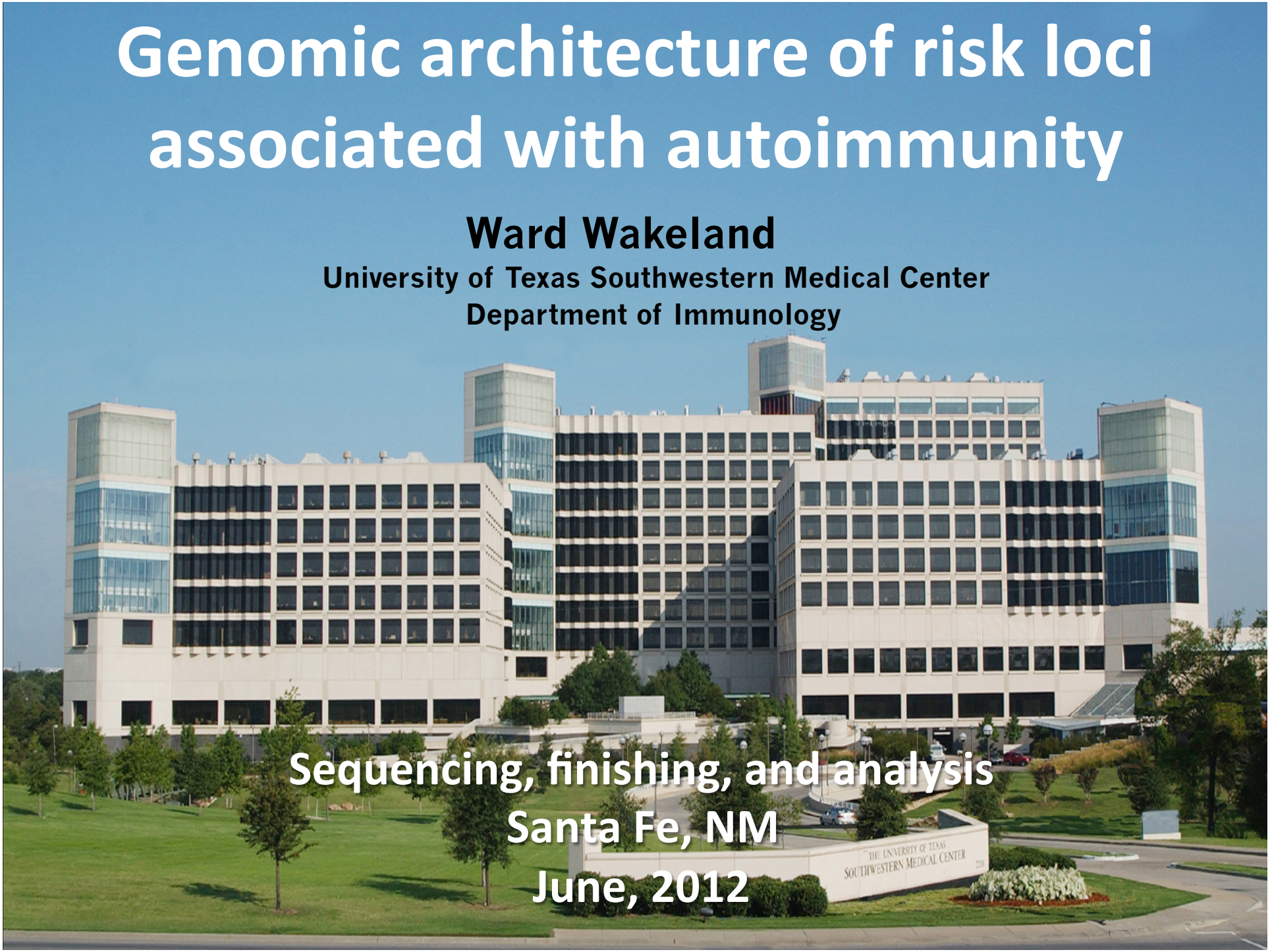
Genomic architecture of risk loci associated with autoimmunity

Ward Wakeland
University of Texas Southwestern Medical Center
Department of Immunology



Sequencing, finishing, and analysis
Santa Fe, NM
June, 2012

More than 25 common alleles in the general population contribute to SLE susceptibility

Table 1 | The main systemic lupus erythematosus candidate genes[4–11]

Gene	Approximate OR*	Potential function
MHC class III region	>2.0	Unknown gene or genes
MHC class II region	>2.0	Antigen presentation
IRF5	1.5–2.0	Innate immune signaling and lymphocyte activation—T cells
ITGAM	1.2–1.5	Antigen recognition and immune complex processing
STAT4	1.2–1.5	Innate immune signaling and lymphocyte activation—T cells
PTPN22	1.2–1.5	Lymphocyte activation—T cells
TNFAIP3	1.2–1.5	Innate immune signaling
TNFSF4	1.2–1.5	Lymphocyte activation—T cells and B cells
FCGR3B	1.2–1.5	Antigen recognition and immune complex processing
BLK	1.2–1.5	Lymphocyte activation—B cells
FCGR2A	1.2–1.5	Antigen recognition and immune complex processing
TNIP1	1.2–1.5	Innate immune signaling
IRAK1 or *MECP2*	1.2–1.5	Innate immune signaling (*IRAK1*) or methylation (*MECP2*)
ATG5	1.2–1.5	Autophagy/apoptosis
UBE2L3	<1.2	Ubiquitination
BANK1	<1.2	Lymphocyte activation—B cells
PHRF1	<1.2	Unknown
PXK	<1.2	Unknown
PRDM1	<1.2	Unknown
JAZF1	<1.2	Unknown
UHRF1BP1	<1.2	Unknown
IL10	<1.2	Immune signaling

*Strong genetic effect, OR >2.1; moderate effect, OR 1.2–2.0; weak effect, OR <1.2. Abbreviations: MHC, major histocompatibility complex; OR, odds ratio.

Nat. Rev. Rheumatol. 2010

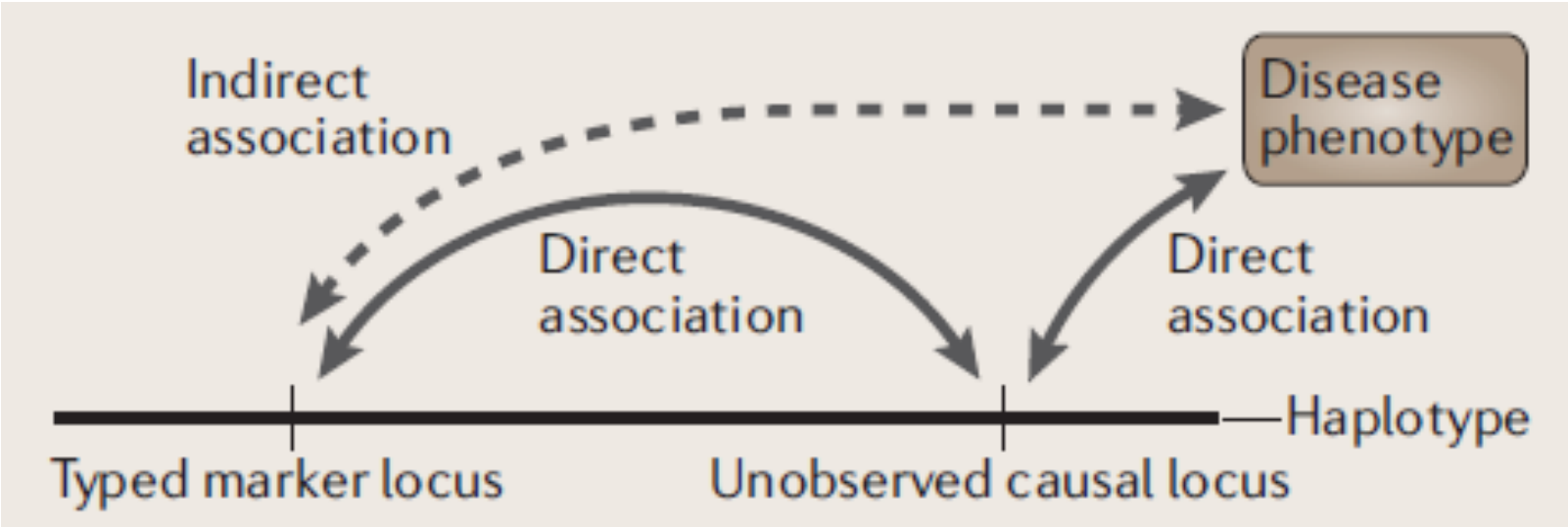
Common allele – weak effect paradigm:

