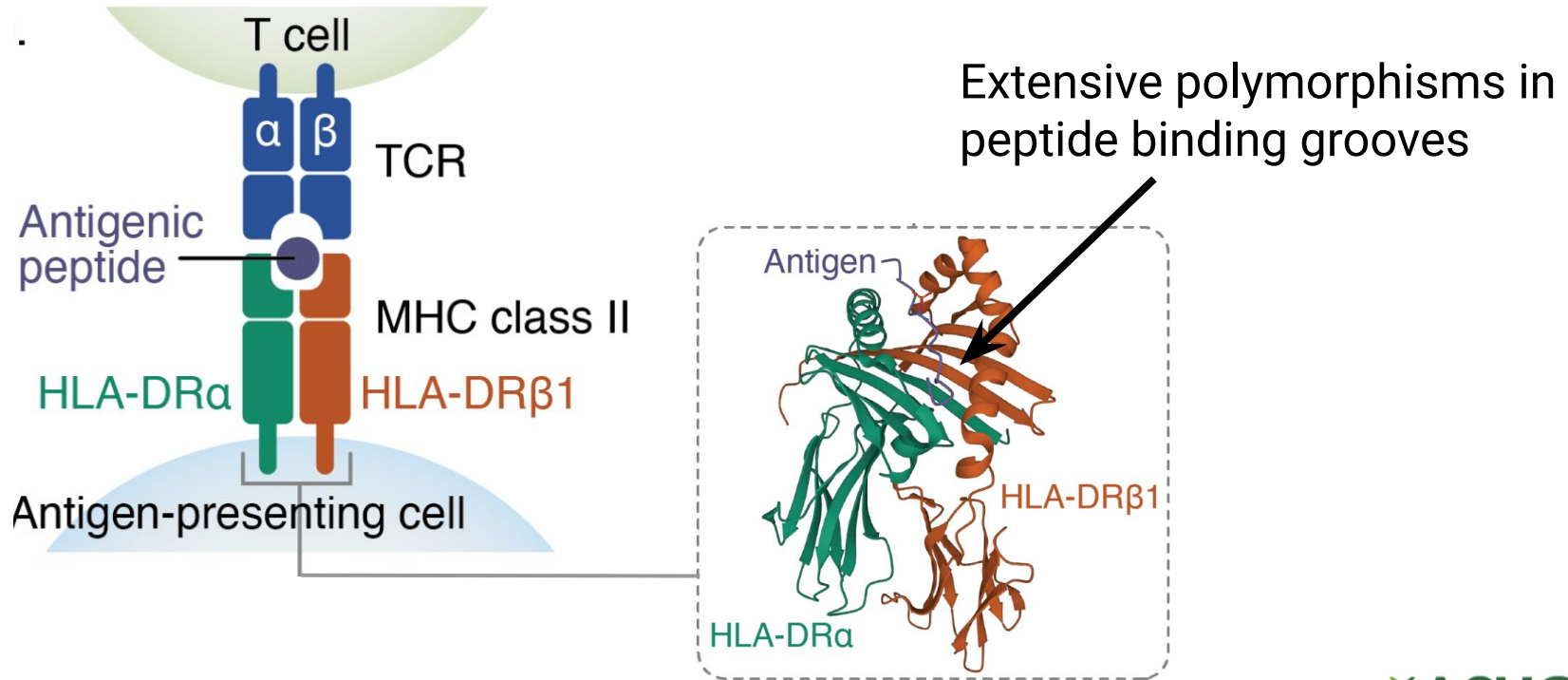# MHC locus is characterized by long and complex linkage disequilibrium structure

