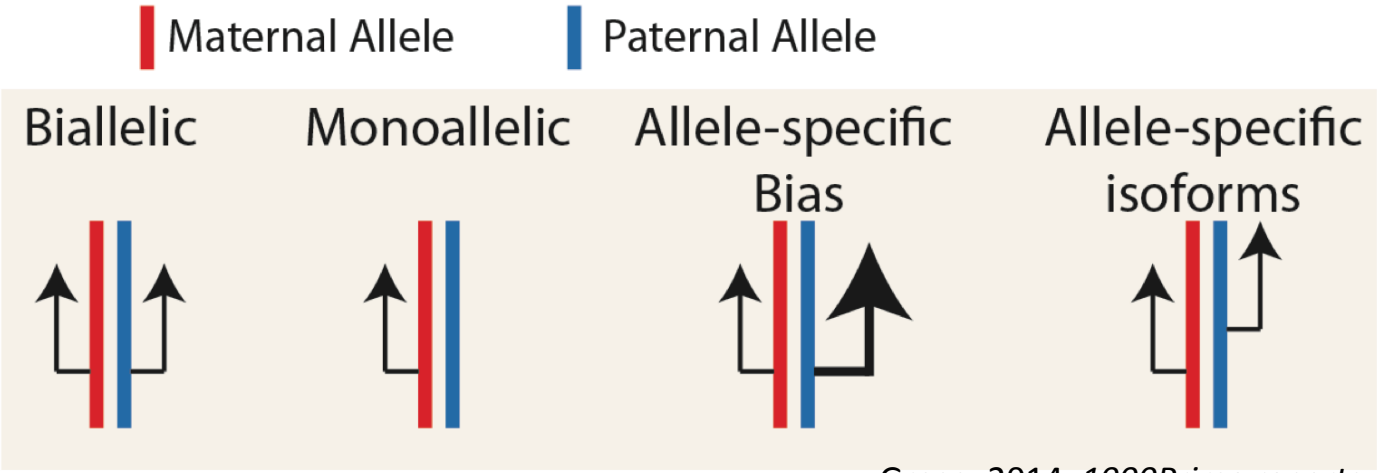
# Studying allele specific expression to capture condition-specific genetic  regulatory effects

# Overview

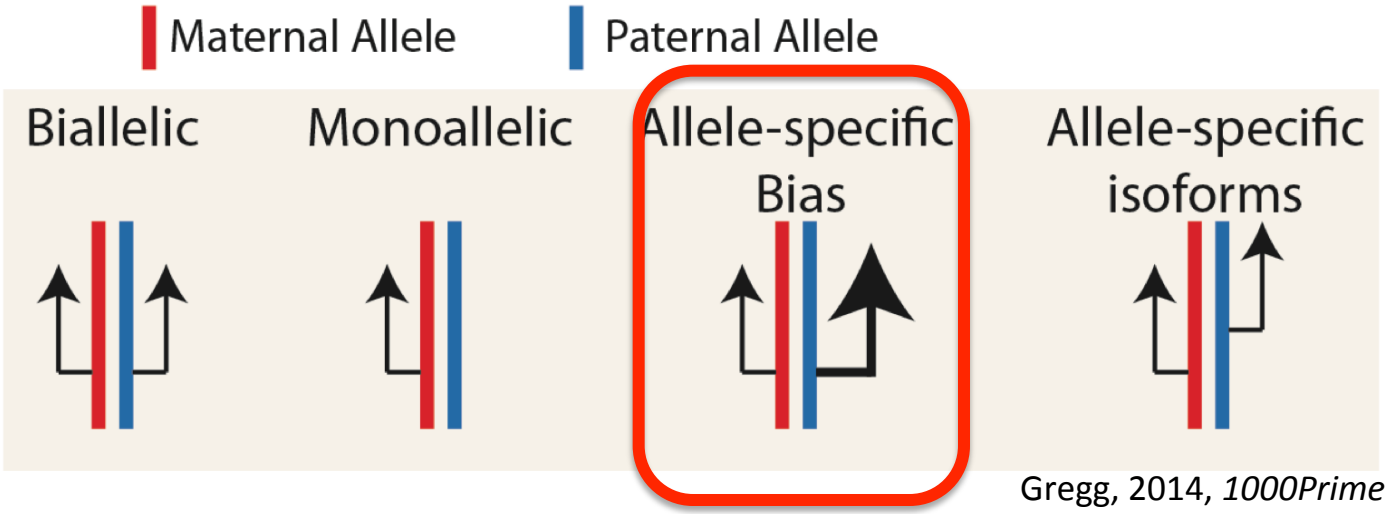
- Quick intro to allele specific expression (again)
- New stuff in pipeline
- TASC ASE
- HLA class II
- STAT pilot

# Types of allelic expression



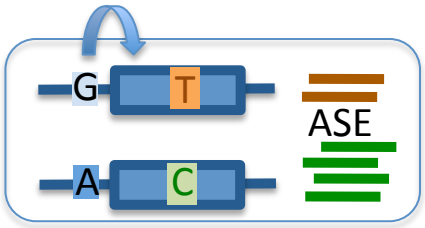

Gregg, 2014, *1000Prime reports*

# Types of allelic expression

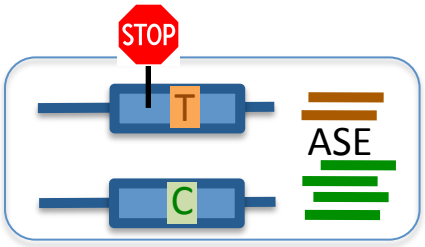



Gregg, 2014, *1000Prime reports*

Genetic regulatory variation

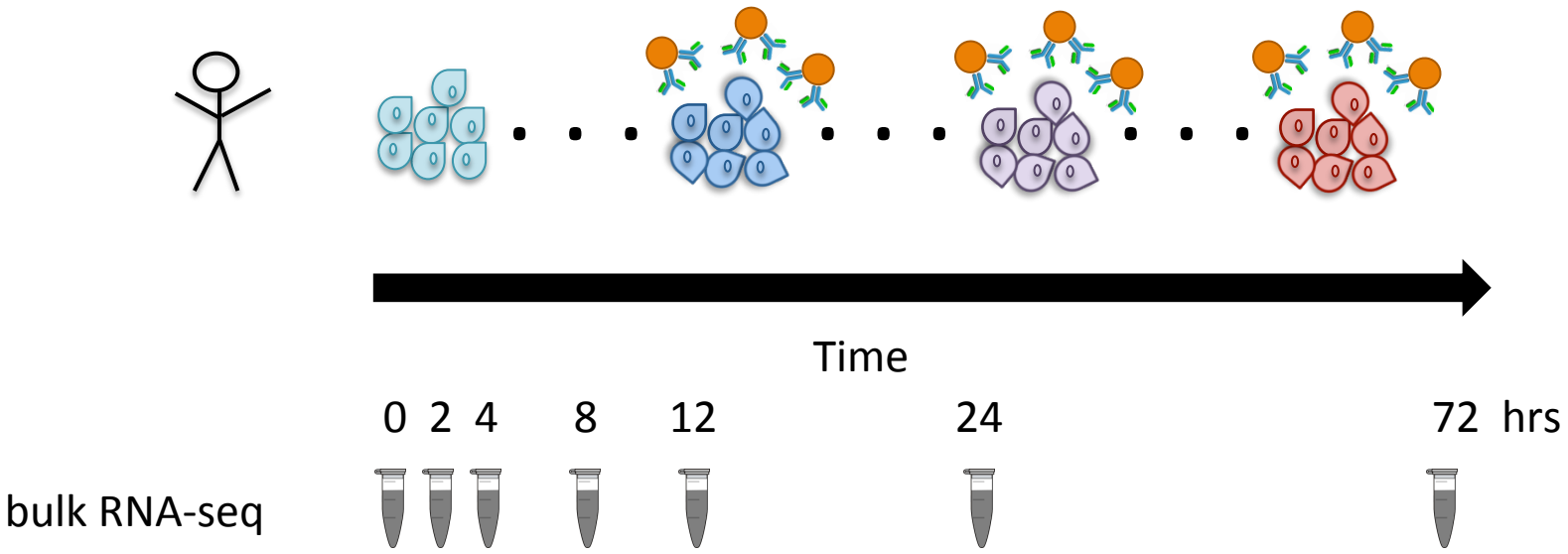



Nonsense mediated decay



# ASE in response to stimuli

CD4+ $T_{MEM}$ cells stimulated with anti-CD3/CD28 beads

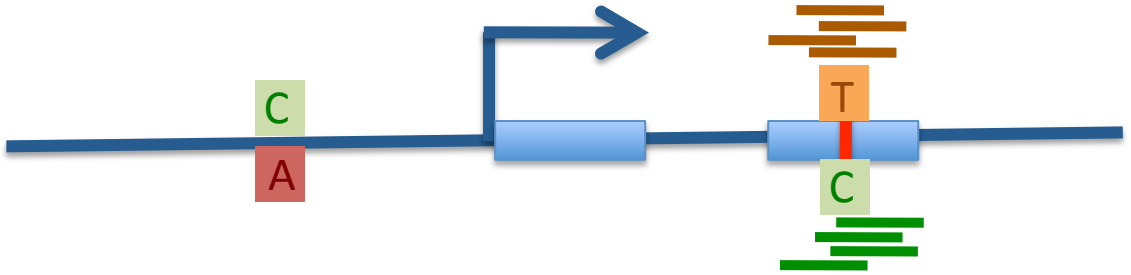



<u>Objective</u>

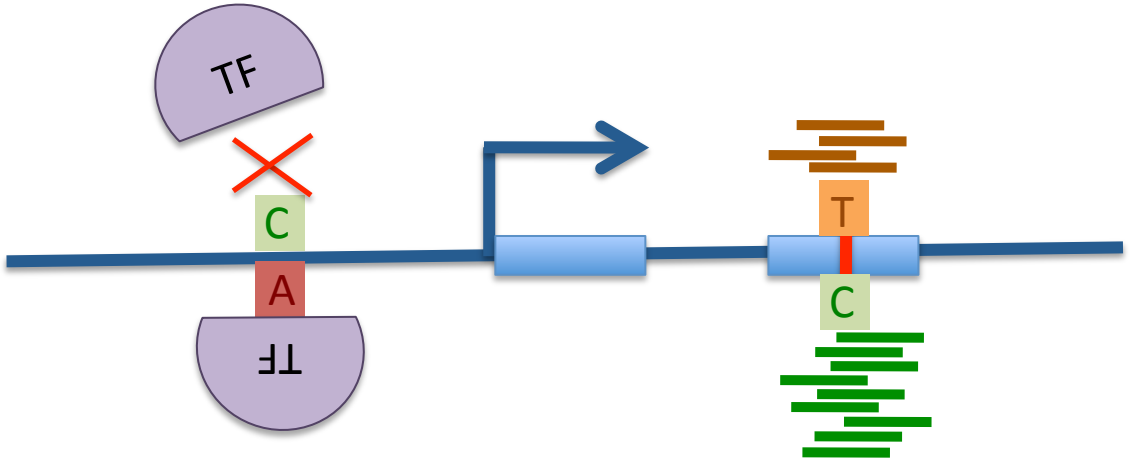
Find time-dependent allele specific expression to capture activated and inactivated genetic regulatory effects upon stimulation

