



National Institute for  
Health Research

# Annotation of GWAS hits

## Jo Knight

Jo.Knight@genetics.kcl.ac.uk

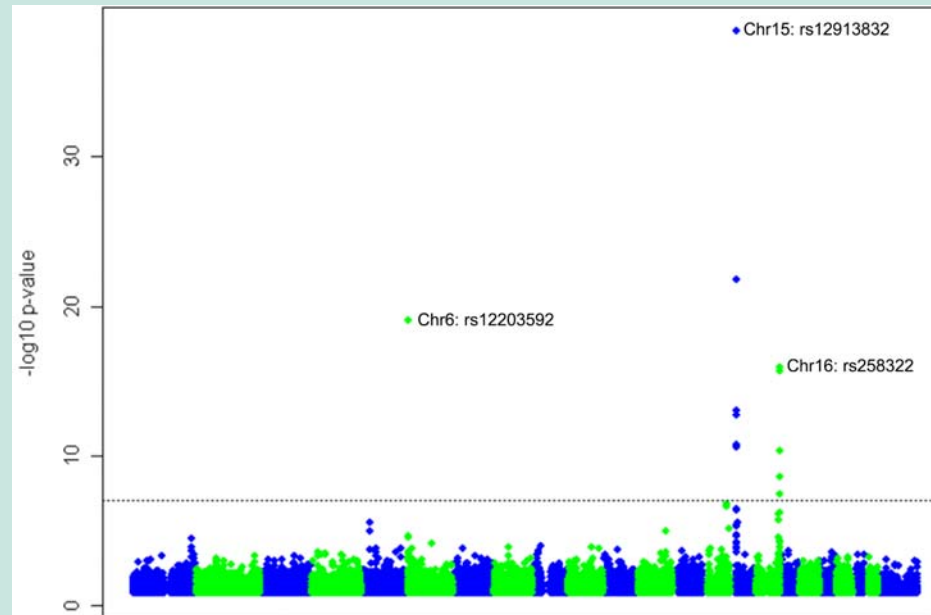


KING'S  
College  
LONDON

Guy's and St Thomas'  
NHS Foundation Trust



# Motivation



- Small number of highly significant results - hundreds of marginally significant results
- Marginal results under investigated
- New methods required to prioritise SNPs for follow up

# Aims

- Current aim
  - To investigate whether GWAS hits are enriched for functionally important annotations relative to random expectation
    - nsSNP
    - eQTL
    - Promoter SNPs
- Future aim
  - Design a weighting scheme using machine learning
  - Data driven weighting scheme

# Background

- Functional characteristics of causal variants
  - Mendelian traits (Cambien et al 1999)
  - Complex traits unknown
    - GWAS data required
    - Now > 1000 hits - enough for quantitative evaluation
- Annotation
  - There is an increasing amount of annotation available which is increasing detailed
    - For review see Karchin (2009)
    - For an example see SNP Nexus (Chelala et al 2008)
- Weighting and prioritisation schemes
  - Currently ad-hoc rather than empirically tested

# Annotations

- nsSNPs
  - SNPs that change the amino acid sequence
  - Polyphen (HGVDbase)
- eQTLs
  - SNPs that are associated with expression
  - GWAS by Dixon et al 2007
    - 55,000 transcripts representing 21,000 genes studied
    - ~15,000 transcripts representing 7,000 genes demonstrated heritable patterns of expression
    - GWAS of these eQTLs was undertaken
    - We selected the top ~ 50,000 SNPs ( $p \sim 10^{-6}$ )
- Regulation
  - SNPs in the promoter region or the first exon
  - First EF track from UCSC genome browser

# Data - Ideal

- All GWAS hits versus all analysed SNPs
  - Duplicate entries included
- Difficulties for hits
  - Panel versus follow up SNP
  - Independence of hits across samples
  - Different SNPs analysed in different sections of the study
- Difficulties for random SNPs
  - Lists of SNPs that passed QC not generally available

# Data - GWAS hits

- Both derived through literature searches
- Panel only SNPs
- NHGRI (Hindorff et al 2009)
  - Ongoing collection
  - Unique SNP N = 1172
  - Study N = 266 (pre Mar 08 = 98)
  - P-values  $<10^{-5}$
- Johnson et al (2009)
  - Data freeze at 1 March 2008
  - We used  $p < 10^{-5}$  for equivalence with NHGRI
  - Unique SNP N = 4086
  - Study N = 107
- Overlap of 85 studies but only 350 SNPs