Combinations of common alleles (>0.1) in the population cause disease

Many (50-100?) genes associated, but all have weak effects

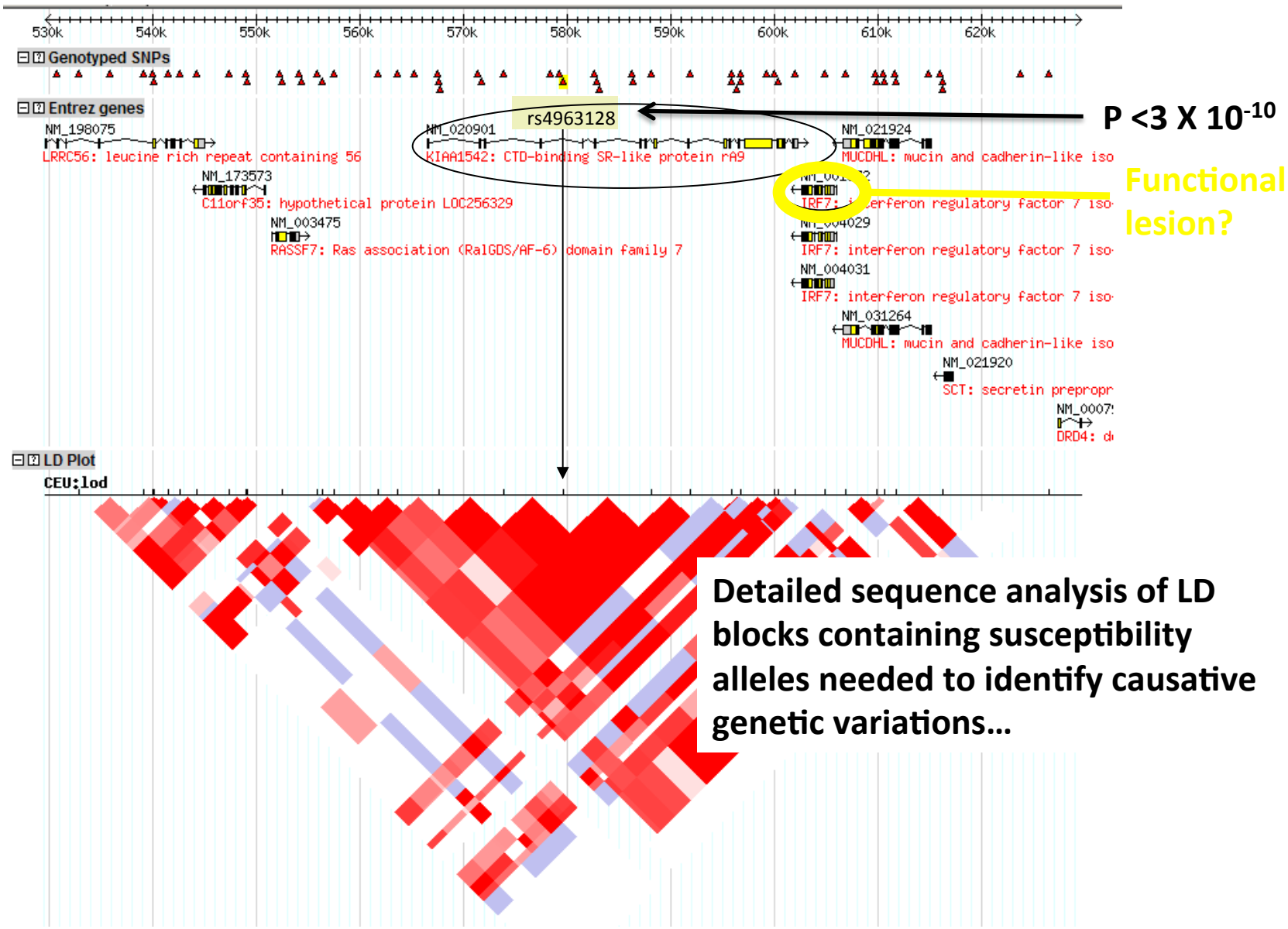
Disease susceptibility caused by cumulative effect of many genes

Detection of disease allele with linked markers



From Balding, *Nat. Rev. Genetics,* 2006

Tagging SNPs and LD blocks



$P < 3 \times 10^{-10}$

Functional lesion?

Detailed sequence analysis of LD blocks containing susceptibility alleles needed to identify causative genetic variations…

Population sequencing of LD blocks to identify causal genetic variants of autoimmune disease

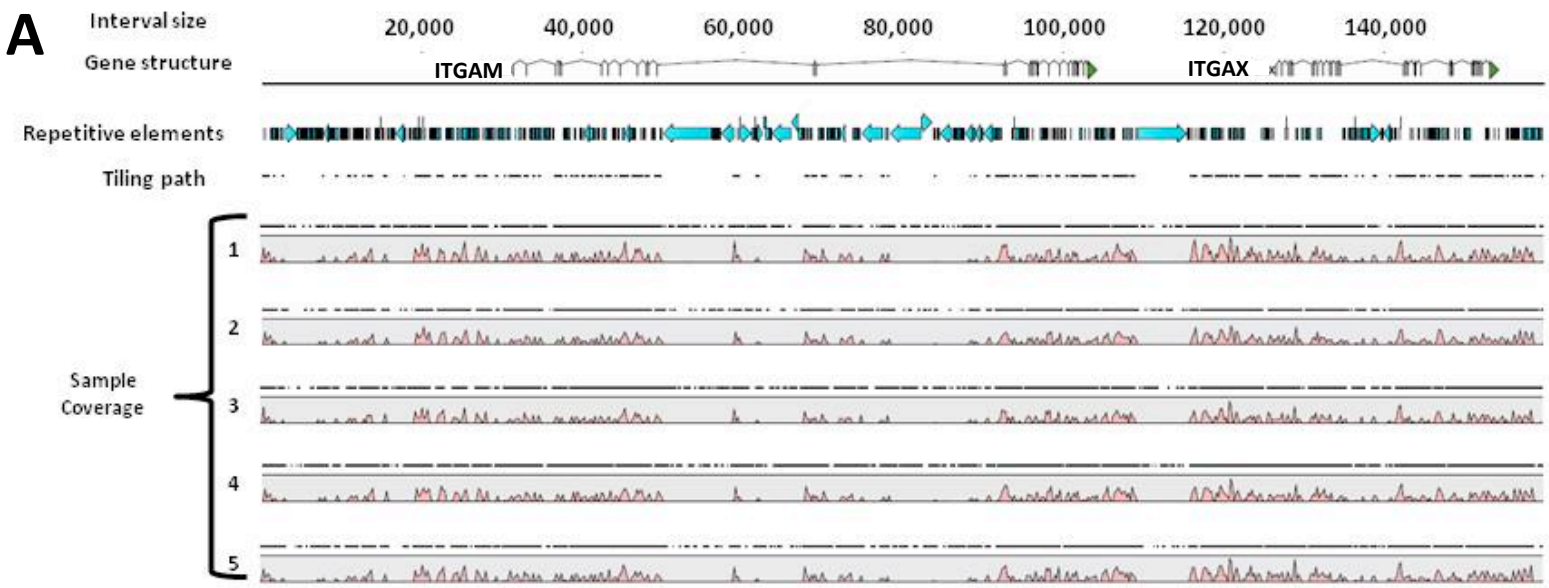
- Paired-end 80/100 bp sequencing using the Illumina GAIIx/HISEQ 2000