# Condition specific allelic expression

Condition 1



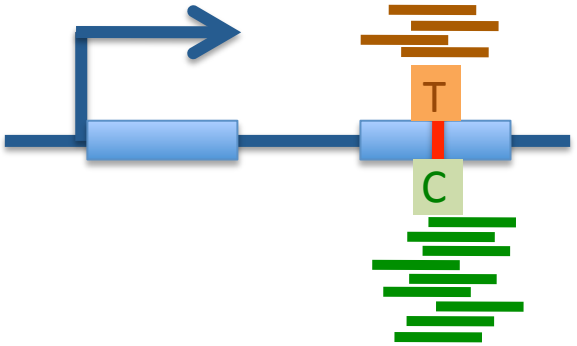

Condition 2



# Assessing allele-specific expression in low input RNA-seq

1) **Input data:**
   - Genotypes: called variants from RNA-seq
   - RNA-seq uniquely aligned reads

2) **Calculate read counts over heterozygous sites**
   - Pipeline from Kukurba *et al.* 2014




3) **Filter and correct for sources of error**
   - Require BAS
   - Remove mapping bias sites

Ref ratio: 4/(4+9) = 0.31

4) **Analyze data**
   - Reference ratio : reference counts / total counts

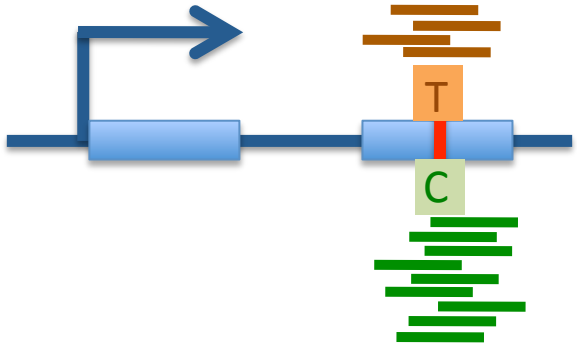
# Assessing allele-specific expression in low input RNA-seq

1) **Input data:**
   - Genotypes: called variants from RNA-seq
   - RNA-seq uniquely aligned reads

☀ pipeline validation

2) **Calculate read counts over heterozygous sites**
   - Pipeline from Kukurba *et al.* 2014

☀ weighted duplicate counts




3) **Filter and correct for sources of error**
   - Require BAS
   - Remove mapping bias sites

Ref ratio: 4/(4+9) = 0.31

☀ map to personalized masked genome

4) **Analyze data**
   - Reference ratio : reference counts / total counts

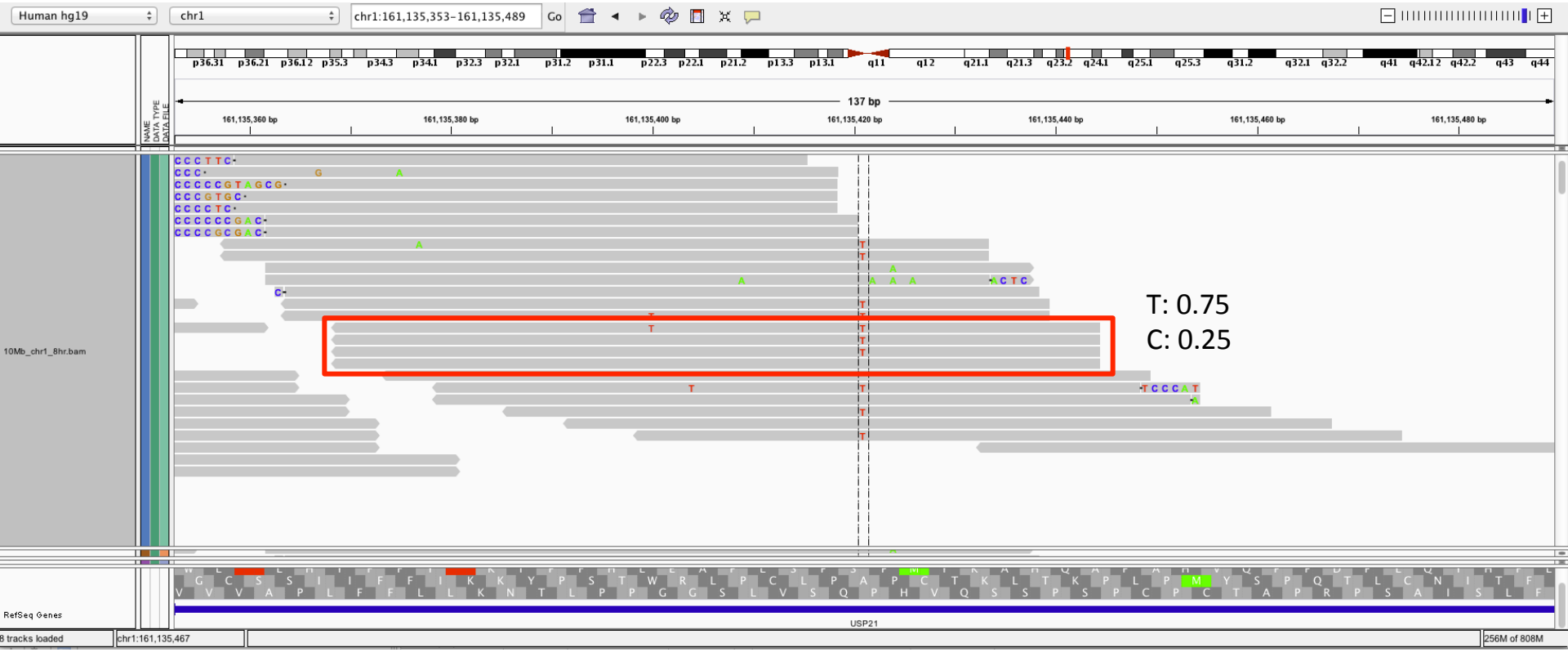
# Validating variant calling pipeline using NA12878

- "reference genotypes" are variants called by best practices of GATK from WGS
- mRNA-seq data from Kilpinen *et al.,*2013: ~18M 50bp PE reads
- Differences with our RNA-seq datasets: less depth since only one condition assayed, paired-end data (so in reality ~36M reads), shorter reads

| | GATK and Piskol filters | filts + overlap dbSNP | ASE tested* |
|---|---|---|---|
| **Total Hets Called** | 13176 | 12566 | 4876 |
| **# in NA12878 calls** | 12000 | 11976 | 4631 |
| **with same alt allele** | 11981 | 11957 | 4629 |
| **with same genotype** | 11911 | 11887 | 4624 |
| **% concordant** | 90.4 | 94.6 | 94.8 |
| **% FP** | 9.6 | 5.4 | 5.2 |