# Data – “Random” Panels

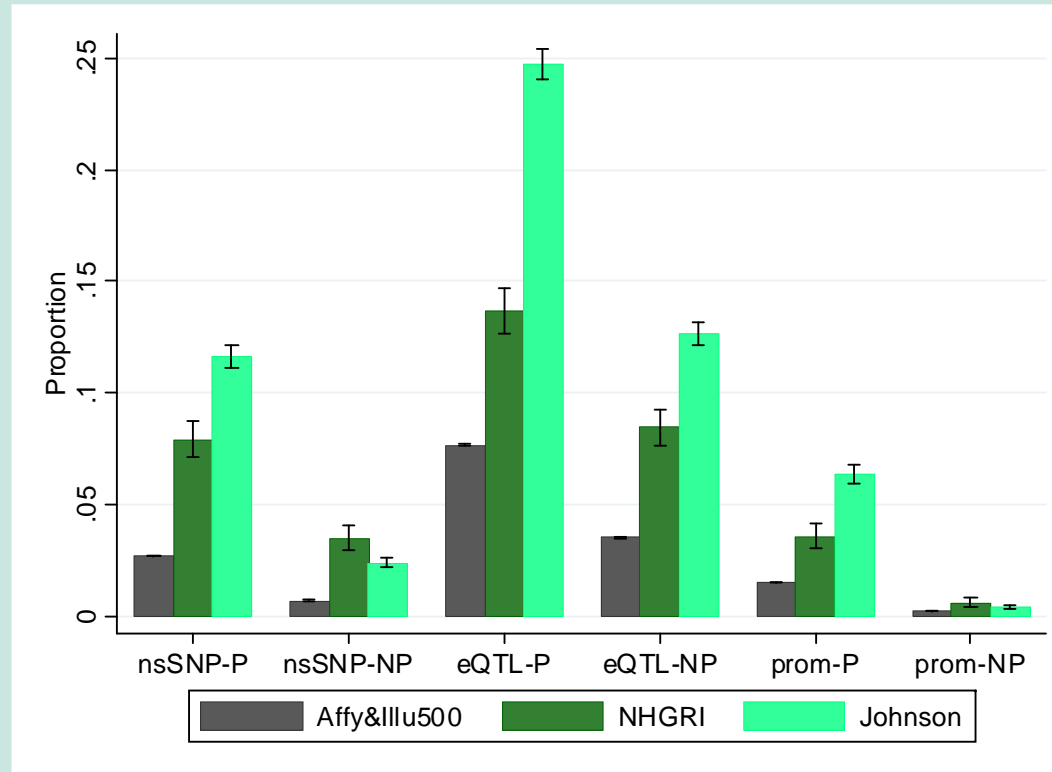
- Robustness check
  - Differences between the way that chips are designed means the list may have different levels of annotation
    - Illu 550
    - Affy 500
    - Affy 500 and Illu 550
    - Affy 6.0 and Illu 1M

# Analysis

- Compare the proportions of annotated SNPs in GWAS vs “random” SNPs
- Use LD proxies of annotated SNPs to improve coverage and account for indirect association
- Test different lists as a sensitivity analysis
- Stratified analysis to investigate different sources of possible bias
  - Chromosomal location of SNPs (esp MHC region)
  - MAF
- Overlap of annotation (in GWAS hits with > 1 annotation)