- Targeted resequencing of ~10 Mb of genome in 88 LD blocks using custom designed SureSelect capture arrays

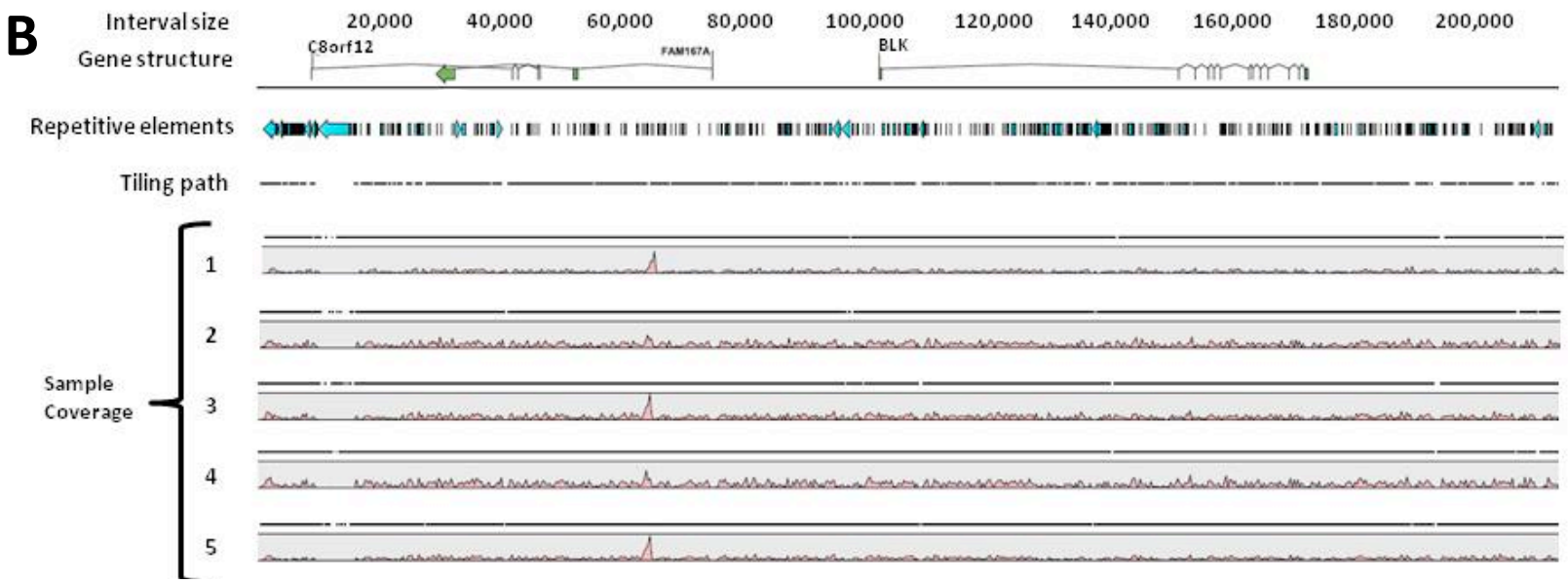
- Produced ~1 Gigabase of high quality, targeted sequence/sample

Identification of sequence variations

- Sequenced 194 Caucasian SLE patients and matched controls for 25 risk loci with an average of 107 fold coverage of non-repetitive regions
- Assembled with BWA-GATK pipeline to HG19 reference human genome
- Called SNPs and Indels using Unified Genotyper
  - SNPs/Indels found in any sample assessed in all samples
  - Developed hard filters to screen data for quality and accuracy by concordance estimates with Immunochip data
  - Developed variant calling pipeline with 99.2% concordance for Immunochip calls