*Both alleles seen + at least 10 dup weighted counts per site

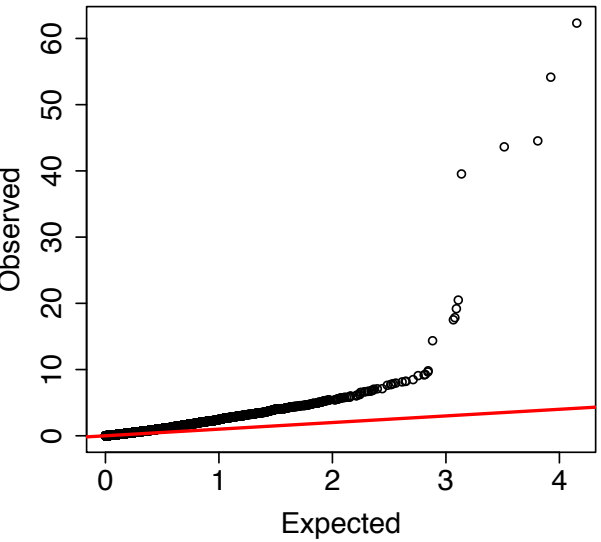
# Weighted duplicate counts



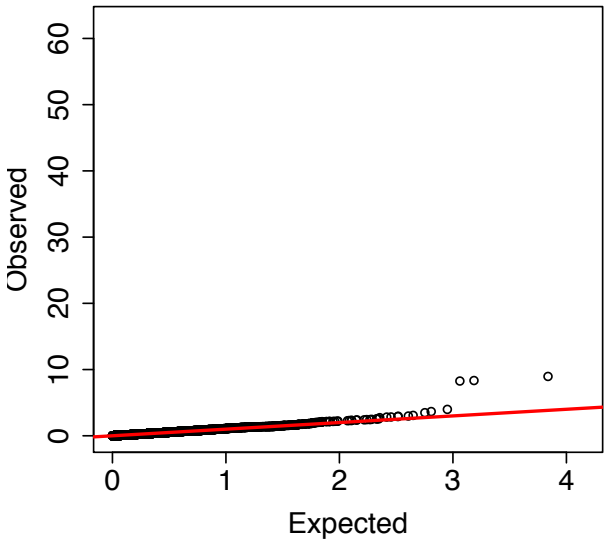

| With dups | Weighted dups |
|---|---|
| C: 7 | C: 6.25 |
| T: 10 | T: 5.75 |

# Logistic regression for time point vs allele counts

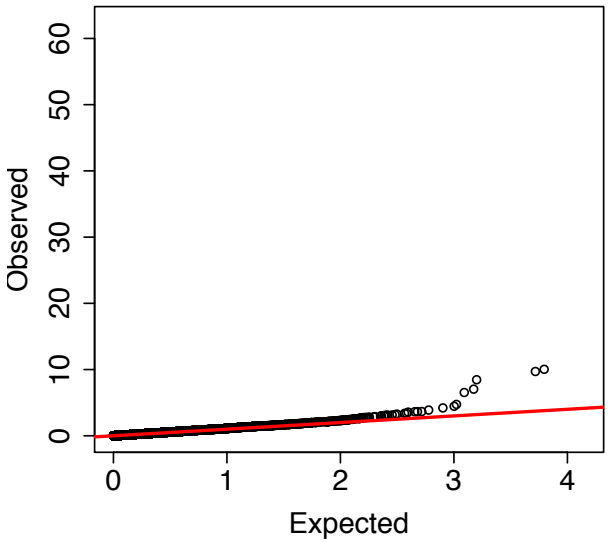
With duplicates
lambda = 3.1




No duplicates
lambda = 1.1

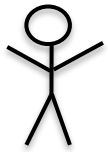



Weighted duplicates
lambda = 1.14




*time as quantitative variable

# Map to personalized masked genome



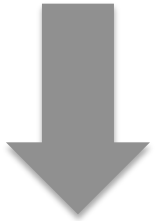
TACGCGATTC**G**GATCCGATAGC

TACGCGATTC**T**GATCCGATAGC

heterozygous site

reference genome

TACGCGATTC**G**GATCCGATAGC

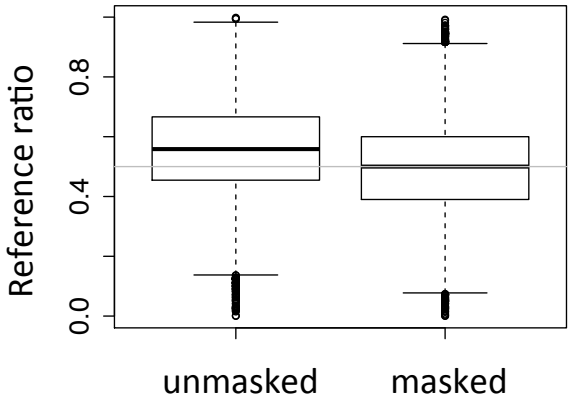


TACGCGATTC**N**GATCCGATAGC

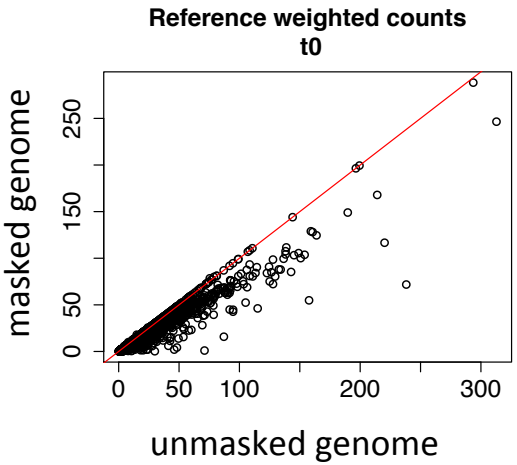
masked reference genome

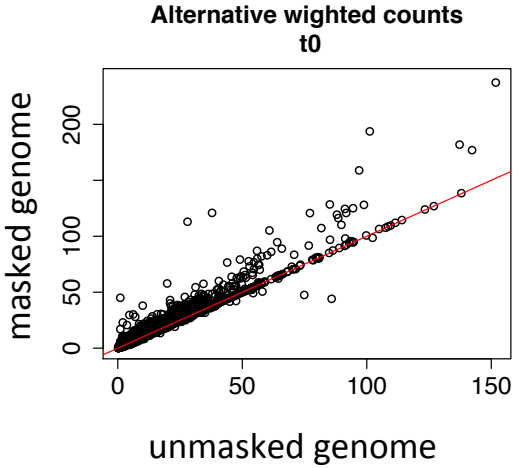
# Map to personalized masked genome

Reference ratio distribution bias disappears




Reference allele tends to lose reads and alternative tends to gain reads

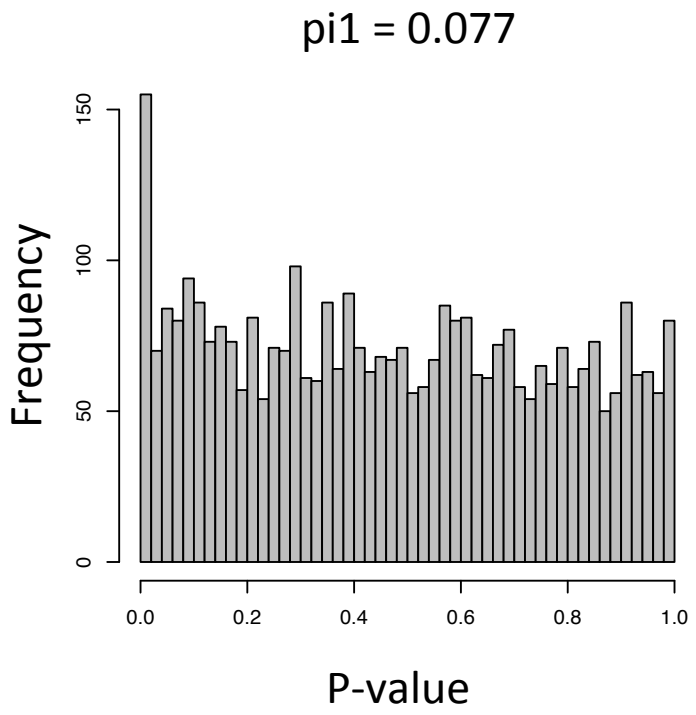


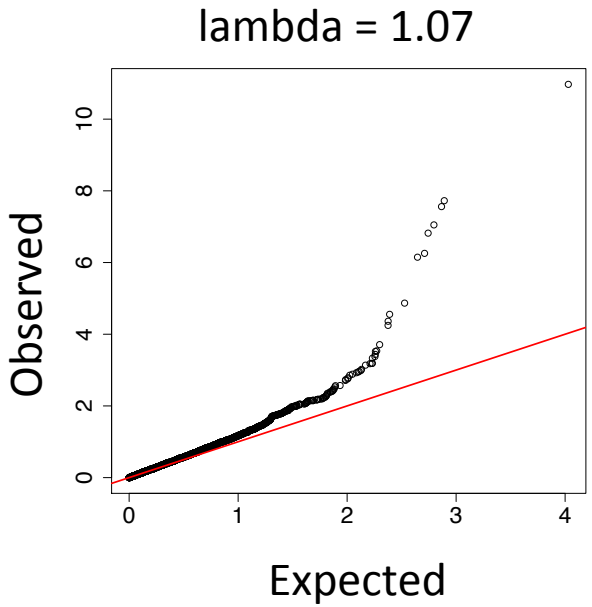



# Logistic regression time vs weighted allele counts

3,578 tested heterozygous sites
(BAS, min 10 weighted counts in all time points, masked genome)

17 significant SNPs (FDR < 10%), spanning 13 protein coding genes, 1 non-coding site

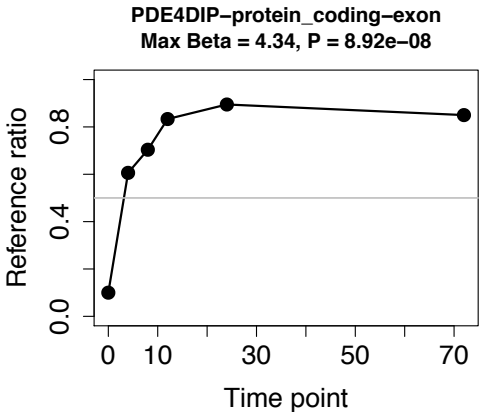


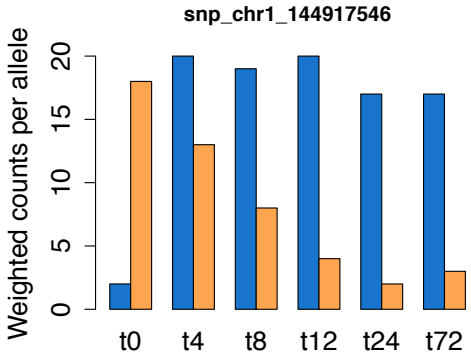




*time as qualitative variable

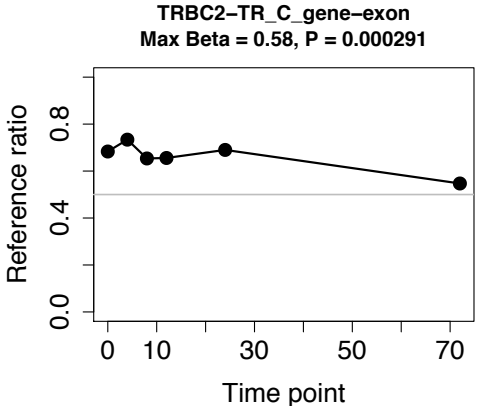
# TASC ASE examples

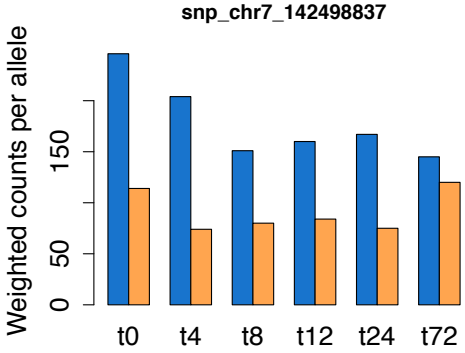
Anchors enzyme to Golgi
Involved in eosonphilia