# Annotation proportion

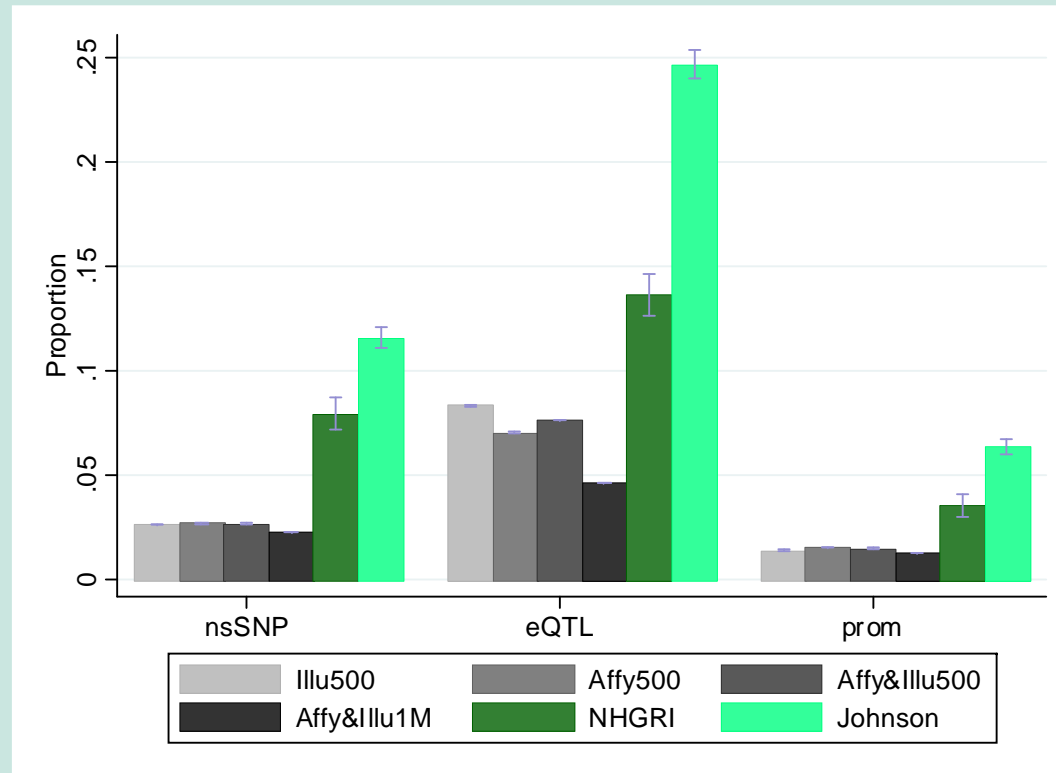
## - with and without LD proxies



- Annotation proportion generally higher in GWAS hits

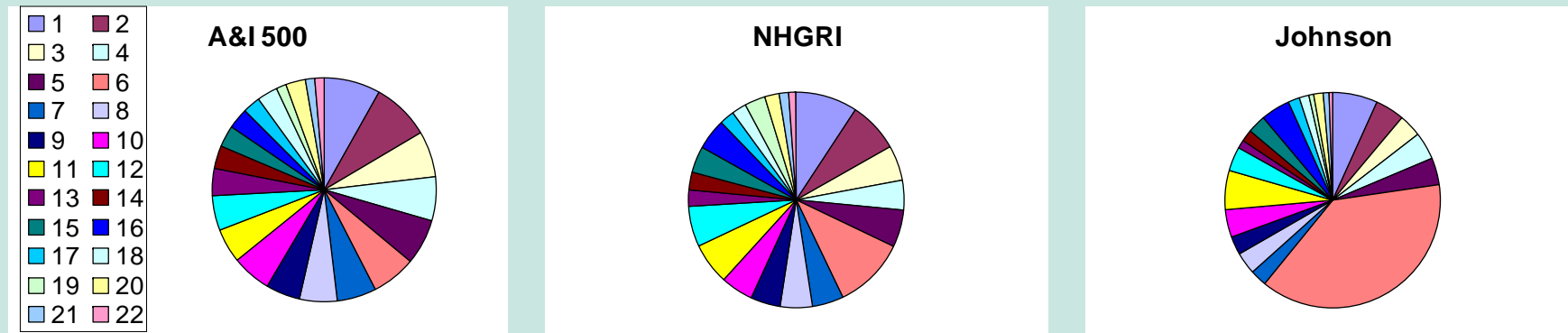
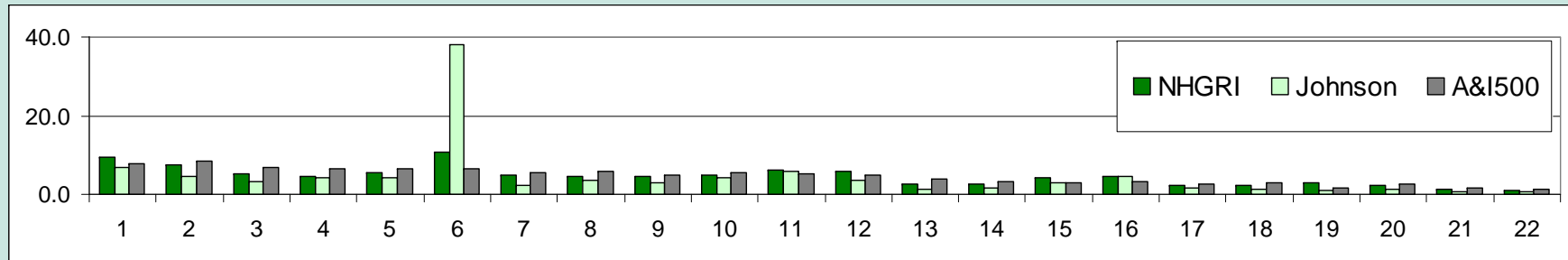
# Annotation proportion