62.8% of 160Kb

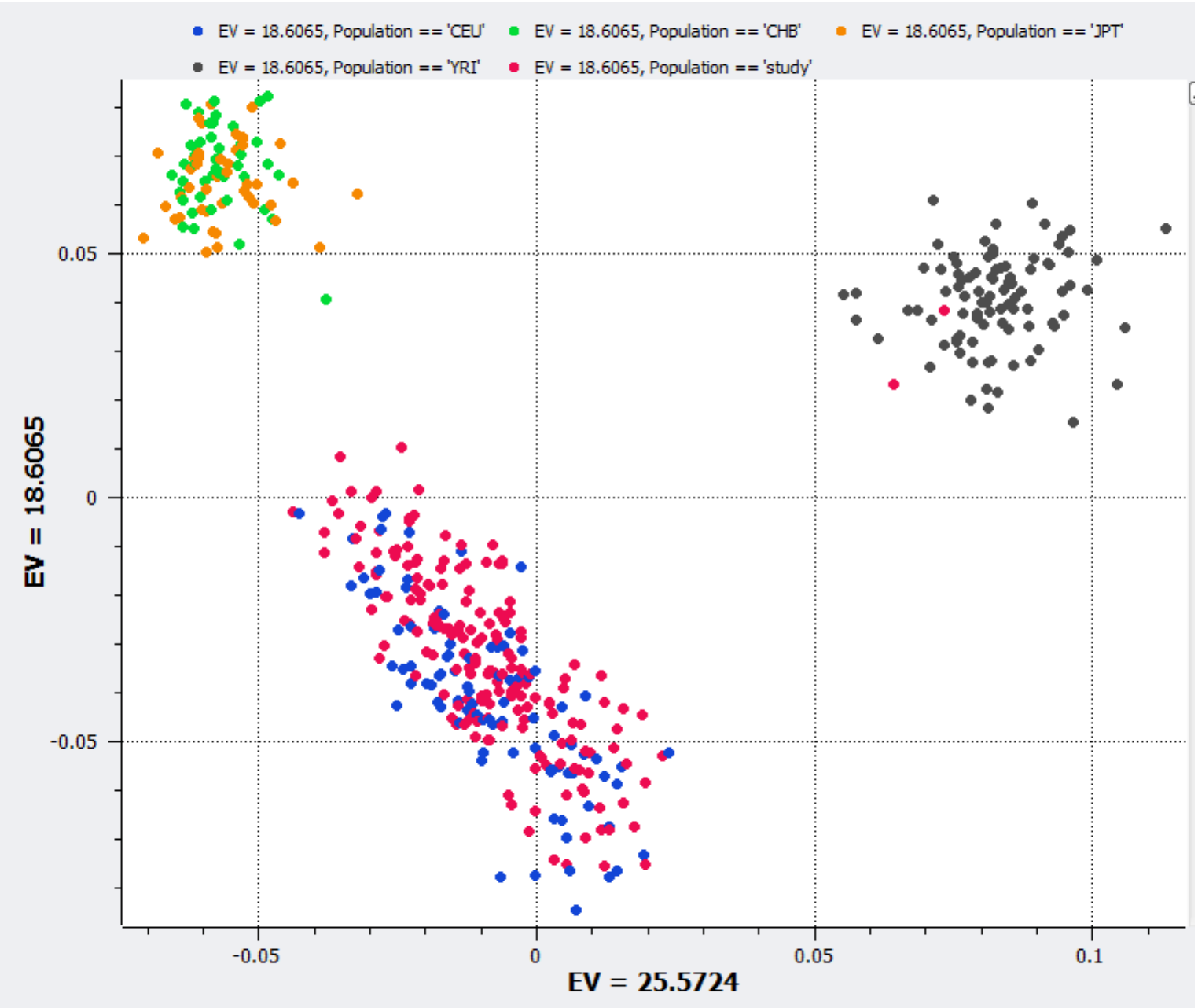


91.98% of 213Kb

Sequence coverage of targeted risk loci

Region Name	Total region size	Region Fold Coverage	% 1X Coverage	Non-repetitive region	Non-Repetitive Fold Coverage	%15X Coverage
ATG5	149342	59.55	87.54	86773	86.79	91.83
Bank1	338212	64.70	87.37	207075	90.21	91.18
BLK	213474	87.37	91.98	146074	110.49	94.32
ETS1	64548	105.61	96.48	51817	118.55	95.87
IKZF1	265152	78.18	77.69	163433	113.64	92.77
IRF5	228276	62.07	82.44	120948	98.93	91.85
IRF7	135041	55.11	81.44	85076	78.74	75.91
ITGAM	159999	48.90	62.80	56449	116.09	94.08
JAZF1	459898	87.06	90.44	311333	112.74	93.50
LYN	197059	73.70	82.73	100826	122.94	95.61
MECP2-IRAK	123088	57.86	82.54	69107	89.92	88.64
MSH5	75834	70.33	83.13	46283	102.64	89.27
NCF2	135998	81.49	90.21	87941	109.20	92.35
PRDM1	72998	98.32	94.11	51678	123.72	97.14
PTPN22	58895	51.31	80.78	27561	89.52	91.46
PTTG1	68031	95.95	91.15	43603	129.32	95.55
PXK	95260	79.21	85.82	53394	122.33	96.53
SCUBE1	296563	71.54	90.24	200825	92.63	89.89
STAT4	124748	65.80	82.77	68977	103.33	93.56
TLR7; TLR8	65679	77.88	89.74	43495	103.67	95.67
TNFAIP3	309999	85.84	87.33	190076	120.71	96.24
TNFSF4	251863	73.64	84.65	146587	108.55	94.83
TNIP1	71998	97.27	95.69	54952	114.86	93.56
UBE2L3	82786	68.93	74.49	38722	127.62	91.44
XKR6	322217	87.68	91.66	217741	113.54	94.96
Dataset Totals	**4,366,958**	**75.44**	**86.0%**	**2,670,746**	**107.58**	**93.09%**

PCA analysis of Affy 6.0 Hapmap 270 samples and our 194 Caucasians



Defining potentially functional variations within the dataset

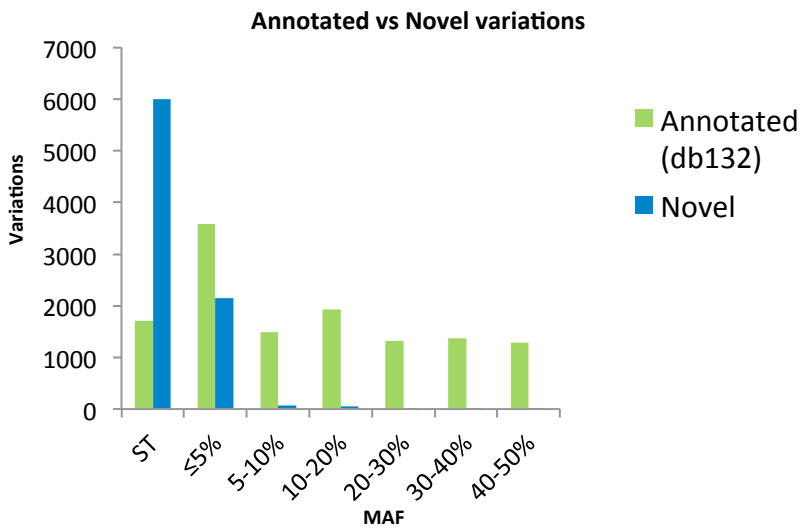
Definition of potentially functional variants

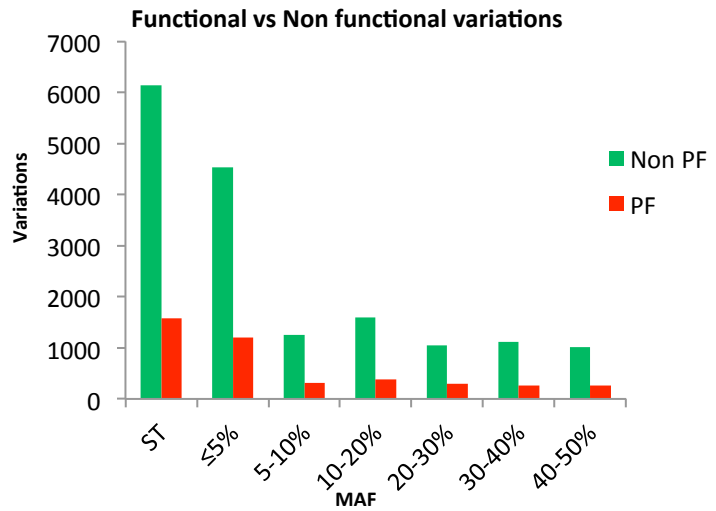
- Variant located in genomic region with annotated functional characteristics
  - NSC, SC, coding del, or frame shift
  - Other annotated genomic regions: TFBS, CpG islands, 3' or 5' UTR, splice sites, DH sites, miRNA, ncRNA, etc.
- eQTL analysis
  - 62 lymphoblastoid cell lines analyzed by RNA-SEQ and targeted genomic sequencing
  - 2036 potentially functional cis-eQTL SNPs detected among 25 genomic intervals assayed (~12%)

Identification of 4030 potentially functional variations in annotated regions of 25 LD blocks in 192 Caucasians

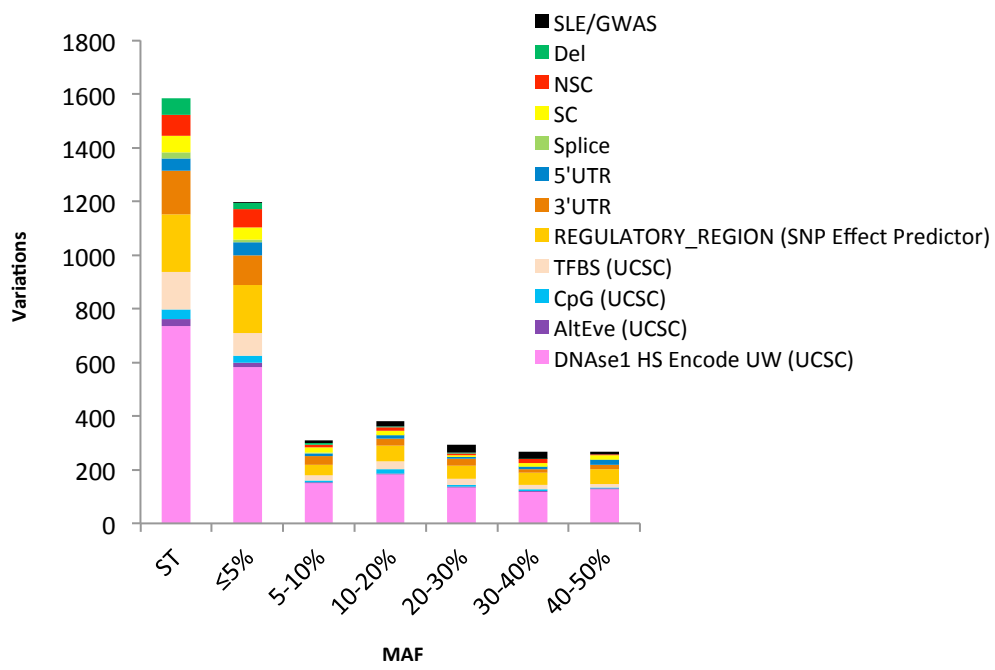
Regions	Chr Coordinates	Genes	Known variations (dbSNP 132)	Novel	Potential functional
ATG5	chr6:106630352-106779695	1	307	219	60
Bank1	chr4:102709764-103047977	1	799	492	47
BLK	chr8:11249633-11463108	4	1226	668	573
CTLA4	chr2:204731509-204738683	1	16	12	8
ETS1	chr11:128328656-128393205	1	182	157	94
IKZF1	chr7:50217647-50482800	1	682	511	275
IRF5	chr7:128543678-128771955	5	574	335	231
IRF7	chr11:557342-692384	9	370	290	221
ITGAM	chr16:31240000-31400000	2	222	185	116
JAZF1	chr7:27803576-28263475	3	1370	1001	135
Lyn	chr8:56790386-56987446	2	478	284	260
MECP2-IRAK	chrX:153264333-153387422	2	100	160	82
MSH5	chr6:31687008-31762843	9	150	106	148
NCF2	chr1:183433757-183569756	3	243	214	108
PRDM1	chr6:106530806-106603805	1	203	147	142
PTPN22	chr1:114356438-114415334	2	67	72	28
PTTG1	chr5:159846208-159914240	2	228	125	159
PXK	chr3:58316617-58411878	1	323	188	85
SCUBE1	chr22:43491269-43787833	5	1244	669	246
STAT4	chr2:191892306-192017055	1	216	158	59
TLR7_TLR8	chrX:12880079-12945759	2	123	93	56
TNFAIP3	chr6:138000000-138310000	1	779	546	378
TNFSF4	chr1:173147870-173399734	2	598	355	128
TNIP1	chr5:150398485-150470484	2	294	135	151
UBE2L3	chr22:21904834-21987621	2	127	105	66
XKR6	chr8:10737657-11059875	1	1292	778	174
25 LD blocks	4,366,958 bp	66	12213	8005	4030