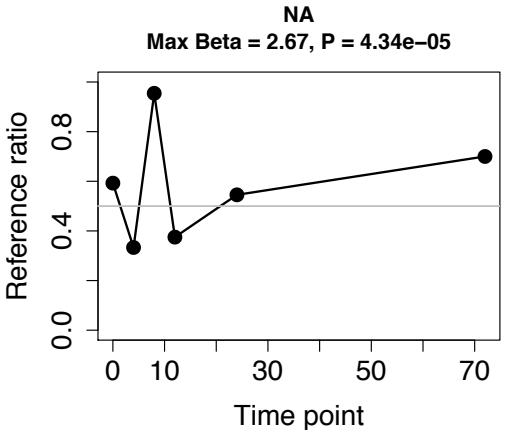



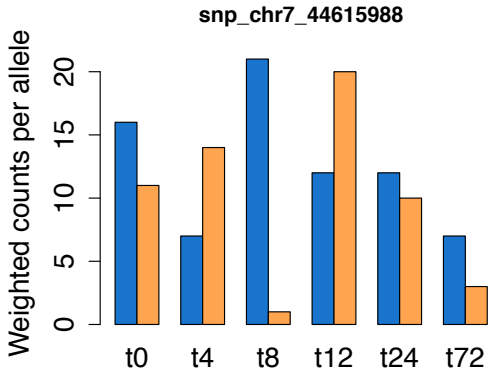

T cell receptor beta constant 2



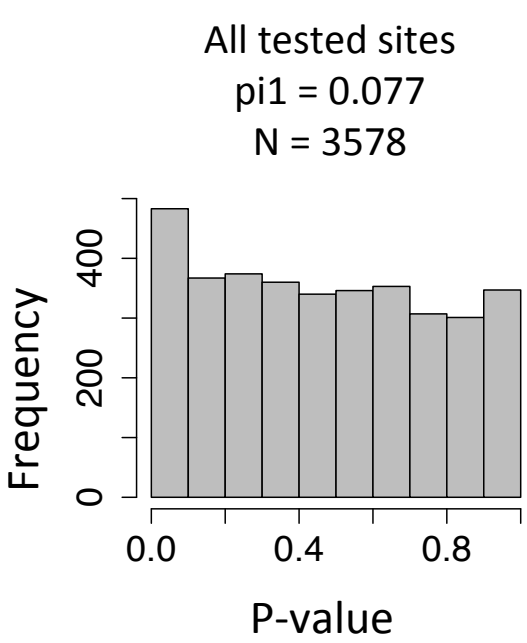

Non-coding
Between 2 genes
In Dnase I HS, 3 cell-types

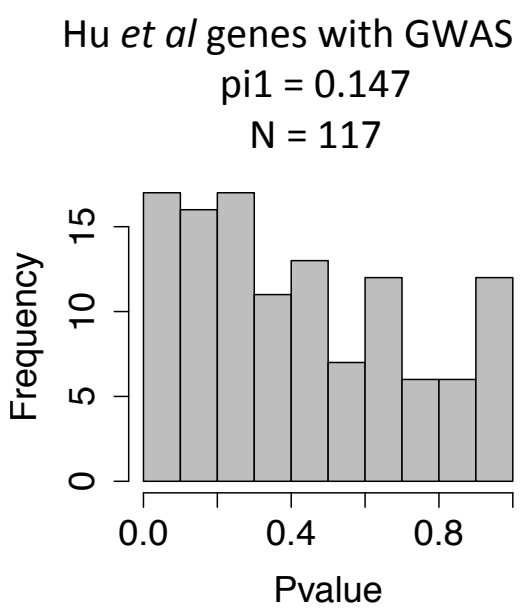


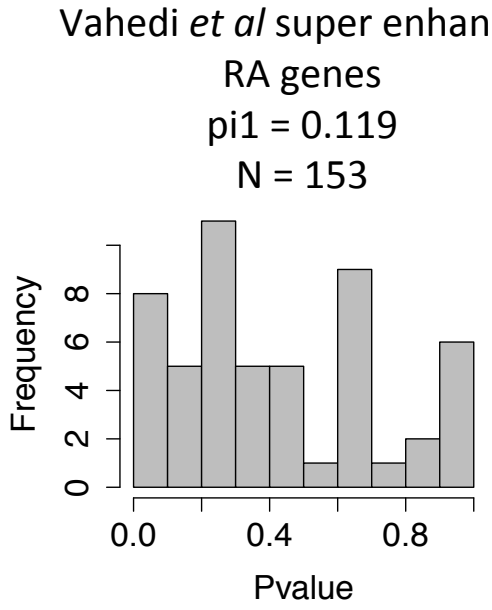




*2 SNPs close to each other in 1 gene, and 1 case for chrX

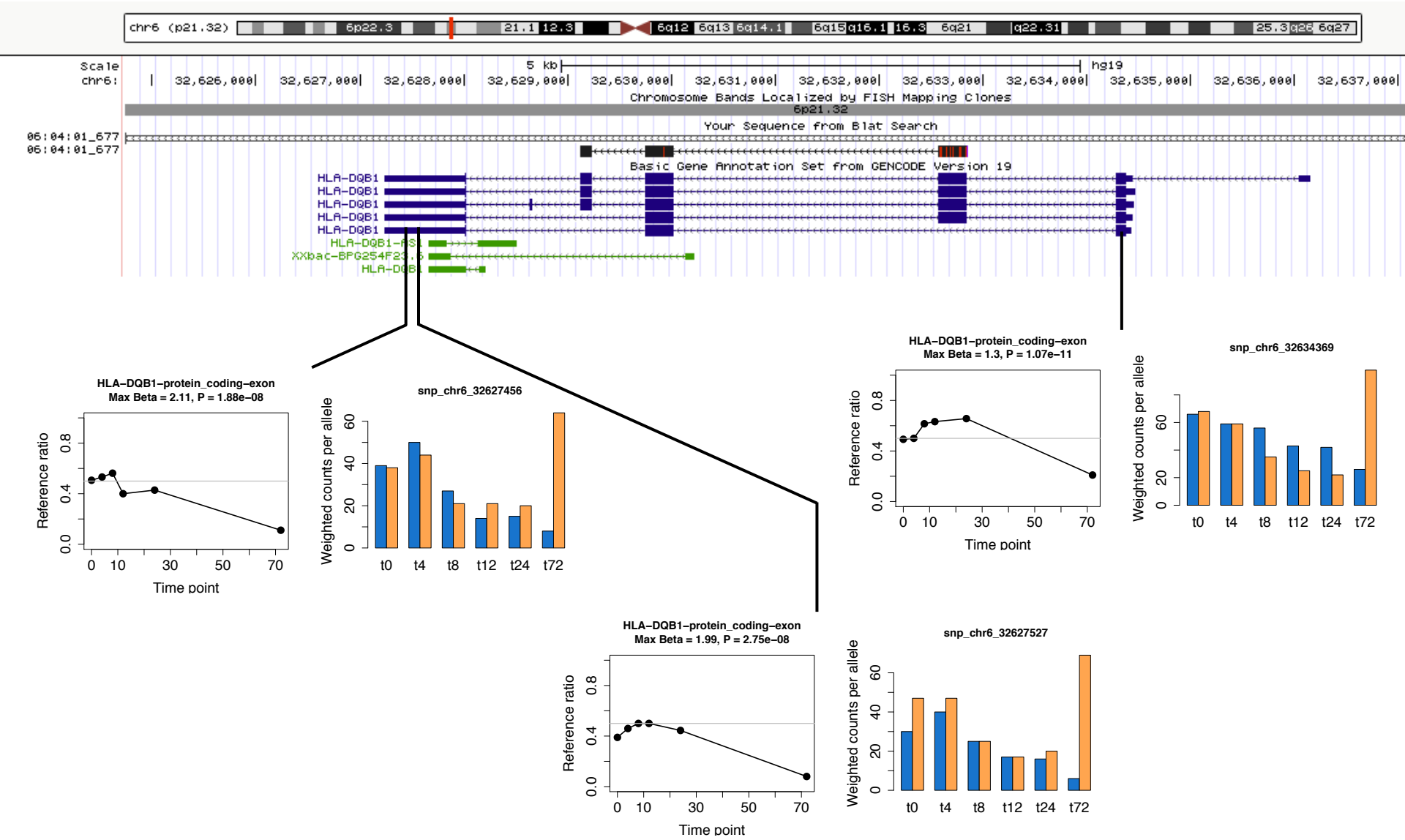
# Enrichment of low P-values for RA genes



# HLA-DQB1: top TASC ASE signal



# HLA class II genotyping

Clinical genotyping

Common alleles:
DQB1*06:04
DQB1*05:03

Rare alleles:

DQB1*06:34/36/38/39/52/58/
85/86/89/93/135/155/158N

DQB1*05:06/08/10/13/15/16/23/
24/38/39/40/41N/42/43/50/56/
67/78

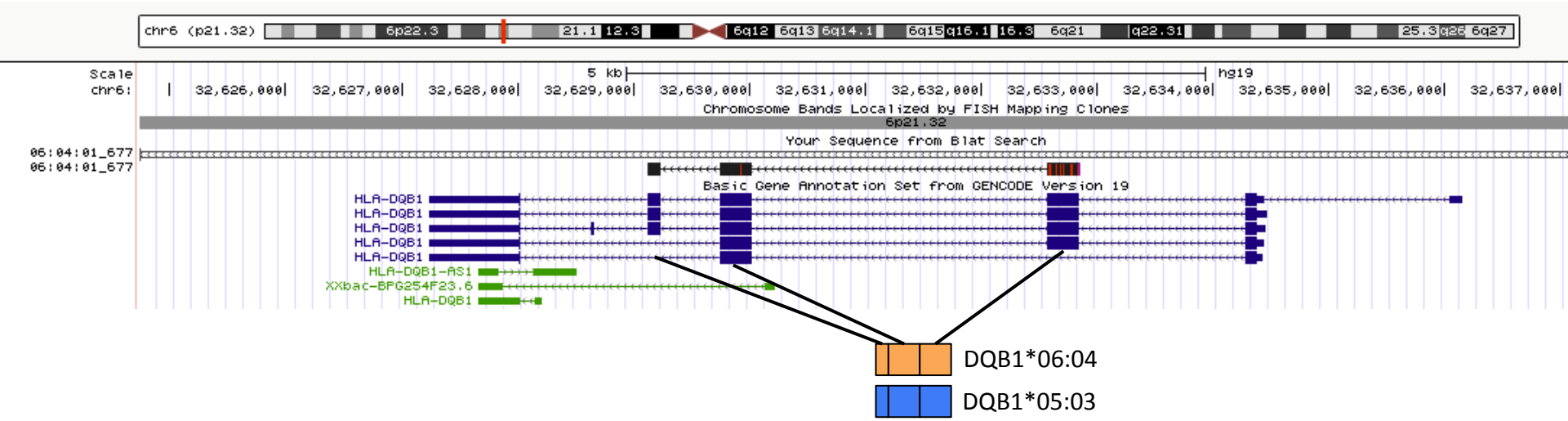
1) Fom each common allele, chose longest cDNA sequence available in ATHLATES IMGT/HLA db

2) Aligned these 2 sequences with BLAST and cut them so that they are the same length and well aligned (677bp, 96% id)

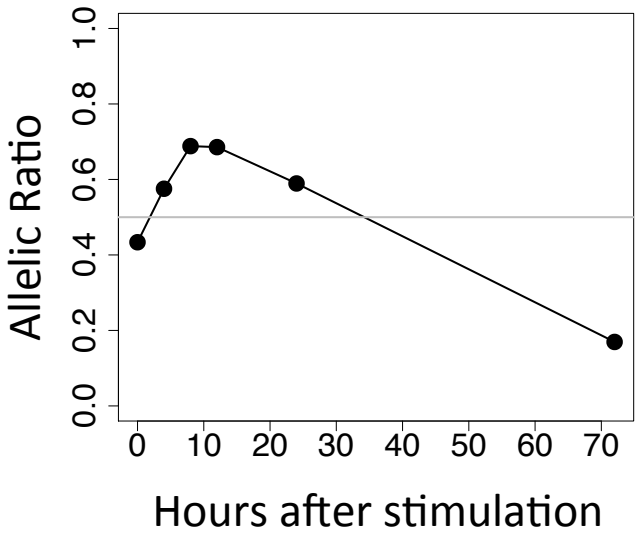
3) Aligned to ref genome with BLAT to get coordinates, and then masked these 3 exons

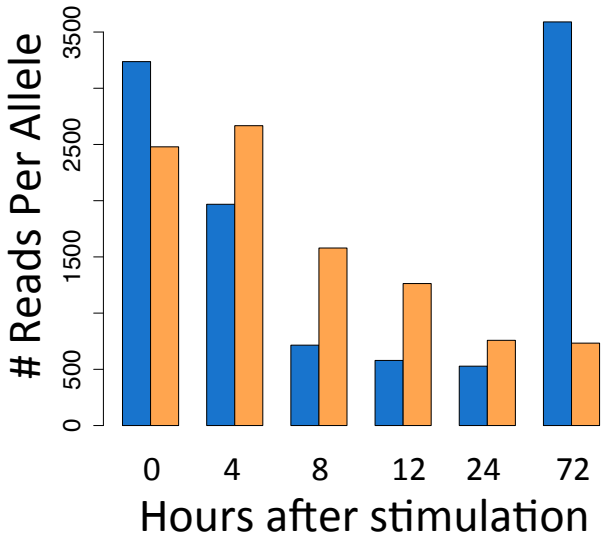
4) Added the 2 cDNA alleles to the masked reg genome and re-mapped all reads

# DQB1 TASC ASE replicates using 3 exons




P < 2.2e-16



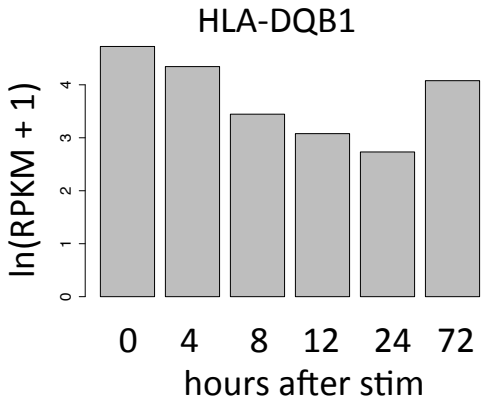



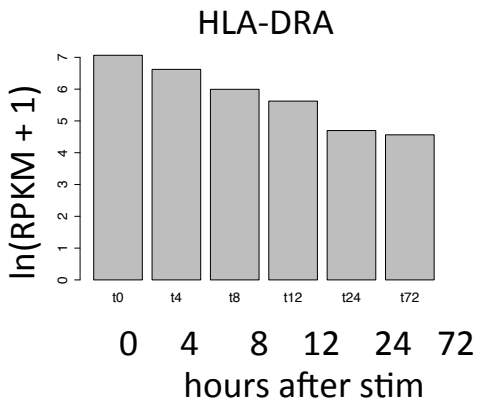

*ASE for DQB1 has been reported in DCs, LCLs and melanoma cell line

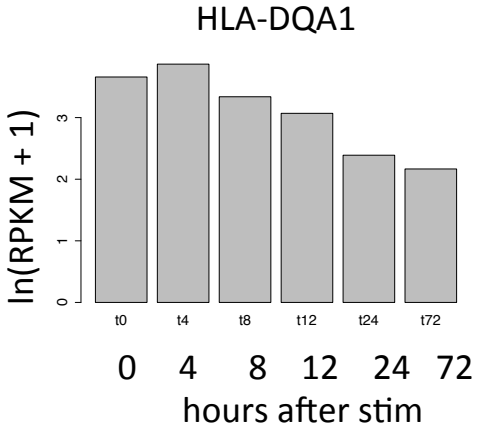
# HLA class II genes in T cells

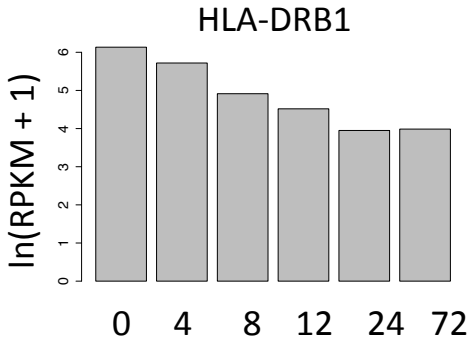
- HLA class II molecules are mainly expressed in professional antigen presenting cells (APCs)
- Normally not in T cells, but in activated human T cells yes (protein level)
- Expression can be induced in non-professional APCs