HLA presents antigens to immune cells
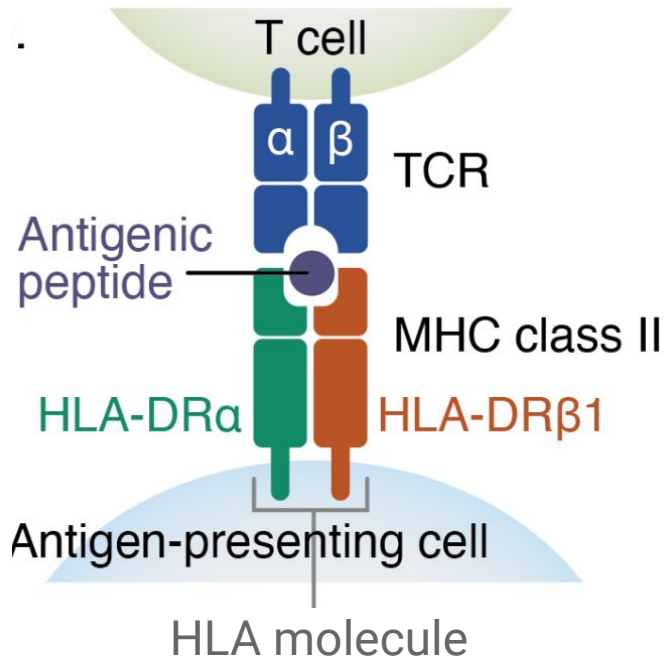
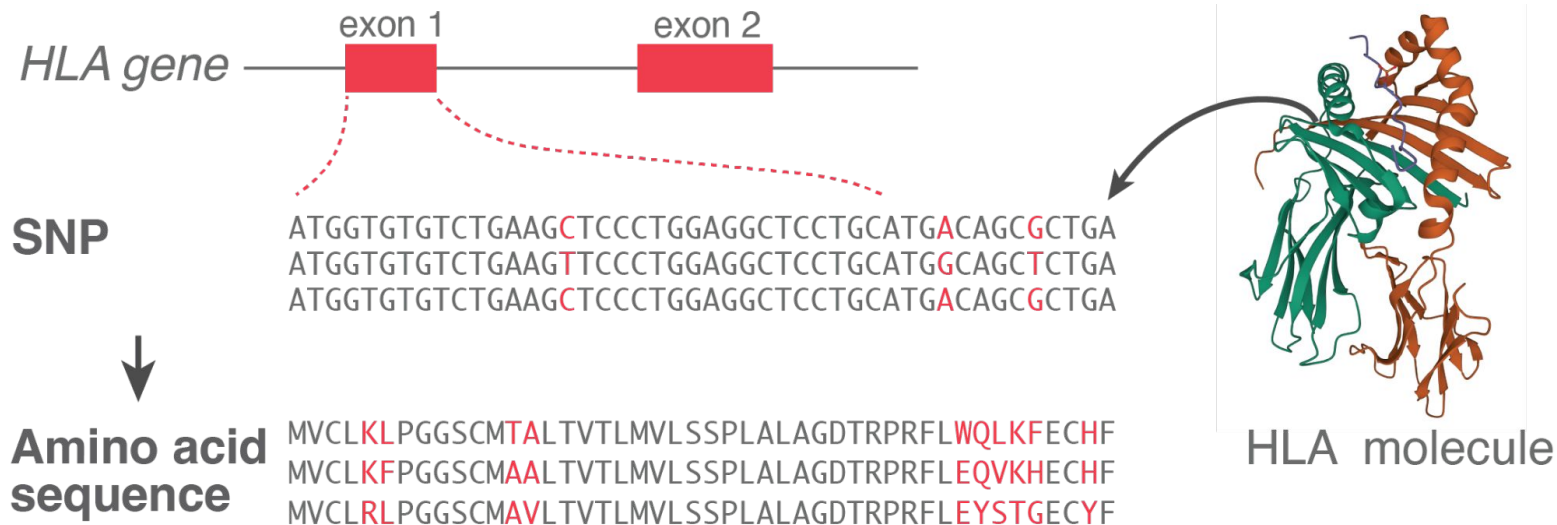# HLA presents antigens to immune cells, and therefore is extremely polymorphic
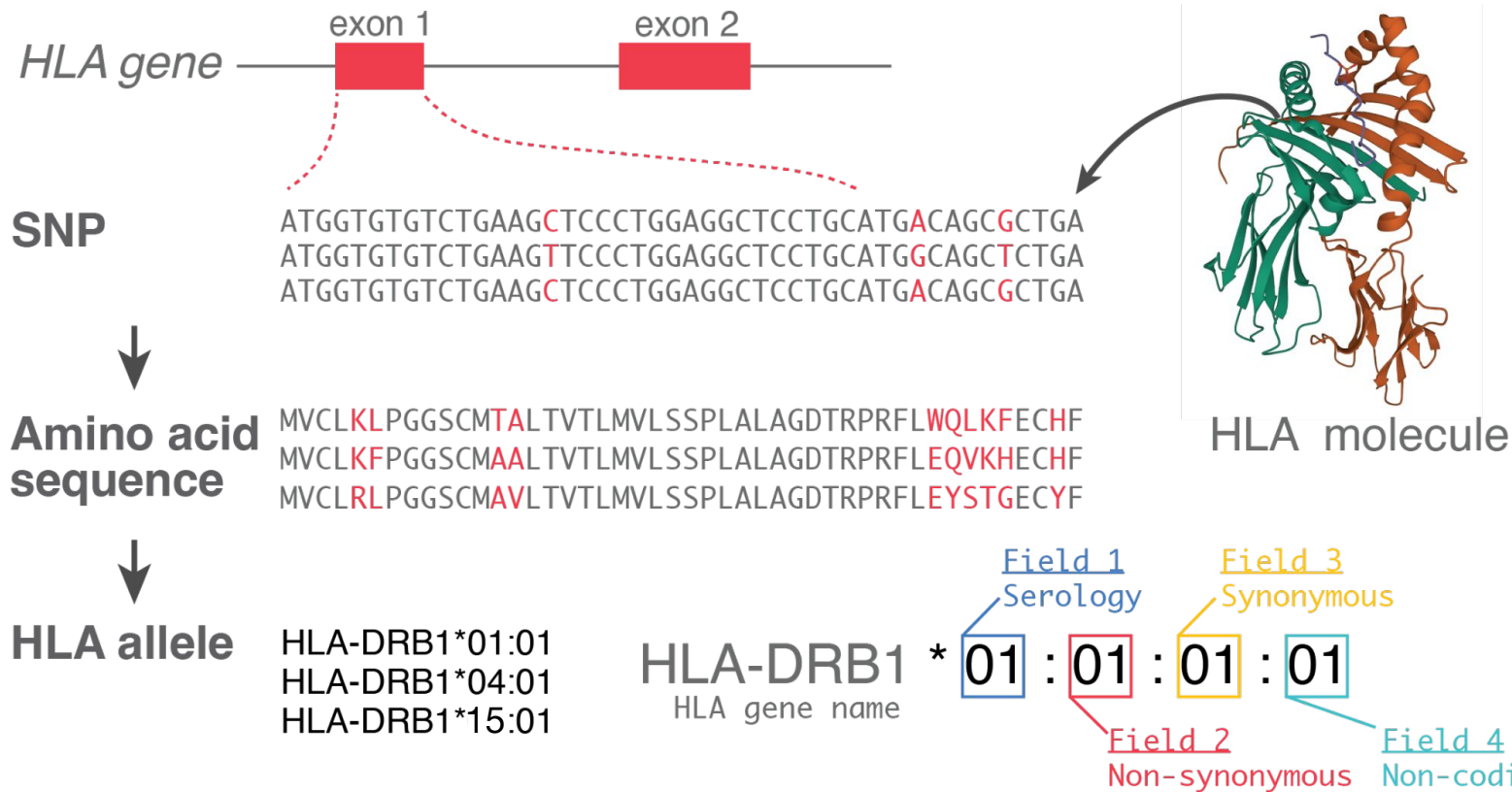