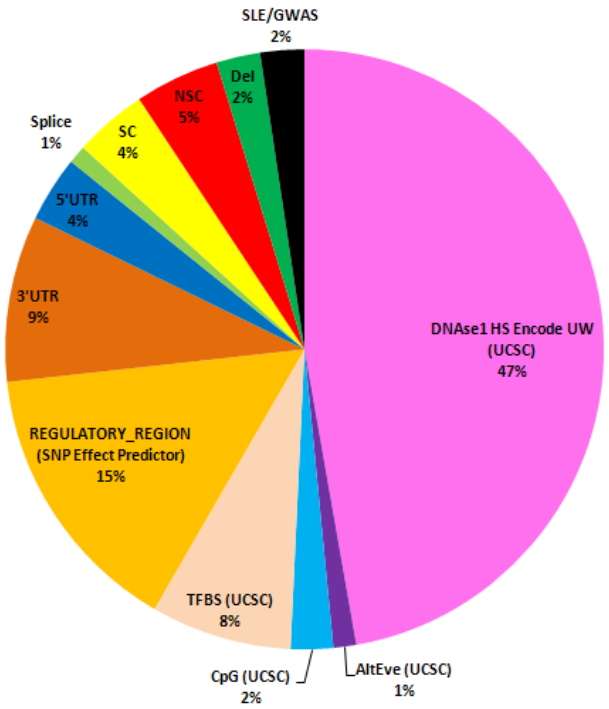
Characteristics of sequence variations identified by targeted sequencing