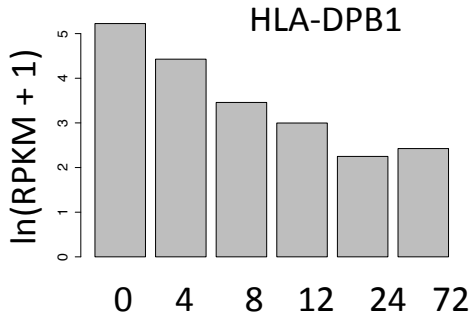
- Transcript expression values are counter-intuitive in our data:



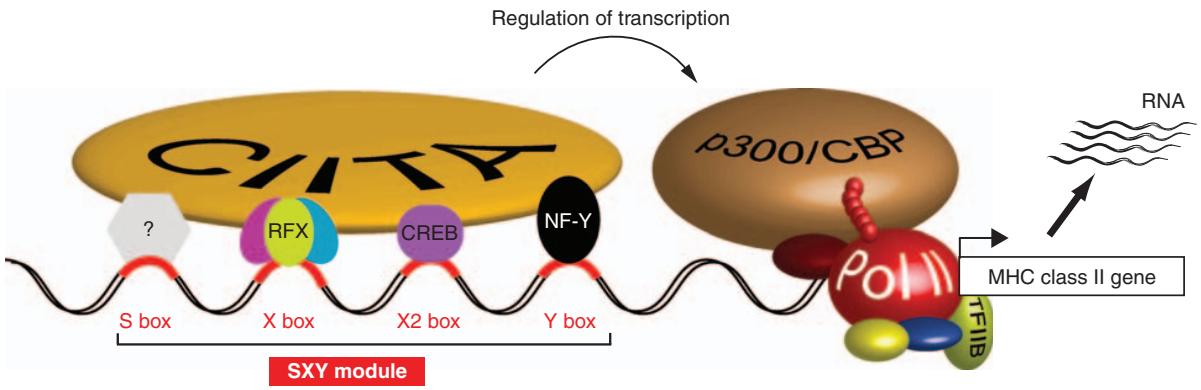



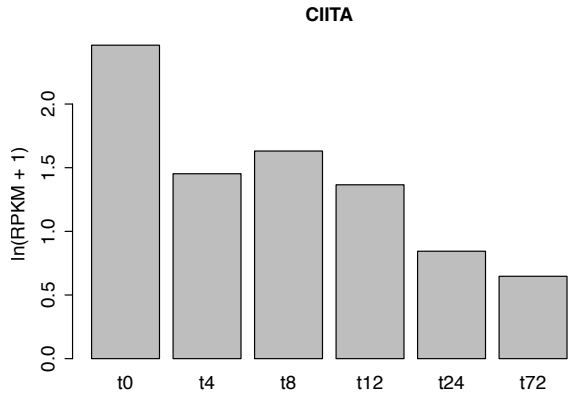



# CIITA: co-activator of HLA class II genes

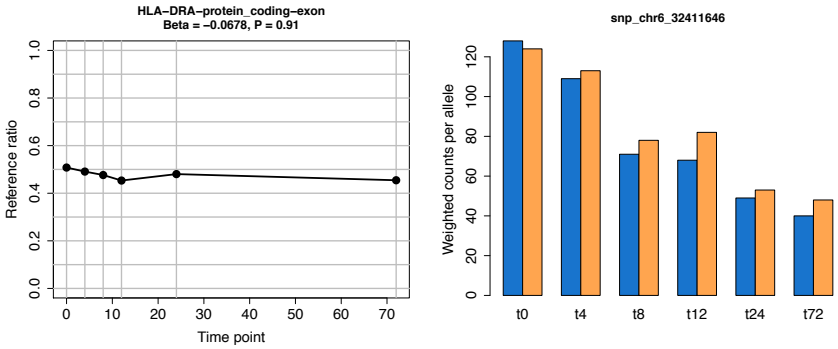



Handunetthi *et al* 2010

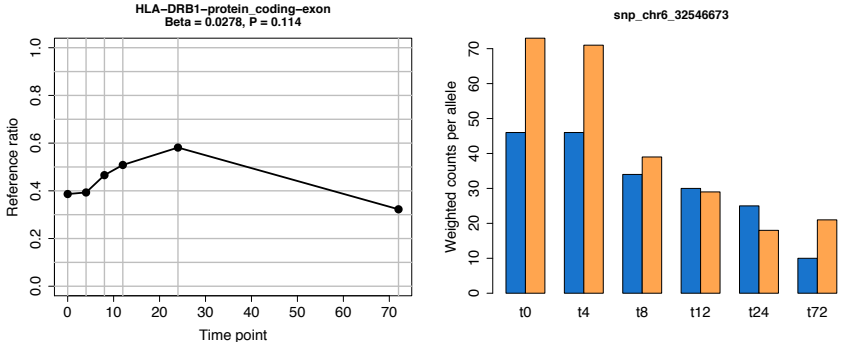


# No significant TASC ASE in other HLA class II genes

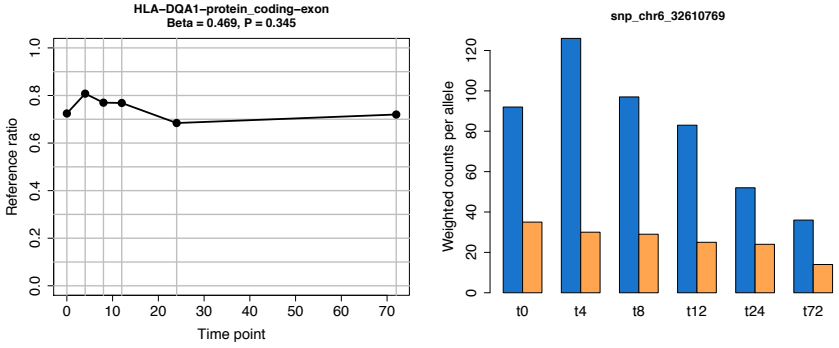
## HLA-DRA



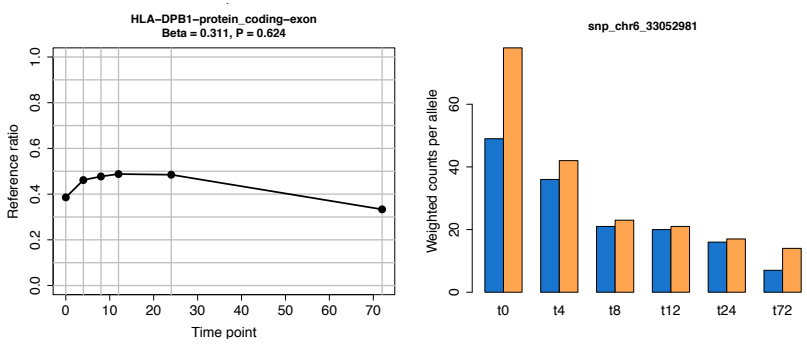

## HLA-DRB1




## HLA-DQA1




## HLA-DPB1



# Next steps for time-dependent ASE and DQB1

- One-off experiment to check if protein expr DQB1 goes up upon CD4+ Tmem stimulation (can we at the same time identify which cellular subtype the signal is coming from?)

- Check transcript levels for HLA class II genes in other T cell datasets (will check in Deepak's)

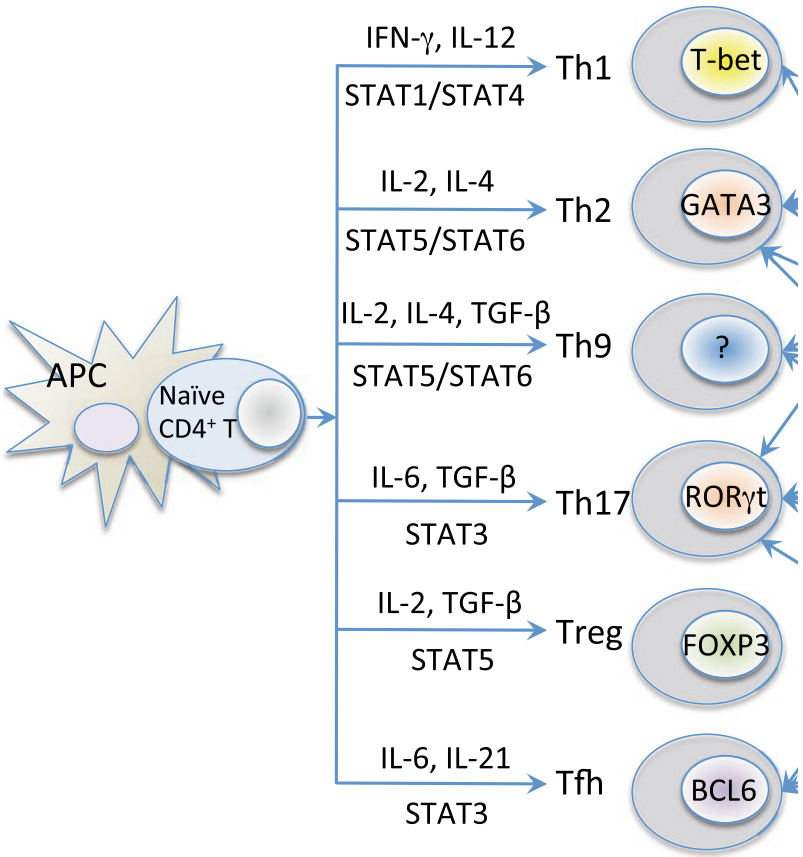
- Make sure DQB1 signal does not come from DQB2 (unlikely)

- Measure TASC ASE in other genes by phasing SNPs and fusing exons if possible (more power for log reg)

- Expand study to many genotyped individuals (chose based on DQB1 genotypes?, same time points or less?)