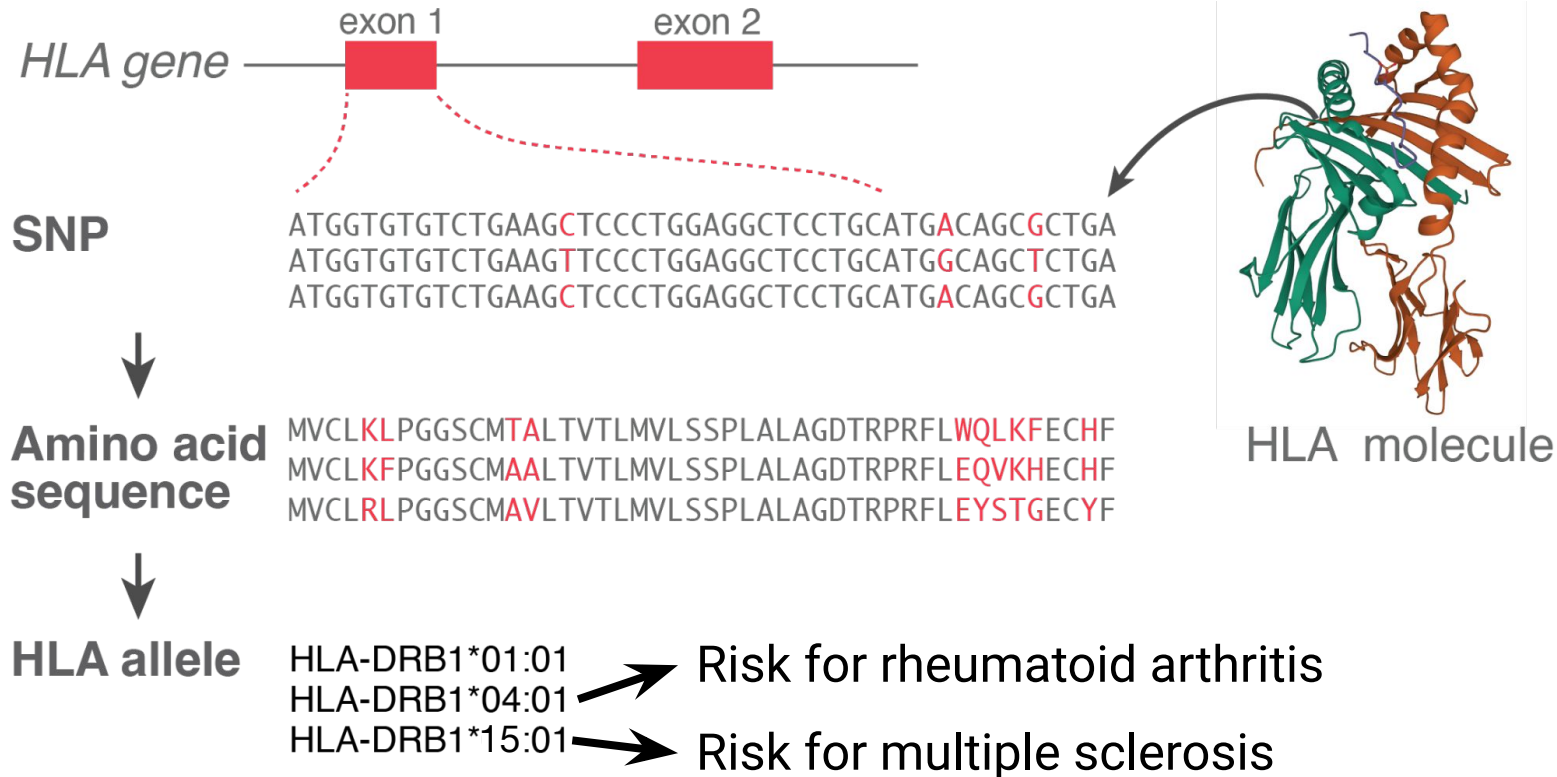Extensive polymorphisms in peptide binding grooves