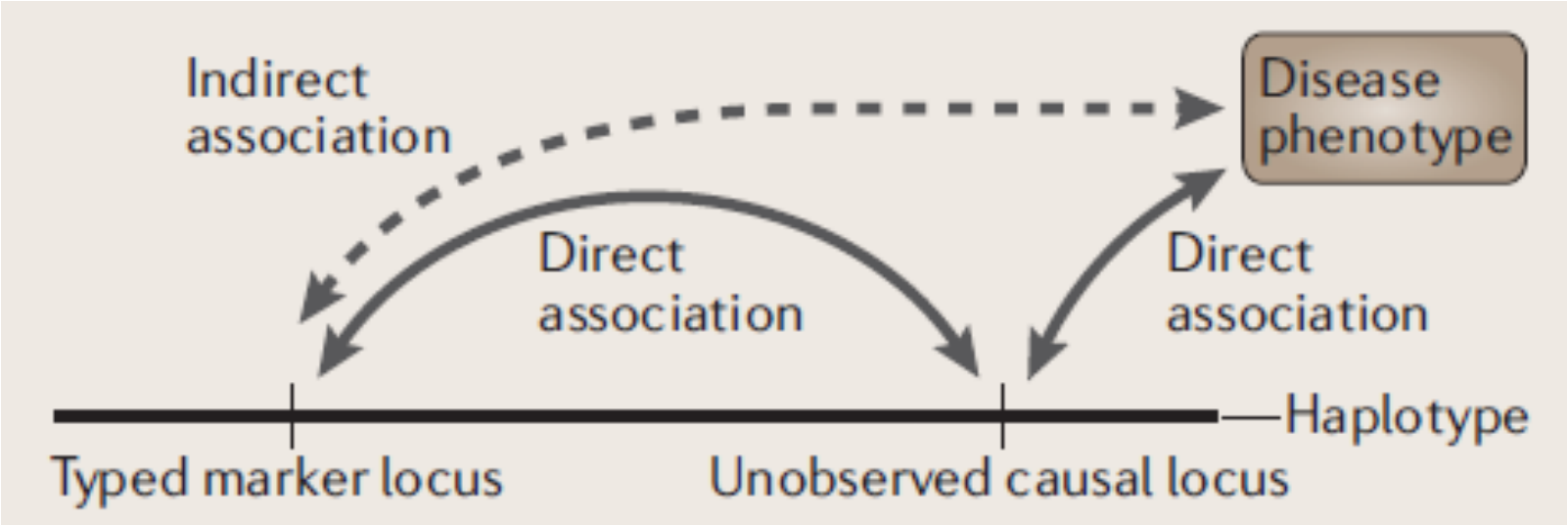


Potentially functional variations





Defining allelic architecture of SLE risk loci

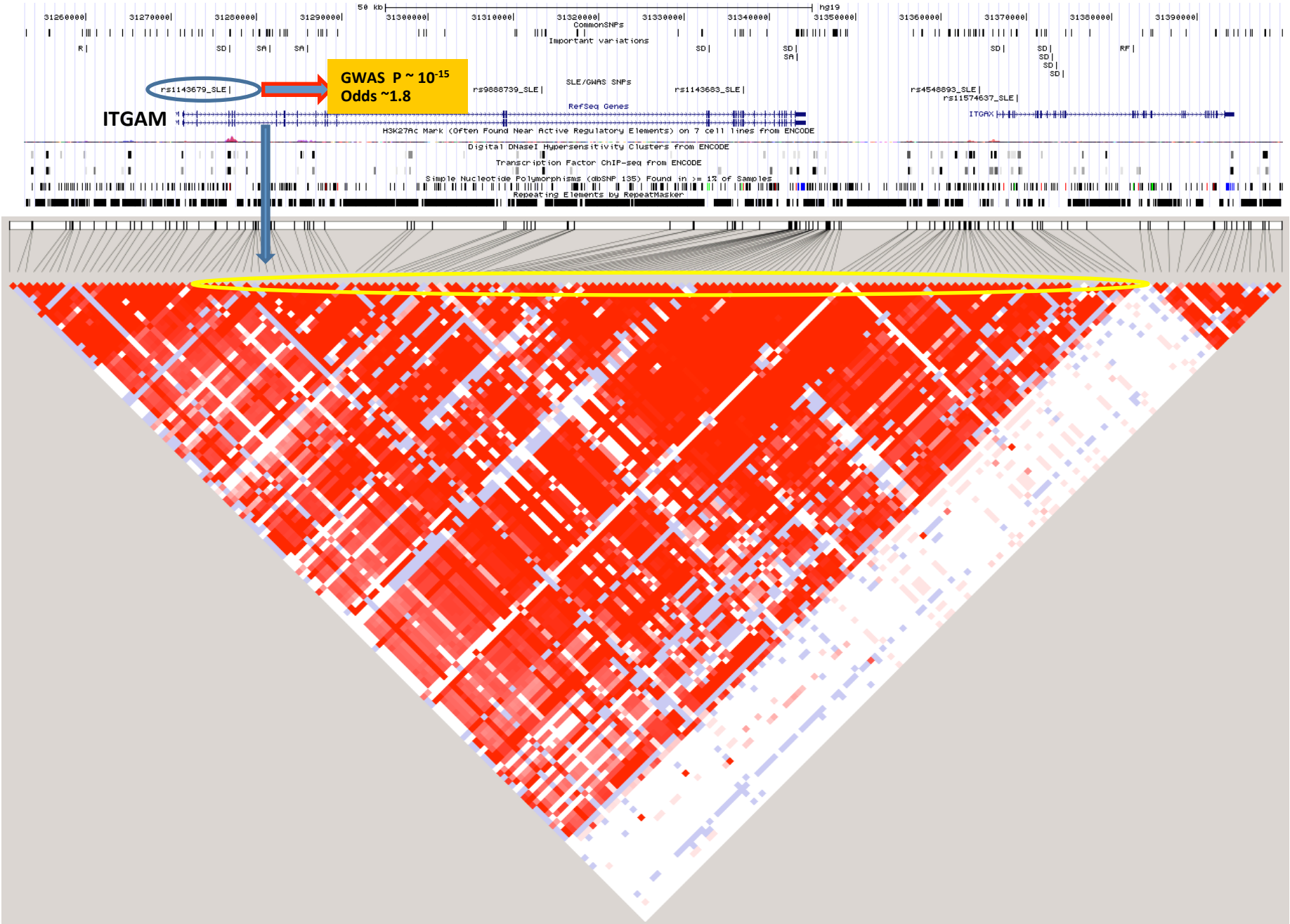


From Balding, *Nat. Rev. Genetics,* 2006

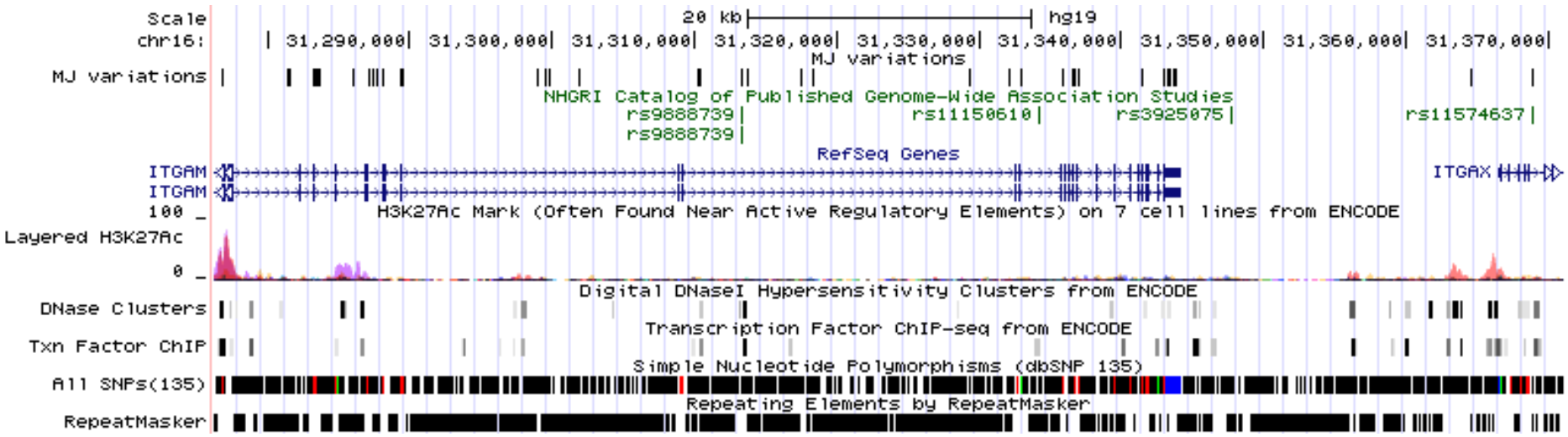
Analytical strategy

- 107 SLE cases and 85 controls (selected for high content of risk alleles)
- Identify genomic segment in tight LD with fine-mapped tagging SNPs
- Identify all potentially functional SNPs in tight LD with tagging SNP
- Define all haplotypes and develop median neighbor joining network of alleles
- Overlay functional and disease-association parameters

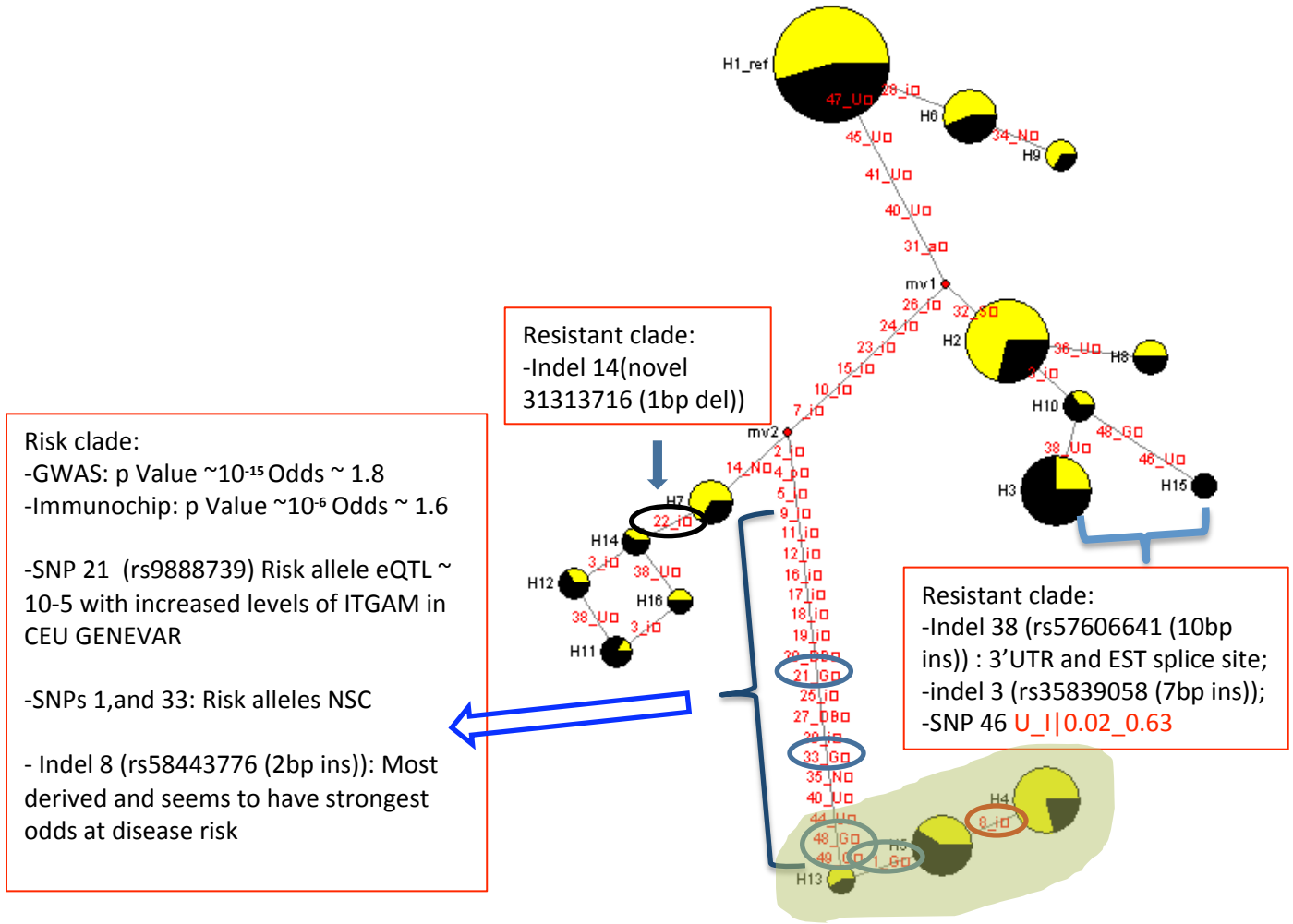
ITGAM (CD11b) is located in a strong LD block spanning a ~120 KB