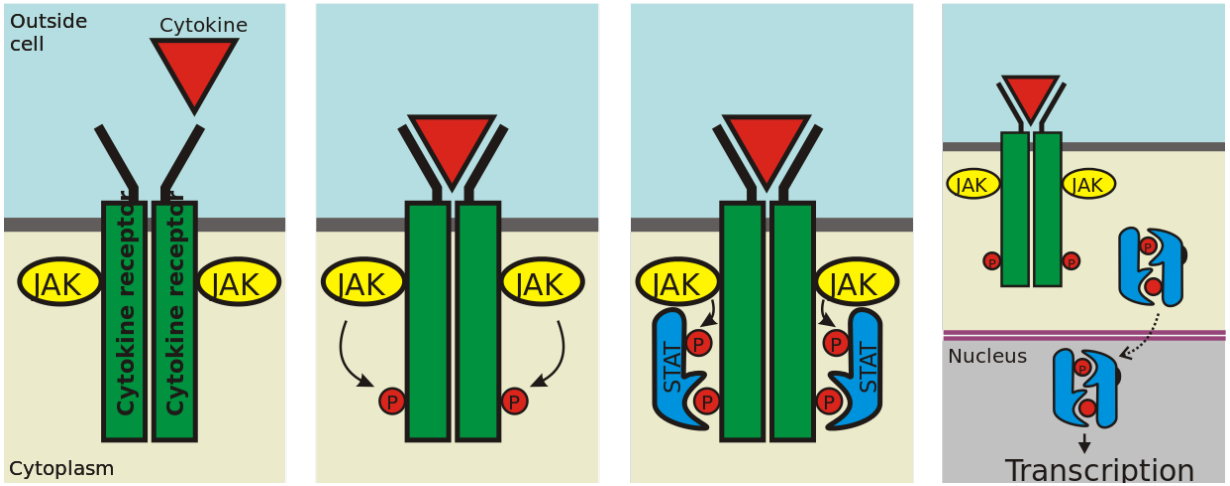
# STAT PILOT

# Signal Transducer and Activator of Transcription (STATs)

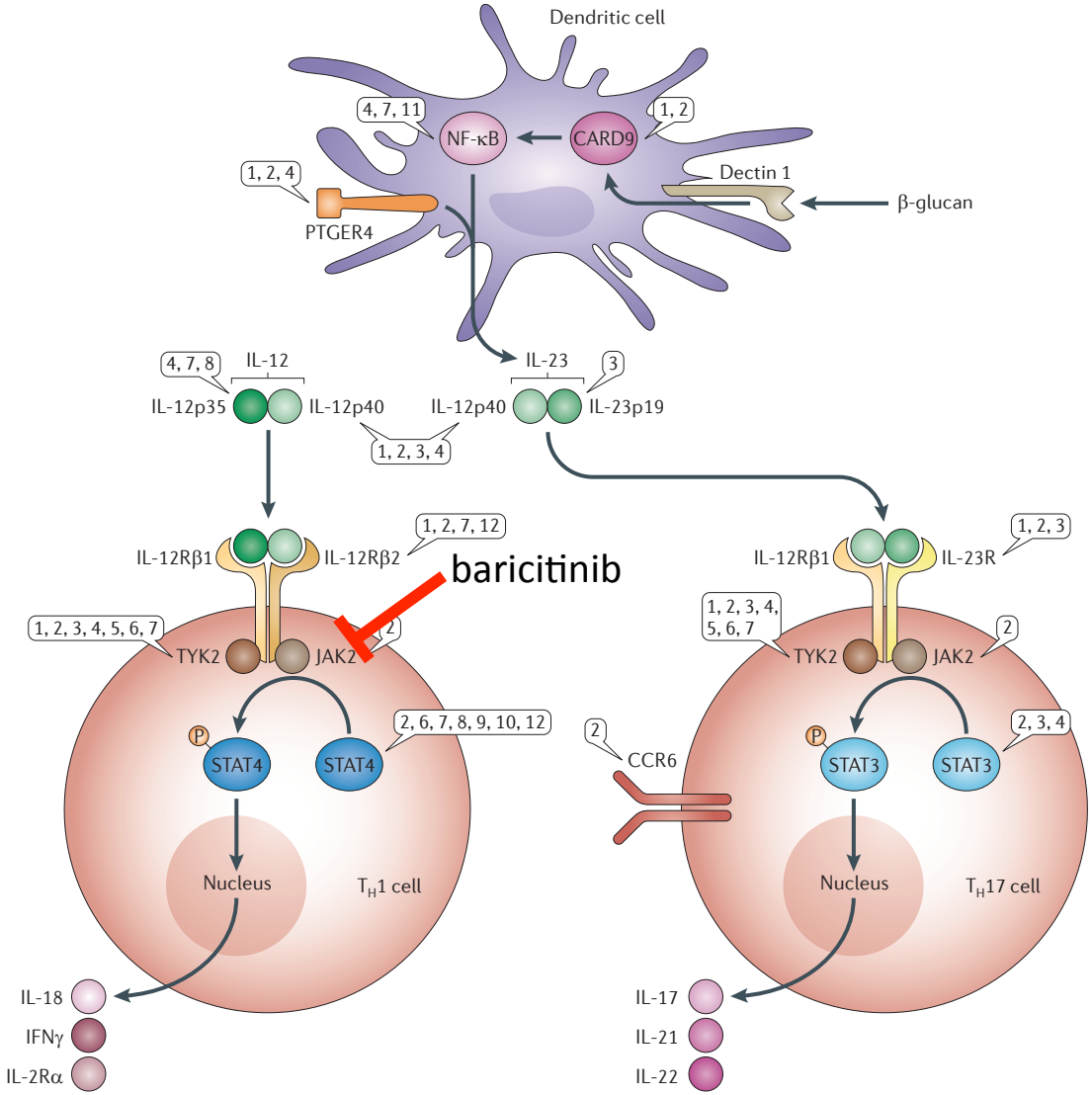



Li 2014

# STAT TFs are activated by Janus Kinase (JAK)



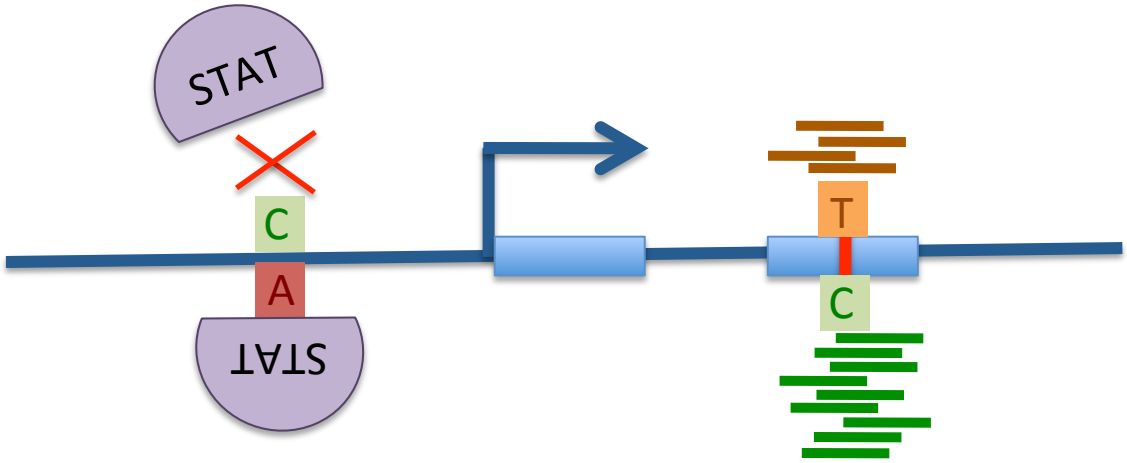

Wikipedia

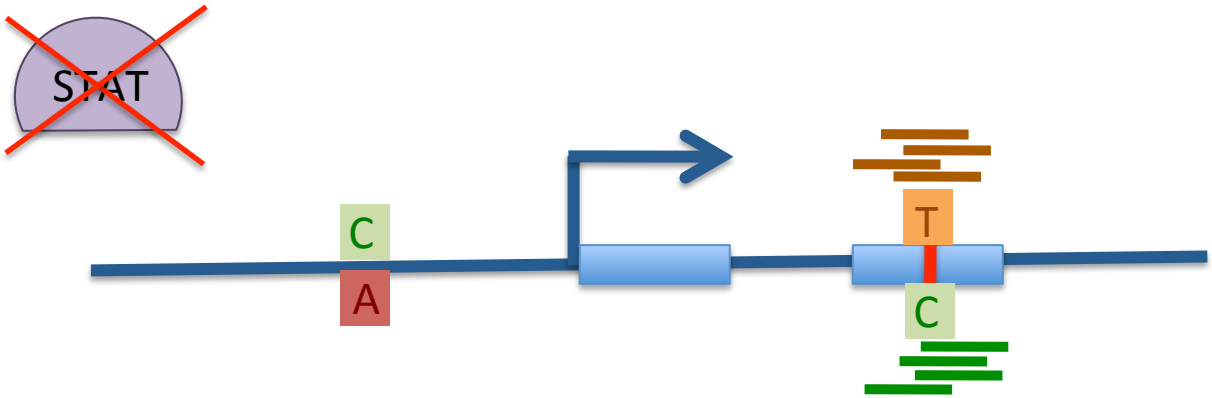

Dendritic cell
4, 7, 11
NF-κB
CARD9
1, 2
1, 2, 4
PTGER4
Dectin 1
β-glucan
4, 7, 8
IL-12
IL-12p35
IL-12p40
IL-23
3
IL-12p40
IL-23p19
1, 2, 3, 4
IL-12Rβ1
IL-12Rβ2
1, 2, 7, 12
baricitinib
1, 2, 3, 4, 5, 6, 7
TYK2
JAK2
2
P
STAT4
STAT4
2, 6, 7, 8, 9, 10, 12
Nucleus
T$_H$1 cell
IL-18
IFNγ
IL-2Rα
IL-12Rβ1
IL-23R
1, 2, 3
1, 2, 3, 4, 5, 6, 7
TYK2
JAK2
2
2
CCR6
P
STAT3
STAT3
2, 3, 4
Nucleus
T$_H$17 cell
IL-17
IL-21
IL-22
1 Ankylosing spondylitis
2 Inflammatory bowel disease
3 Psoriasis
4 Multiple sclerosis
5 Type 1 diabetes
6 Rheumatoid arthritis
7 Primary biliary cirrhosis
8 Coeliac disease
9 Systemic lupus erythematosus
10 Sjögren's syndrome
11 Ulcerative colitis
12 Behçet's disease