## - with alternate random panels



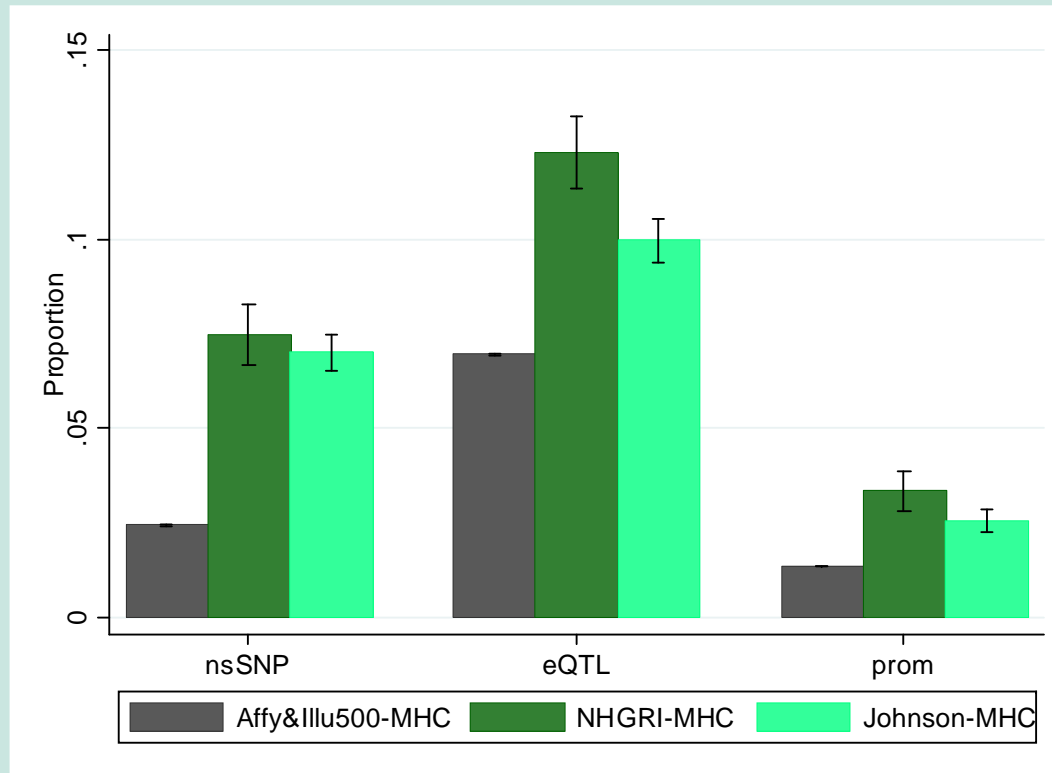
- Alternate random SNP panels show similar effects