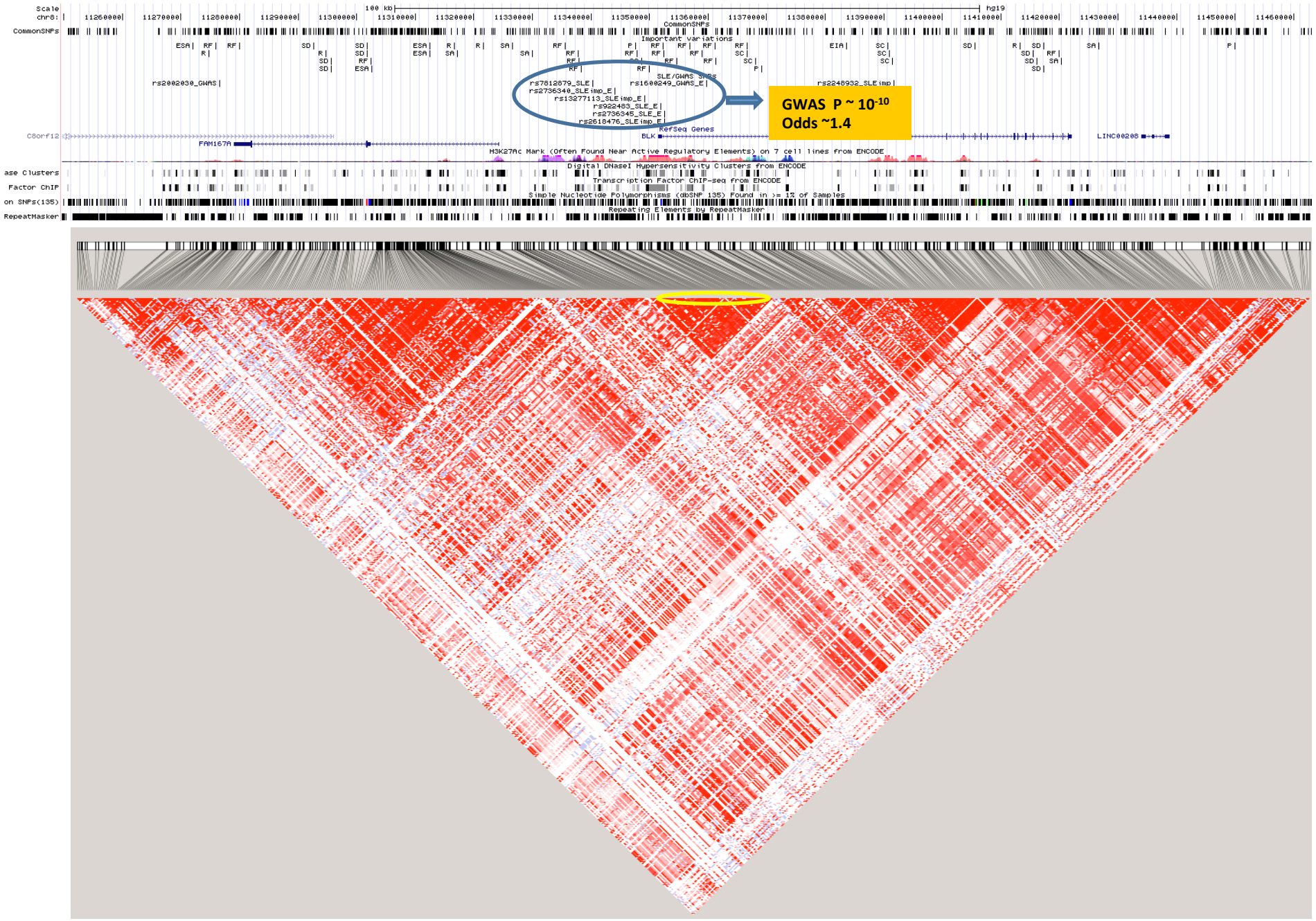
46 potentially functional SNVs in tight LD with ITGAM tagging SNP

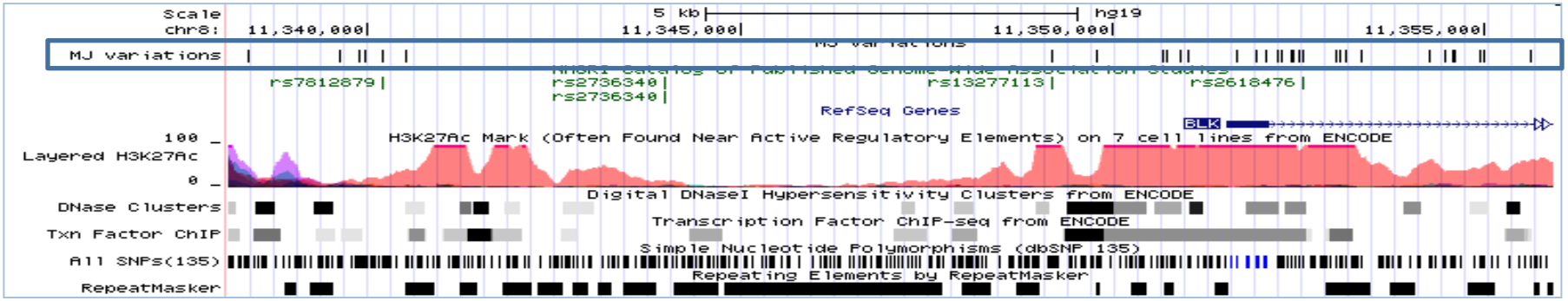


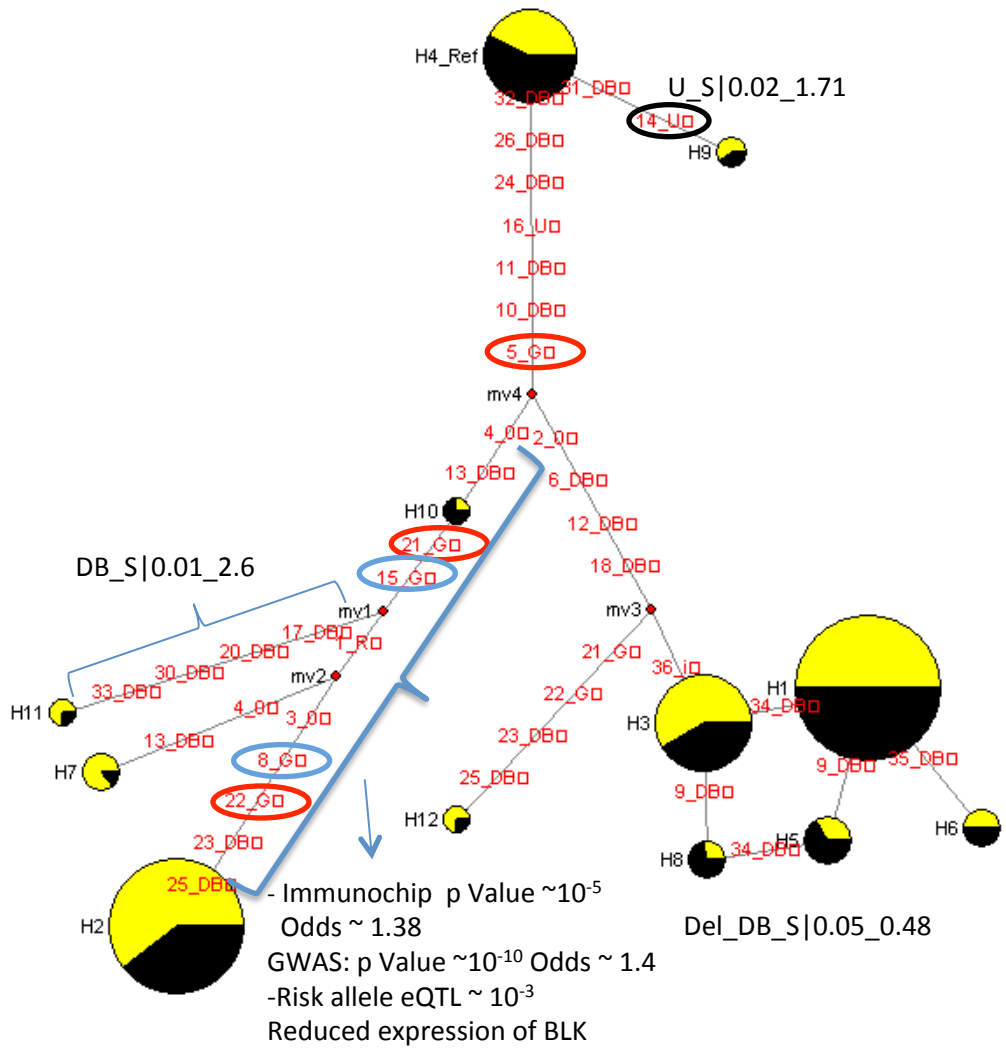
Sixteen ITGAM haplotypes formed among 384 chromosomes