HLA variations determine HLA "alleles" and function

HLA molecule

HLA variations determine HLA "alleles" and function

HLA variations determine HLA "alleles" and function

HLA gene

exon 1   exon 2

SNP

ATGGTGTGTCTGAAGCTCCCTGGAGGCTCCTGCATGACAGCGCTGA
ATGGTGTGTCTGAAGTTCCCTGGAGGCTCCTGCATGGCAGCTCTGA
ATGGTGTGTCTGAAGCTCCCTGGAGGCTCCTGCATGACAGCGCTGA

Amino acid sequence

MVCLKLPGGSCMTALTVTLMVLSSPLALAGDTRPRFLWQLKFECHF
MVCLKFPGGSCMAALTVTLMVLSSPLALAGDTRPRFLEQVKHECHF
MVCLRLPGGSCMAVLTVTLMVLSSPLALAGDTRPRFLEYSTGECYF

HLA molecule

HLA allele

HLA-DRB1*01:01
HLA-DRB1*04:01
HLA-DRB1*15:01

HLA-DRB1 * 01 : 01 : 01 : 01

HLA gene name

Field 1
Serology

Field 2
Non-synonymous

Field 3
Synonymous

Field 4
Non-coding

ASHG
American Society of Human Genetics

HLA variations determine HLA "alleles" and function

HLA gene

exon 1   exon 2

SNP

ATGGTGTGTCTGAAGCTCCCTGGAGGCTCCTGCATGACAGCGCTGA
ATGGTGTGTCTGAAGTTCCCTGGAGGCTCCTGCATGGCAGCTCTGA
ATGGTGTGTCTGAAGCTCCCTGGAGGCTCCTGCATGACAGCGCTGA

Amino acid sequence

MVCLKLPGGSCMTALTVTLMVLSSPLALAGDTRPRFLWQLKFECHF
MVCLKFPGGSCMAALTVTLMVLSSPLALAGDTRPRFLEQVKHECHF
MVCLRLPGGSCMAVLTVTLMVLSSPLALAGDTRPRFLEYSTGECYF

HLA molecule

HLA allele

HLA-DRB1*01:01
HLA-DRB1*04:01 → Risk for rheumatoid arthritis
HLA-DRB1*15:01 → Risk for multiple sclerosis

ASHG
American Society of Human Genetics

Genotype imputation in a nutshell

Genotype imputation

Known Genotyped SNPs

Given Reference haplotype with
whole-genome SNPs

**Unknown
(Impute!)** Untyped SNPs

HLA imputation in a nutshell

	Genotype imputation	HLA imputation
Known	Genotyped SNPs	Genotyped SNPs
Given	Reference haplotype with whole-genome SNPs	Reference haplotype with HLA alleles and amino acid sequences
Unknown (Impute!)	Untyped SNPs	Untyped HLA alleles and amino acid sequences

HLA imputation in a nutshell

(Known) Input

Genotyped SNP in MHC	HLA alleles	HLA amino acid sequences
CGA.ATCT..GTCTTCTGT.CTAA	?	?
CAA.ATCT..GTCCT.TGT.CTAA	?	?
CAA.ATTT..TGCTTCAGT.CTAA	?	?