# SNP Chromosome Distribution



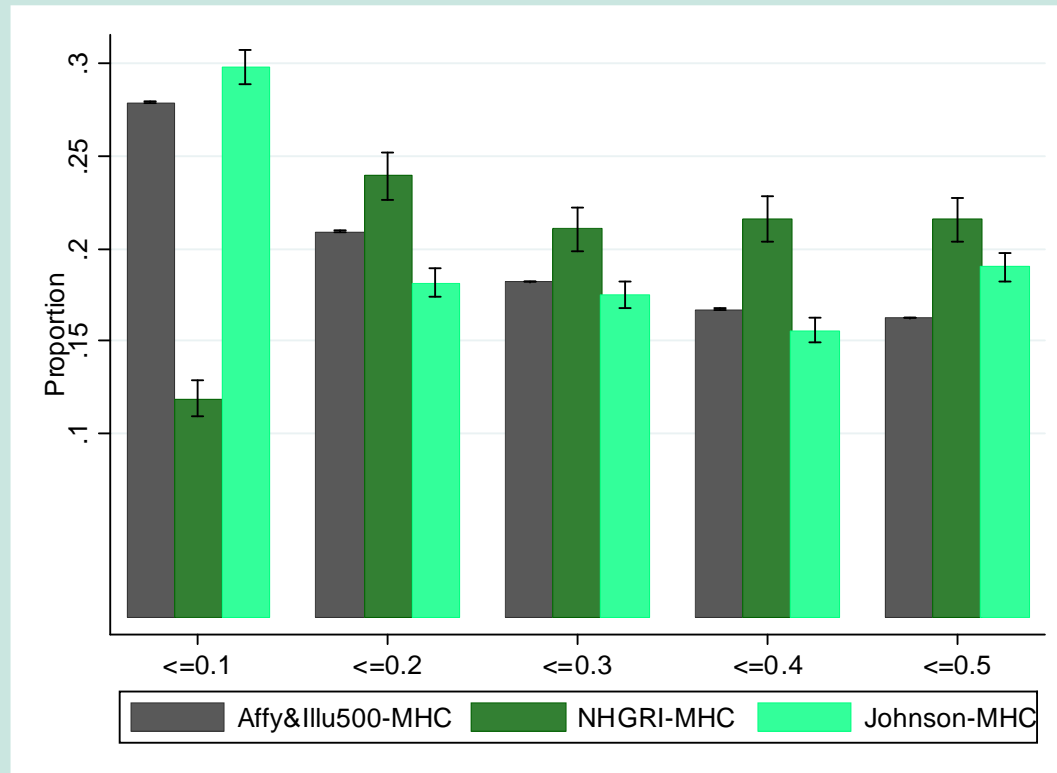
- Over representation of SNPs on chromosome 6 in the Johnson panel

# Annotation proportion - without MHC



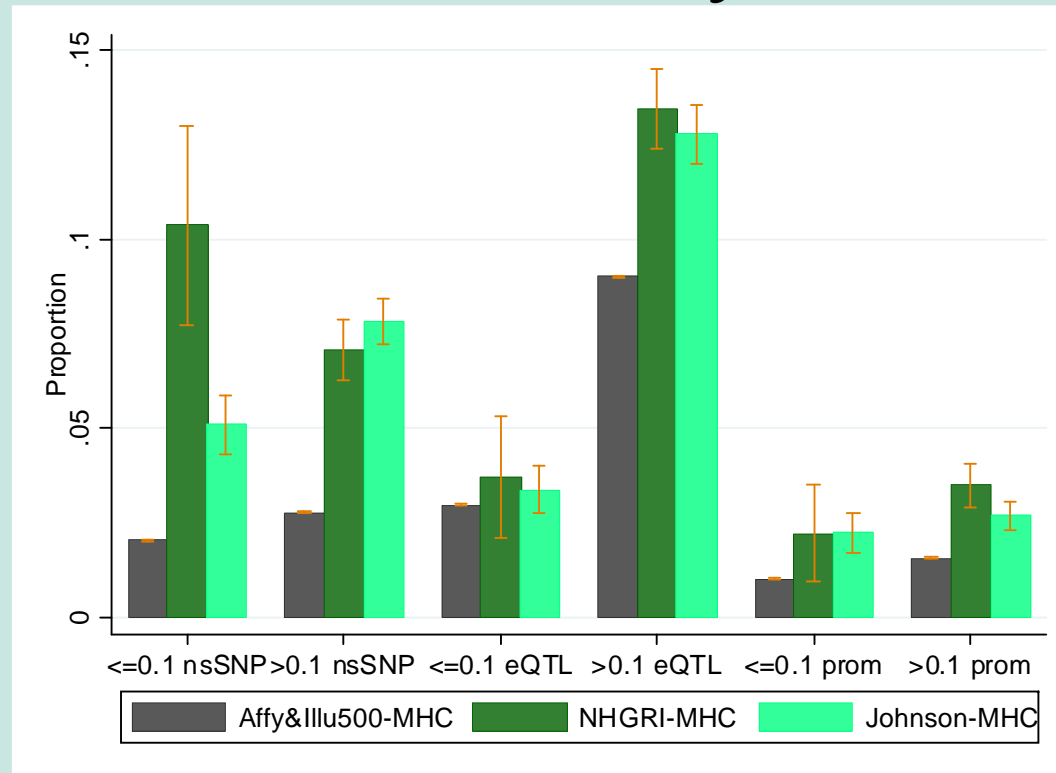
- Patterns similar without MHC region SNPs