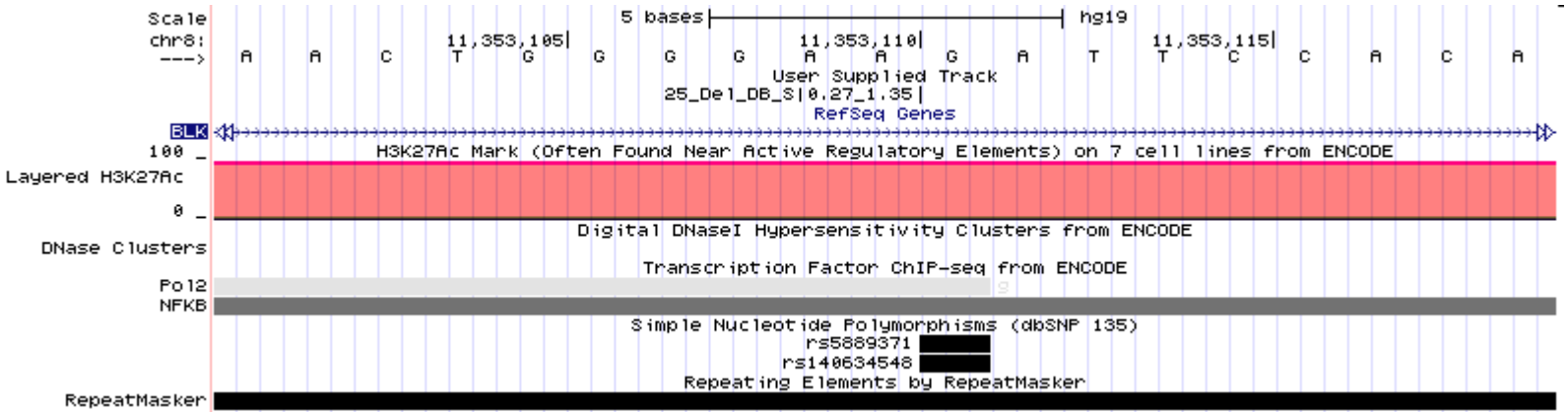
B lymphocyte kinase (BLK) located in genomic segment with multiple LD blocks

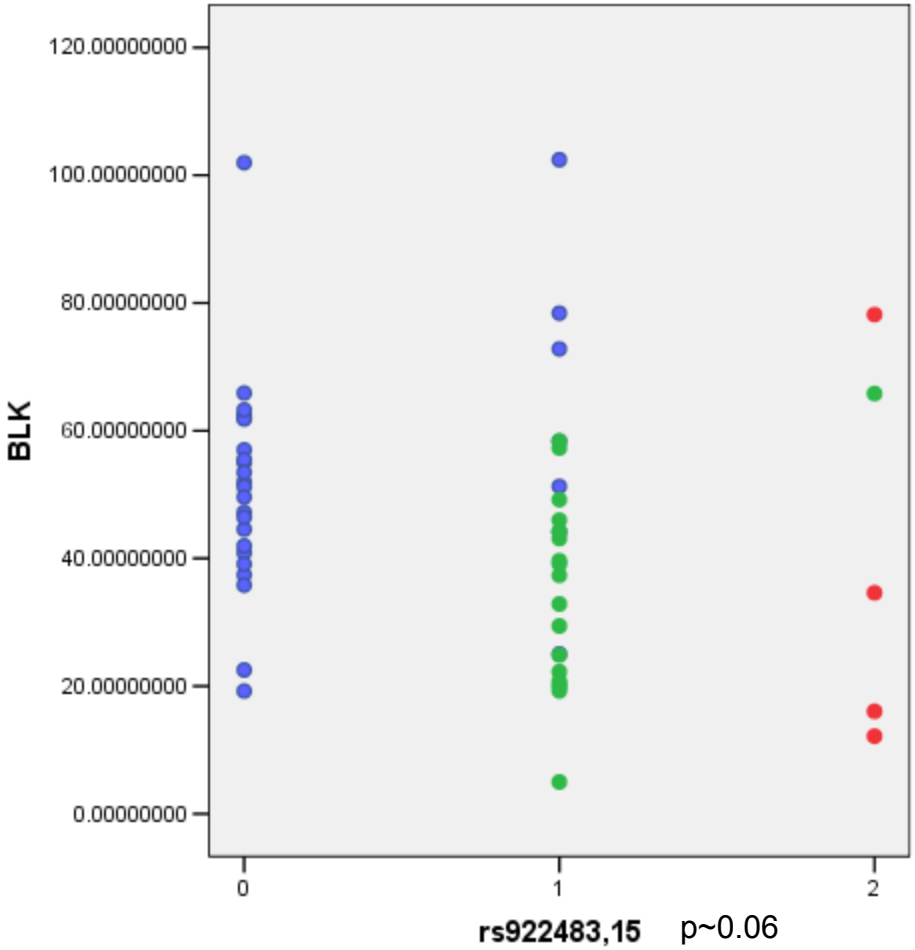






- Immunochip p Value ~10^{-5}
  Odds ~ 1.38
GWAS: p Value ~10^{-10} Odds ~ 1.4
-Risk allele eQTL ~ 10^{-3}
Reduced expression of BLK





p~0.06

rs5889371,25 G/-

0

1

2

$p \sim 10^{-3}$

Summary and Future Directions

- **GWAS tagging SNPs often identify multiple functional haplotypes that form a complex series of risk and resistant haplotypes**
  - We are expanding the study set to >1500 patients and controls and
    - Assess the association of specific haplotypes with disease
    - Assess the role of minor variants in disease pathogenesis
  - We will integrate data from the 1000 genomes project into the haplotype analysis
- **Functional variations responsible for disease allele associations are commonly associated with combinations of SNVs**
  - We will expand the sample size of the eQTL panel
  - We will develop dynamic eQTL panels for monocyte cell lineages
  - We will develop a detailed isoform, ncRNA, and miRNA database
- **eQTL SNPs in tight LD with tagging SNPs can be associated with transcriptional changes in genes that are not in tight LD**

Research fellows, students, and collaborators

Post doctoral fellows

Sun-Hee Hwang
Ekta Rai
Shaheen Khan
Prithvi Raj
Pratikumar Doshi
Ran Song
Andrew Wang

Technical Staff

Chaoying Liang
Carlos Arana
Ferdicia Carr-Johnson
Ben Wakeland

Students

Chukwuemika Aroh

UT Collaborators

Quan Zhen Li
Chandra Mohan
Jose Casco
Laurie Davis
David Karp
Nan Yan
Jeffrey Kahn

OMRF

Patrick Gaffney
Graham Wiley
Swapan Nath
Judith James

Penn State

Nancy Olsen

Saint-Luc Universite Belgium

Bernerd Lauwerys