(Given) HLA imputation reference panel

CGAGATCTCAGTCTTCTGTTCTAA ●—● DRB1*04:01 ●—● GGSCMAALTVTLMVL
CAAGATTTCCTTCATCTGTTCTAA ●—● DRB1*01:01 ●—● GGSCMTALTVTLMVL
CGAGATCTCCTGCTTCAGTTCTAA ●—● DRB1*01:02 ●—● GGSCMTALTVTLMVL
CAAGATCTCCGTCCTCTGTTCTAA ●—● DRB1*15:01 ●—● GGSCMTALTVTLMVL

Phasing + Imputation

Output

CGAGATCTCAGTCTTCTGTTCTAA ●—● DRB1*04:01 ●—● GGSCMAALTVTLMVL
CAAGATCTCCGTCCTCTGTTCTAA ●—● DRB1*15:01 ●—● GGSCMTALTVTLMVL
CAAGATTTCCTGCTTCAGTTCTAA ●—● DRB1*01:01 ●—● GGSCMTALTVTLMVL

HLA imputation in a nutshell

(Known) Input

Genotyped SNP in MHC | HLA alleles | HLA amino acid sequences

```
CGA.ATCT..GTCTTCTGT.CTAA          ?              ?
CAA.ATCT..GTCCT.TGT.CTAA          ?              ?
CAA.ATTT..TGCTTCAGT.CTAA          ?              ?
```

(Given) HLA imputation reference panel

```
CGAGATCTCAGTCTTCTGTTCTAA •—• DRB1*04:01 •—• GGSCMAALTVTLMVL
CAAGATTTCCTTCATCTGTTCTAA •—• DRB1*01:01 •—• GGSCMTALTVTLMVL
CGAGATCTCCTGCTTCAGTTCTAA •—• DRB1*01:02 •—• GGSCMTALTVTLMVL
CAAGATCTCCGTCCTCTGTTCTAA •—• DRB1*15:01 •—• GGSCMTALTVTLMVL
```

Multi-ancestry panel
(EUR, EAS, SAS, AFR, LAT)

Luo et al. Nat Genet 2021, Sakaue et al. bioRxiv 2022

Phasing + Imputation

Output

```
CGAGATCTCAGTCTTCTGTTCTAA •—• DRB1*04:01 •—• GGSCMAALTVTLMVL
CAAGATCTCCGTCCTCTGTTCTAA •—• DRB1*15:01 •—• GGSCMTALTVTLMVL
CAAGATTTCCTGCTTCAGTTCTAA •—• DRB1*01:01 •—• GGSCMTALTVTLMVL
```

American Society of Human Genetics

HLA imputation in a nutshell

(Known) Input

Genotyped SNP in MHC	HLA alleles	HLA amino acid sequences
CGA.ATCT..GTCTTCTGT.CTAA	?	?
CAA.ATCT..GTCCT.TGT.CTAA	?	?
CAA.ATTT..TGCTTCAGT.CTAA	?	?

(Given) HLA imputation reference panel

CGAGATCTCAGTCTTCTGTTCTAA •——• DRB1*04:01 •——• GGSCMAALTVTLMVL
CAAGATTTCCTTCATCTGTTCTAA •——• DRB1*01:01 •——• GGSCMTALTVTLMVL
CGAGATCTCCTGCTTCAGTTCTAA •——• DRB1*01:02 •——• GGSCMTALTVTLMVL
CAAGATCTCCGTCCTCTGTTCTAA •——• DRB1*15:01 •——• GGSCMTALTVTLMVL

Multi-ancestry panel
(EUR, EAS, SAS, AFR, LAT)

Phasing + Imputation

Beagle   Beagle
SHAPEIT   Minimac
Eagle

Output

CGAGATCTCAGTCTTCTGTTCTAA •——• DRB1*04:01 •——• GGSCMAALTVTLMVL
CGAGATCTCCGTCCTCTGTTCTAA •——• DRB1*15:01 •——• GGSCMTALTVTLMVL
CAAGATTTCCTGCTTCAGTTCTAA •——• DRB1*01:01 •——• GGSCMTALTVTLMVL

HLA imputation in a nutshell

(Known) Input

Genotyped SNP in MHC   HLA alleles   HLA amino acid sequences

```
CGA.ATCT..GTCTTCTGT.CTAA
CAA.ATCT..GTCCT.TGT.CTAA
CAA.ATTT..TGCTTCAGT.CTAA
```

?	?
?	?
?	?

(Given) HLA imputation reference panel

```
CGAGATCTCAGTCTTCTGTTCTAA ——— DRB1*04:01 ——— GGSCMAALTVTLMVL
CAAGATTTCCTTCATCTGTTCTAA ——— DRB1*01:01 ——— GGSCMTALTVTLMVL
CGAGATCTCCTGCTTCAGTTCTAA ——— DRB1*01:02 ——— GGSCMTALTVTLMVL
CAAGATCTCCGTCCTCTGTTCTAA ——— DRB1*15:01 ——— GGSCMTALTVTLMVL
```

Multi-ancestry panel
(EUR, EAS, SAS, AFR, LAT)

Phasing + Imputation

SNP2HLA (Jia et al. PLoS One 2013)