# SNP MAF Distribution



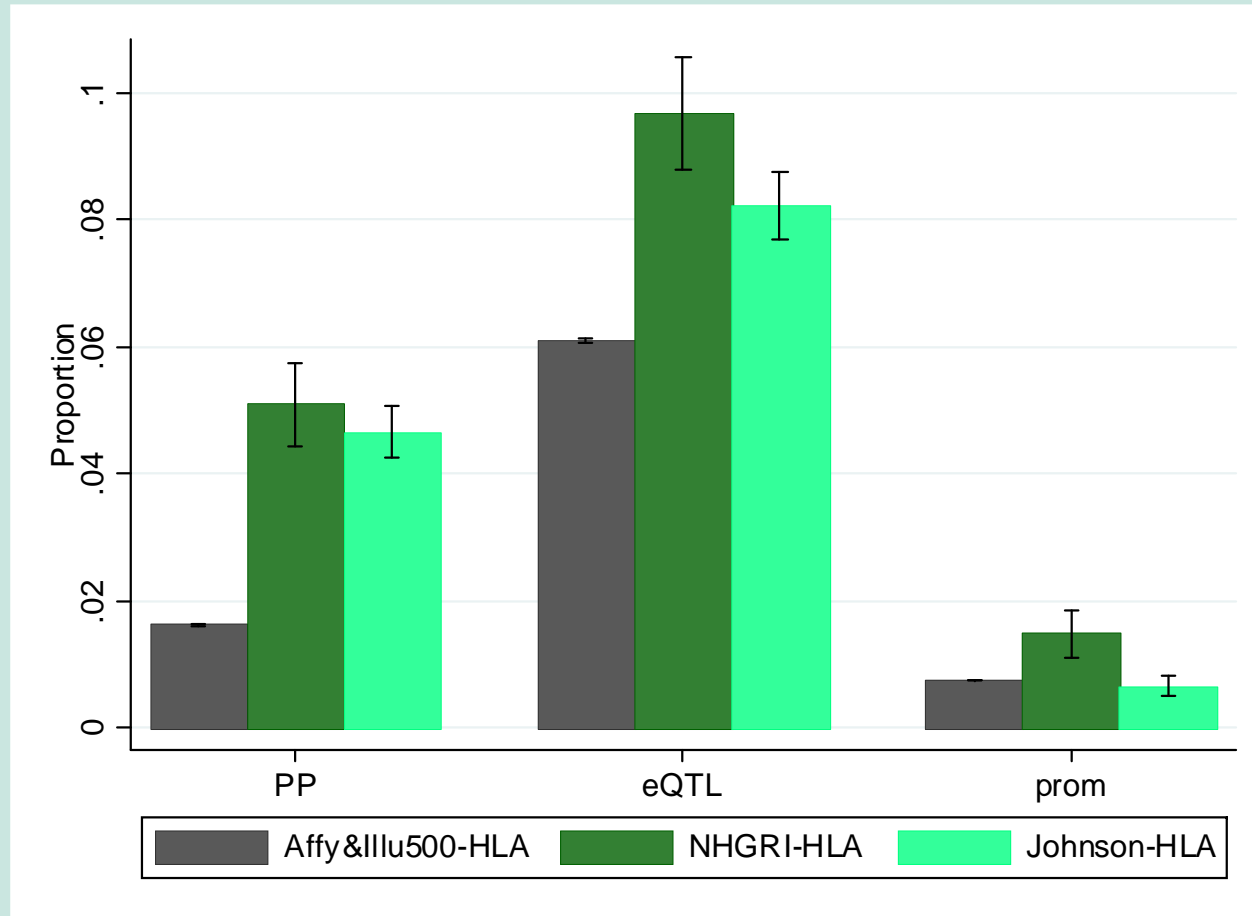
- MAF distribution similar in Johnson and random SNPs but different in NHGRI hits

# Annotation proportion - stratified by MAF



- Patterns remain in most MAF categories

# Annotation proportion - without annotation overlap



# Conclusions

- There is a difference in the proportion of annotated SNPs between GWAS and random SNPs for nsSNPs and eQTLs
- Promoter SNPs show a trend in this direction (Small N)
- Robust to
  - Different random SNP lists
  - Different GWAS hit lists
  - Chromosomal distribution
  - MAF
  - None overlapping annotation
- Possible weighting – ratio of annotation, GWAS/Random
  - nsSNPs 3
  - eQTLs 1.5
  - Promoter 1.4

# Acknowledgements

- Mike Weale, KCL
- Mike Barnes, GSK
- Gerome Breen, KCL