Parkes 2013

# Looking for genetic differential regulation mediated by STATs

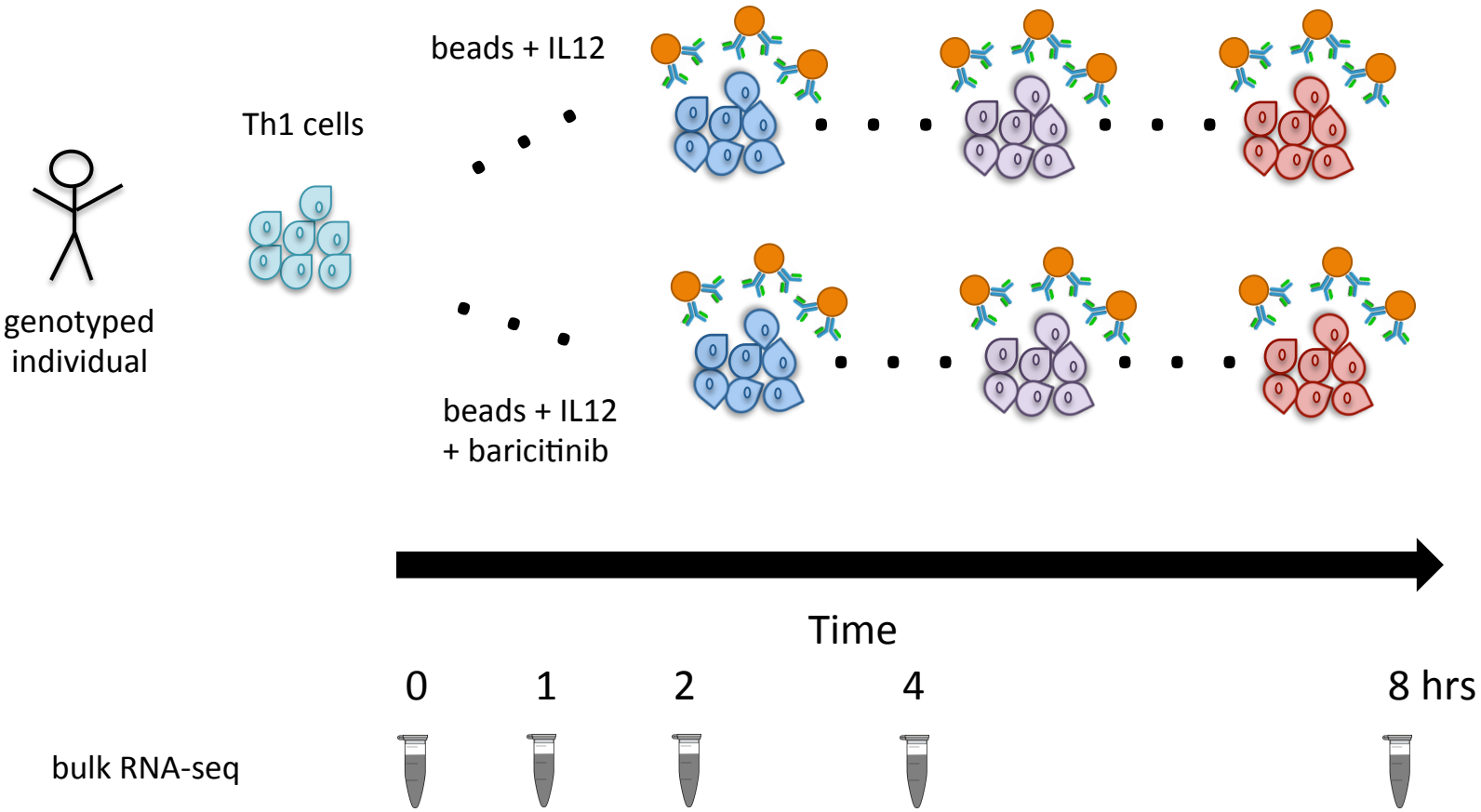
Condition 1




Condition 2

# Pilot experiment design



# Pre-experiments

- RT-PCR

At which time points transcription levels of direct STAT4 target genes peak?

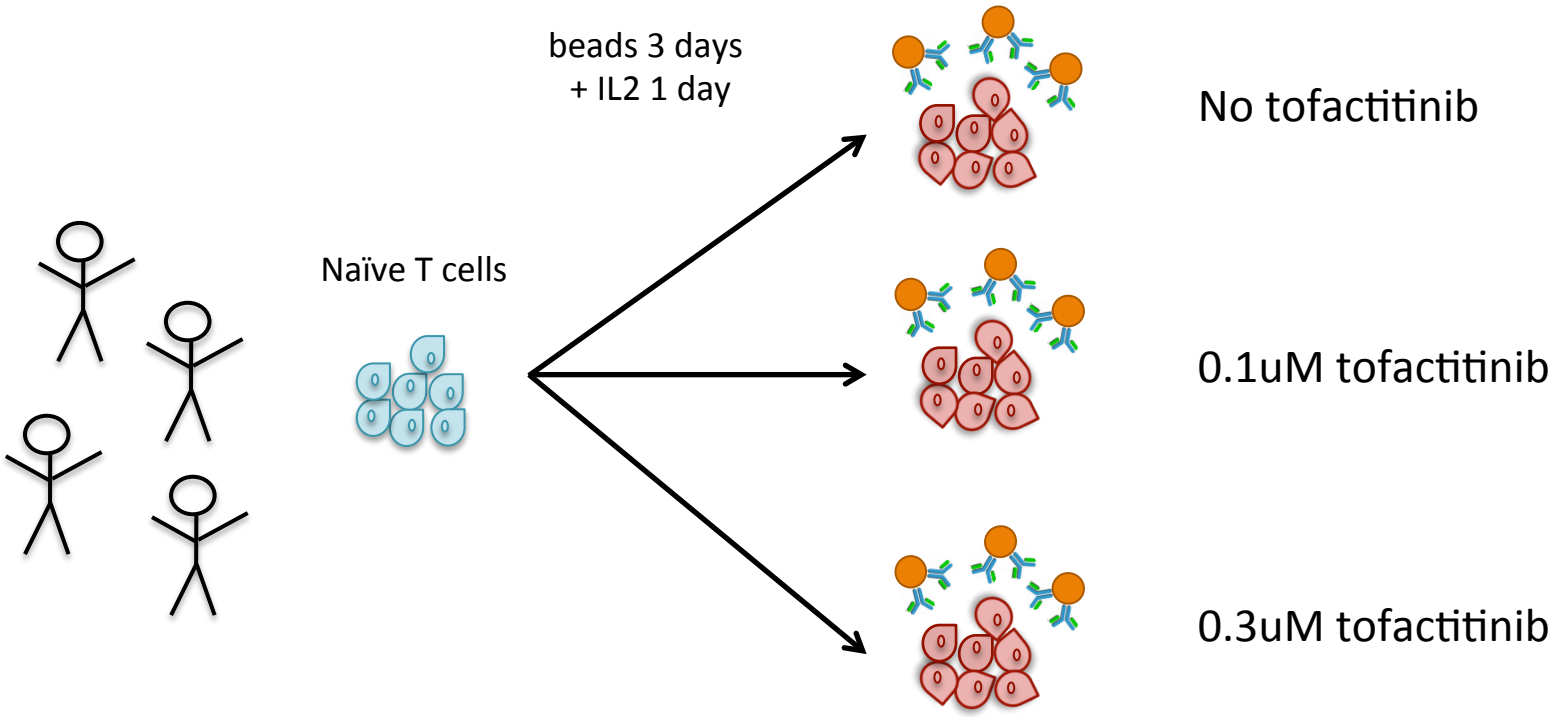
IFNgama
TNF
IL12RB2/Tbet

- Phospho flow

Which STATs get phosphorilated in Th1 upon stimulation
Which STATs are affected by baricitinib?

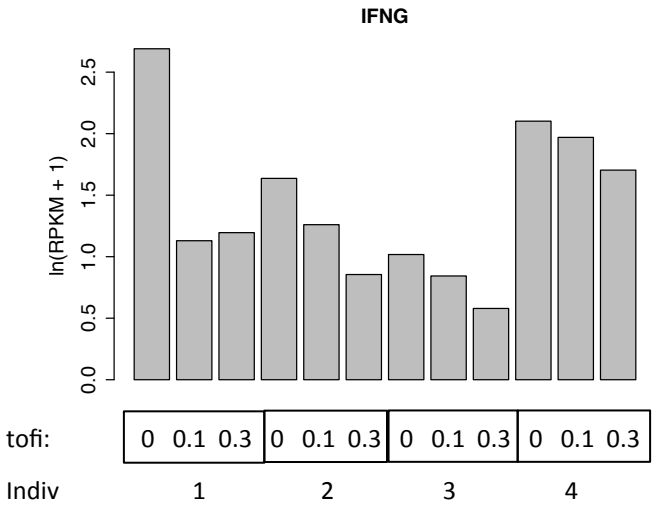
# Analyses on  published dataset using tofi on T cells