Beagle	Beagle
SHAPEIT	Minimac
Eagle	

MICHIGAN IMPUTATIONSERVER

ASHG
American Society of Human Genetics

HLA imputation in MIS

HLA imputation in MIS

Latest HLA reference panel

HLA imputation in MIS

HLA imputation in MIS

Can be prephased or phasing by Eagle at MIS

Let's look at the output from MIS!


```
ssakaue@wmbed-37d:~/Downloads/chr_6 (2)$ zcat chr6.dose.vcf.gz | less -S
```

Let's look at the output from MIS!


```
##fileformat=VCFv4.1
##filedate=2022.11.14
##contig=<ID=6>
##INFO=<ID=AF,Number=1,Type=Float,Description="Estimated Alternate Allele Frequency">
##INFO=<ID=MAF,Number=1,Type=Float,Description="Estimated Minor Allele Frequency">
##INFO=<ID=R2,Number=1,Type=Float,Description="Estimated Imputation Accuracy (R-square)">
##INFO=<ID=ER2,Number=1,Type=Float,Description="Empirical (Leave-One-Out) R-square (available only for genotyped variants)">
##INFO=<ID=IMPUTED,Number=0,Type=Flag,Description="Marker was imputed but NOT genotyped">
##INFO=<ID=TYPED,Number=0,Type=Flag,Description="Marker was genotyped AND imputed">
##INFO=<ID=TYPED_ONLY,Number=0,Type=Flag,Description="Marker was genotyped but NOT imputed">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]">
##FORMAT=<ID=HDS,Number=2,Type=Float,Description="Estimated Haploid Alternate Allele Dosage ">
##FORMAT=<ID=GP,Number=3,Type=Float,Description="Estimated Posterior Probabilities for Genotypes 0/0, 0/1 and 1/1 ">
##pipeline=michigan-imputationserver-1.5.8
##imputation=minimac4-1.0.2
##phasing=eagle-2.4
##panel=apps@multiethnic-hla-panel-4digit-v2@1.0.0
##r2Filter=0.0
```

Imputed SNPs within MHC

```
#CHROM   POS       ID       REF    ALT    QUAL   FILTER  INFO       FORMAT    010061321010_R01C01_10007854   010061321010_R02C01_10020793   010061321010_
6        27970031  rs149946        G      T      .      PASS    AF=0.22479;MAF=0.22479;R2=0.99214;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6        27976200  rs9380032       G      T      .      PASS    AF=0.02975;MAF=0.02975;R2=0.97885;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6        27979188  rs4141691       A      G      .      PASS    AF=0.11754;MAF=0.11754;R2=0.94642;IMPUTED       GT:DS:HDS:GP    0|0:0:0.003:0.0
6        27979625  rs10484402      A      G      .      PASS    AF=0.04041;MAF=0.04041;R2=0.92706;IMPUTED       GT:DS:HDS:GP    0|0:0:0.003:0.0
6        27981673  rs9368540       G      A      .      PASS    AF=0.03706;MAF=0.03706;R2=0.98634;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6        27984726  rs74505854      A      C      .      PASS    AF=0.00719;MAF=0.00719;R2=0.94391;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6        27984907  rs17765055      T      C      .      PASS    AF=0.04632;MAF=0.04632;R2=0.99897;ER2=0.97293;TYPED     GT:DS:HDS:GP    0|0:0
6        27986199  rs72848791      C      T      .      PASS    AF=0.04361;MAF=0.04361;R2=0.99732;ER2=0.97465;TYPED     GT:DS:HDS:GP    0|0:0
6        27986529  rs9368544       A      C      .      PASS    AF=0.04631;MAF=0.04631;R2=0.99885;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6        27998258  rs149990        G      A      .      PASS    AF=0.11782;MAF=0.11782;R2=0.99974;ER2=0.99765;TYPED     GT:DS:HDS:GP    0|0:0
6        27999044  rs9368545       A      T      .      PASS    AF=0.04627;MAF=0.04627;R2=0.99825;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6        27999421  rs16893573      C      T      .      PASS    AF=0.02852;MAF=0.02852;R2=0.98601;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6        28001003  rs17708949      A      C      .      PASS    AF=0.03124;MAF=0.03124;R2=0.98465;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6        28001610  rs149942        T      C      .      PASS    AF=0.26963;MAF=0.26963;R2=0.99868;ER2=0.98731;TYPED     GT:DS:HDS:GP    0|0:0
6        28002388  rs149943        G      A      .      PASS    AF=0.11781;MAF=0.11781;R2=0.99984;ER2=0.99889;TYPED     GT:DS:HDS:GP    0|0:0
6        28003271  rs183926        T      A      .      PASS    AF=0.00996;MAF=0.00996;R2=0.99551;ER2=0.95540;TYPED     GT:DS:HDS:GP    0|0:0
```

Let's look at the output from MIS!


```
ssakaue@wmbed-37d:~/Downloads/chr_6 (2)$ zcat chr6.dose.vcf.gz | less -S
ssakaue@wmbed-37d:~/Downloads/chr_6 (2)$ zcat chr6.dose.vcf.gz | grep HLA | less -S
```

Imputed HLA alleles

6	29910247	HLA_A*01	A	T	.	PASS	AF=0.15456;MAF=0.15456;R2=0.99719;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910248	HLA_A*01:01	A	T	.	PASS	AF=0.15234;MAF=0.15234;R2=0.99517;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910249	HLA_A*01:02	A	T	.	PASS	AF=0.00110;MAF=0.00110;R2=0.85198;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910250	HLA_A*01:136	A	T	.	PASS	AF=0.00002;MAF=0.00002;R2=0.04644;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910251	HLA_A*02	A	T	.	PASS	AF=0.26026;MAF=0.26026;R2=0.99741;IMPUTED	GT:DS:HDS:GP	0	1:0.998:0.0
6	29910252	HLA_A*02:01	A	T	.	PASS	AF=0.22914;MAF=0.22914;R2=0.98881;IMPUTED	GT:DS:HDS:GP	0	1:0.996:0.0
6	29910253	HLA_A*02:02	A	T	.	PASS	AF=0.00471;MAF=0.00471;R2=0.99686;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910254	HLA_A*02:03	A	T	.	PASS	AF=0.00094;MAF=0.00094;R2=0.88473;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910255	HLA_A*02:04	A	T	.	PASS	AF=0.00017;MAF=0.00017;R2=0.89691;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910256	HLA_A*02:05	A	T	.	PASS	AF=0.01556;MAF=0.01556;R2=0.99839;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910257	HLA_A*02:06	A	T	.	PASS	AF=0.00364;MAF=0.00364;R2=0.98705;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910258	HLA_A*02:07	A	T	.	PASS	AF=0.00133;MAF=0.00133;R2=0.94265;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910259	HLA_A*02:10	A	T	.	PASS	AF=0.00001;MAF=0.00001;R2=0.06774;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910260	HLA_A*02:11	A	T	.	PASS	AF=0.00072;MAF=0.00072;R2=0.79302;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910261	HLA_A*02:135	A	T	.	PASS	AF=0.00005;MAF=0.00005;R2=0.22275;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910262	HLA_A*02:17	A	T	.	PASS	AF=0.00108;MAF=0.00108;R2=0.91078;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910263	HLA_A*02:195	A	T	.	PASS	AF=0.00003;MAF=0.00003;R2=0.00997;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910264	HLA_A*02:20	A	T	.	PASS	AF=0.00019;MAF=0.00019;R2=0.41177;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910265	HLA_A*02:22	A	T	.	PASS	AF=0.00026;MAF=0.00026;R2=0.84742;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910266	HLA_A*02:279	A	T	.	PASS	AF=0.00008;MAF=0.00008;R2=0.24552;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910267	HLA_A*02:55	A	T	.	PASS	AF=0.00001;MAF=0.00001;R2=0.02045;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910268	HLA_A*02:56	A	T	.	PASS	AF=0.00007;MAF=0.00007;R2=0.08360;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910269	HLA_A*02:60	A	T	.	PASS	AF=0.00004;MAF=0.00004;R2=0.49123;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910270	HLA_A*02:76	A	T	.	PASS	AF=0.00030;MAF=0.00030;R2=0.40308;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910271	HLA_A*02:87	A	T	.	PASS	AF=0.00001;MAF=0.00001;R2=0.00984;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910272	HLA_A*03	A	T	.	PASS	AF=0.12657;MAF=0.12657;R2=0.99727;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910273	HLA_A*03:01	A	T	.	PASS	AF=0.12061;MAF=0.12061;R2=0.99073;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910274	HLA_A*03:02	A	T	.	PASS	AF=0.00441;MAF=0.00441;R2=0.97833;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910275	HLA_A*03:36N	A	T	.	PASS	AF=0.00063;MAF=0.00063;R2=0.40664;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910276	HLA_A*03:89	A	T	.	PASS	AF=0.00006;MAF=0.00006;R2=0.05589;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910277	HLA_A*11	A	T	.	PASS	AF=0.06332;MAF=0.06332;R2=0.99759;IMPUTED	GT:DS:HDS:GP	1	0:1.000:1.0
6	29910278	HLA_A*11:01	A	T	.	PASS	AF=0.05966;MAF=0.05966;R2=0.96821;IMPUTED	GT:DS:HDS:GP	1	0:0.958:0.9
6	29910279	HLA_A*11:02	A	T	.	PASS	AF=0.00059;MAF=0.00059;R2=0.84655;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910280	HLA_A*11:32	A	T	.	PASS	AF=0.00021;MAF=0.00021;R2=0.04209;IMPUTED	GT:DS:HDS:GP	0	0:0:0.024:0.0
6	29910281	HLA_A*11:50Q	A	T	.	PASS	AF=0.00250;MAF=0.00250;R2=0.41268;IMPUTED	GT:DS:HDS:GP	0	0:0:0.018:0.0
6	29910282	HLA_A*23	A	T	.	PASS	AF=0.02822;MAF=0.02822;R2=0.99569;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910283	HLA_A*23:01	A	T	.	PASS	AF=0.02797;MAF=0.02797;R2=0.99150;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0
6	29910284	HLA_A*23:15	A	T	.	PASS	AF=0.00000;MAF=0.00000;R2=0.00005;IMPUTED	GT:DS:HDS:GP	0	0:0:0,0:1,0