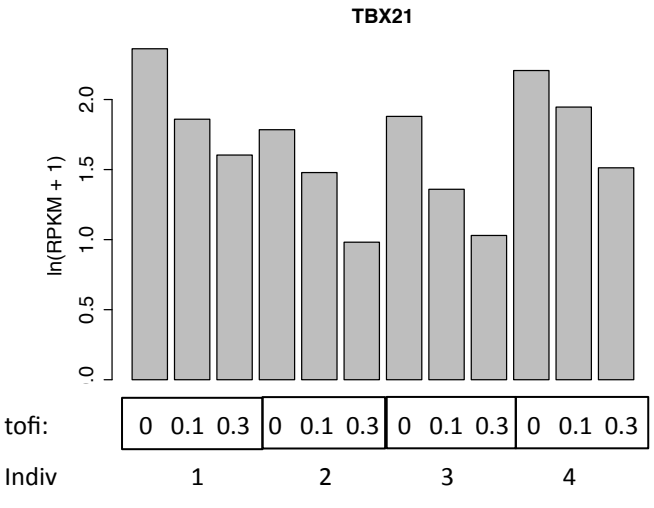
Vahedi *et al* 2015

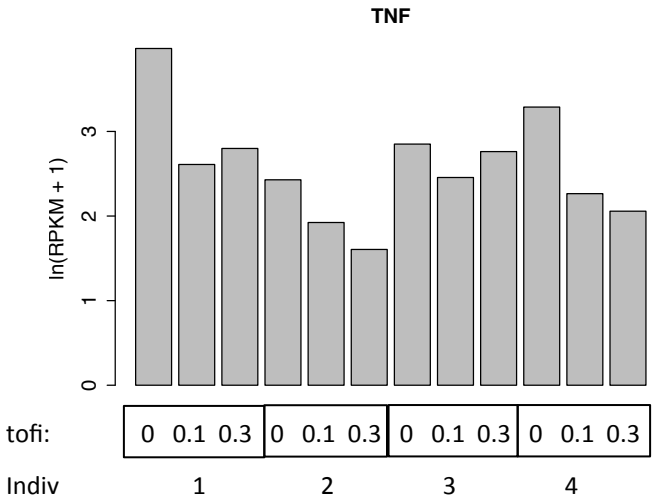


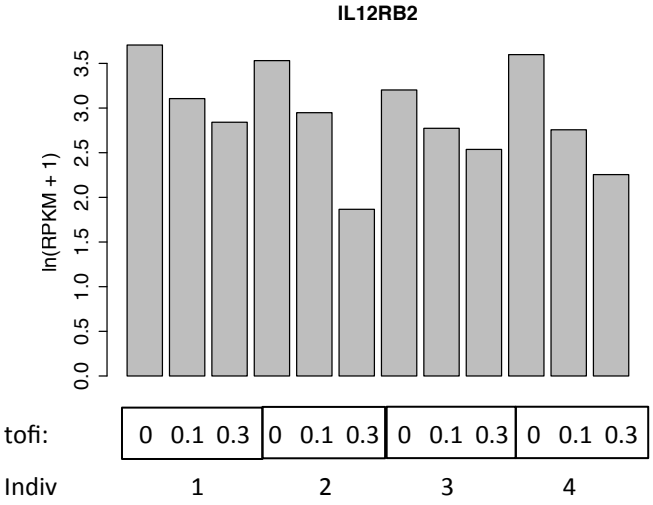

RNA-seq at 3 days, for 3 conditions and 4 indivs

# STAT4 target genes are mostly down-regulated by tofi



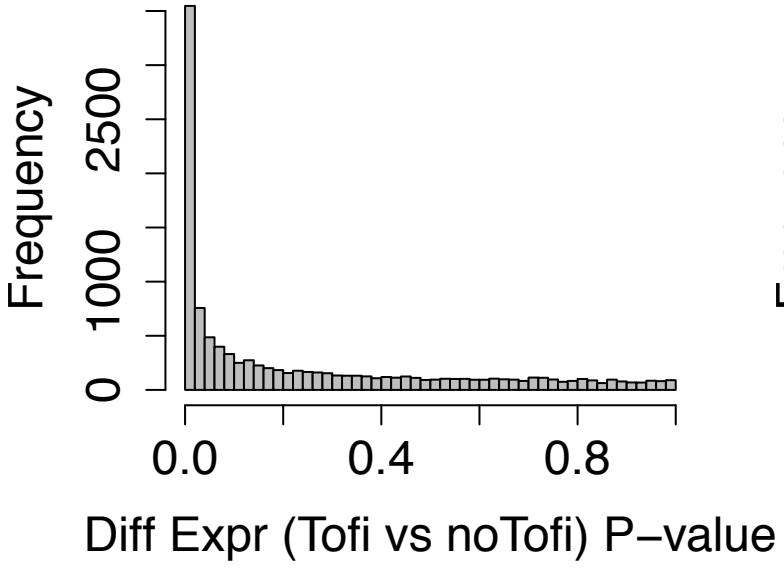



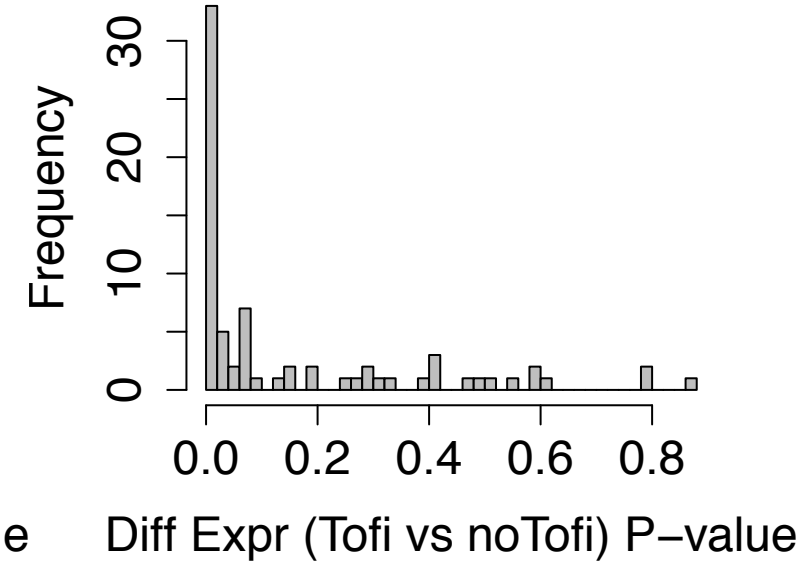
# Tofi regulated genes are enriched for RA genes

P-values for differential expression analysis

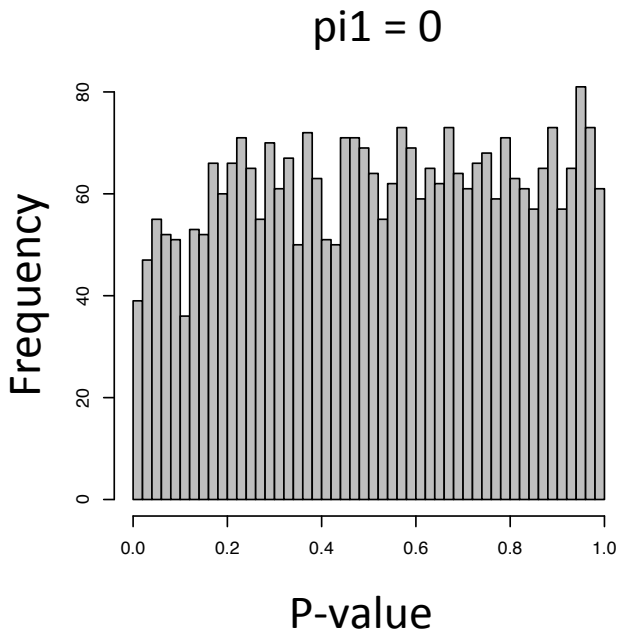
All 11K expressed genes
pi1 = 0.645

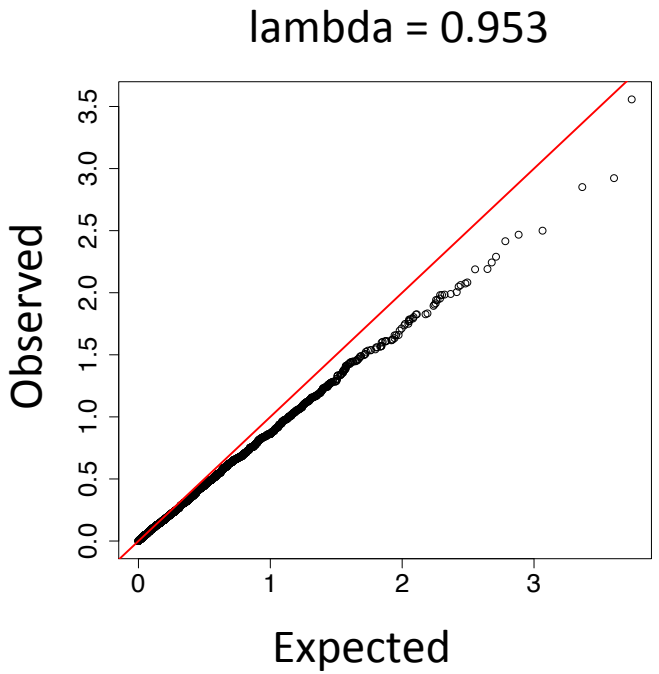



RA genes Okada *et al*
pi1 = 0.966



# No enrichment for significant Tofi-dep ASE events



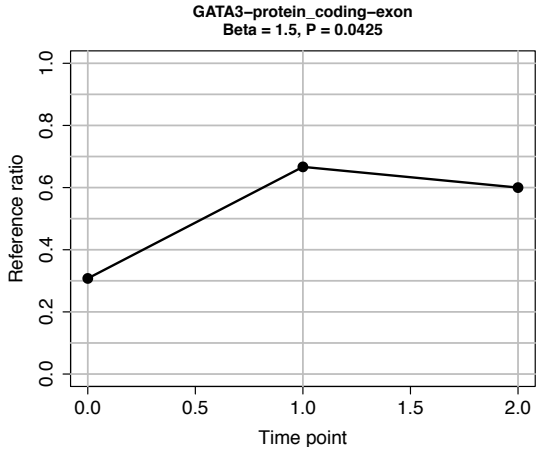

N = 3,090 tested sites

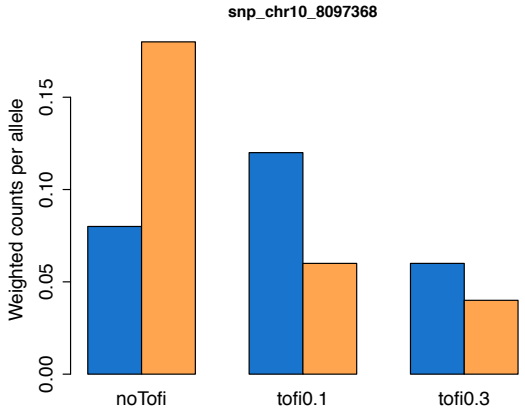
# 3 SNPs in RA genes with P < 0.05

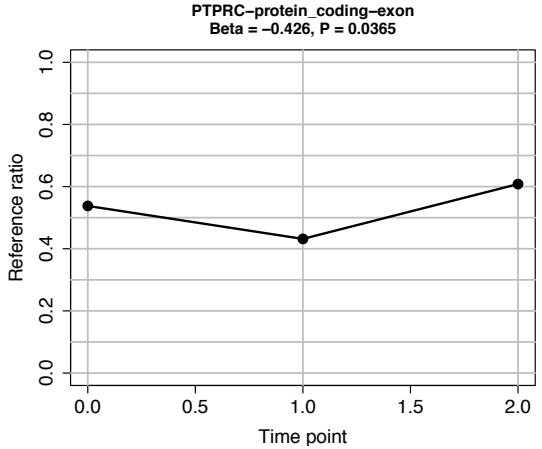
GATA3 TF

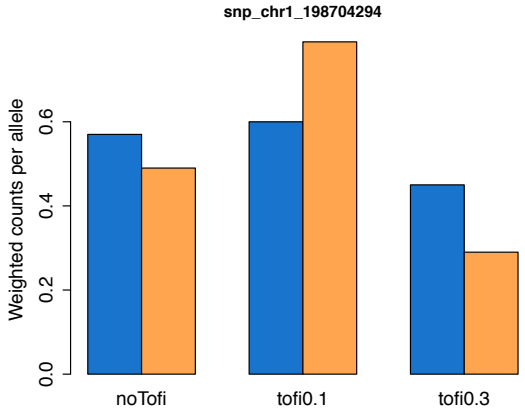
PTPRC
(CD45, regulator of cytokine receptor signaling)

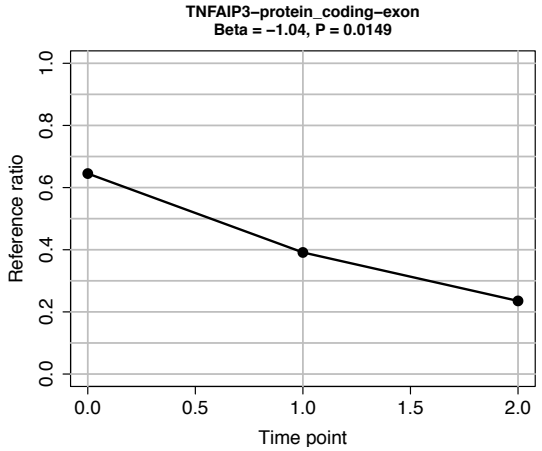
TNF induced protein
(zinc finger, inhibits NFKB activation)

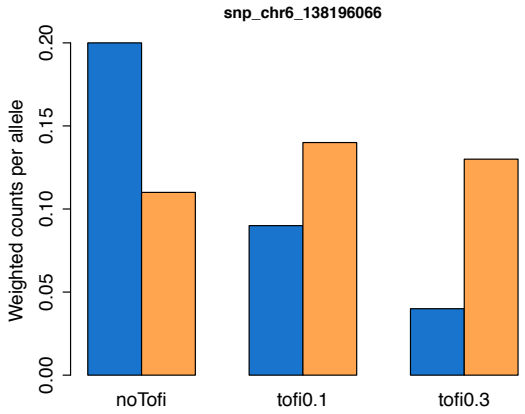


# Conclusions

Condition specific ASE can be detected in a single individual (if sequenced at enough depth and at enough conditions)

Condition specific ASE can be used to detect genetic regulatory effects of disease genes and dissect at which cellular states they are active

Condition specific ASE in DQB1 seems to be real (but needs replication in additional individuals)