REF   ALT
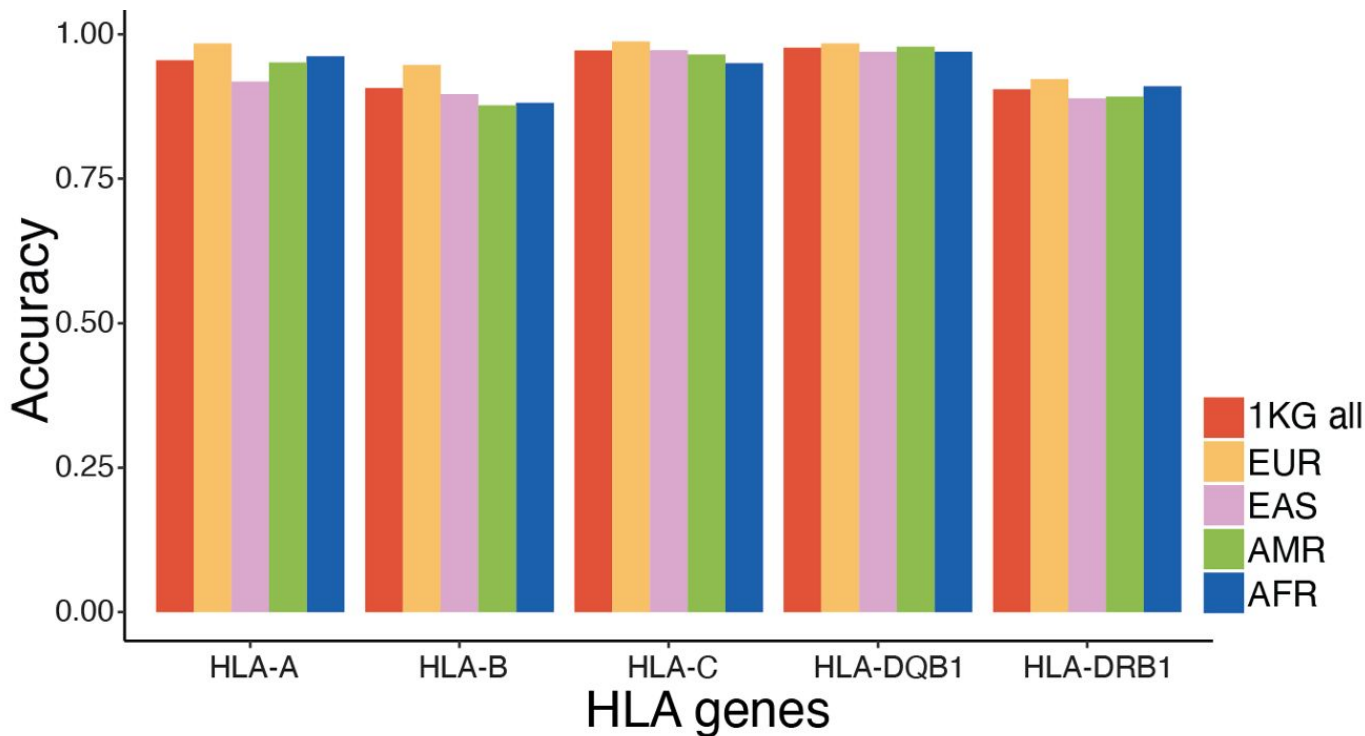
"T": Presence of the allele
"A": Absence of the allele

Imputed HLA alleles

REF   ALT

"T": Presence of the allele
"A": Absence of the allele

HLA imputation in MIS is accurate!

Summary

1. HLA amino acid sequences and alleles characterize antigen presentation and disease risk within HLA.
2. HLA imputation enables imputation of HLA amino acid sequences and alleles from genotyped SNPs.

Summary

1. HLA amino acid sequences and alleles characterize antigen presentation and disease risk within HLA.
2. HLA imputation enables imputation of HLA amino acid sequences and alleles from genotyped SNPs.

For more information and downstream statistical tests…

ASHG
American Society of Human Genetics