Overlapping genes and pathways identified by GWAS and EWAS

Thomas Battram

Brief methods

GWAS and EWAS summary data for seven traits were extracted. The hits from these studies were mapped to genes and pathways (GO and KEGG). Analyses were undertaken to determine whether there was more overlap of genes/pathways than expected by chance using Fisher's exact test and 1000 permutations. Odds ratio were extracted, e.g. for gene overlap OR = odds of EWAS gene being a GWAS gene / odds of EWAS gene not being a GWAS gene.

Then, to determine what scenario would give the empirical results, simulations were conducted. Parameters that varied in the simulations: The proportion of the total number of genes that are causal (prop_causal_genes) The proportion of the total number of genes that are consequential (prop_consequent_genes) The proportion of EWAS genes that cause changes in the trait (percent_ewas_causal)

Parameters determined by empirical results: Number of EWAS genes (n_ewas_genes) Number of GWAS genes (n_gwas_genes)

Results

Empirical analyses

Little evidence of more overlap than expected by chance

Empirical compared to simulations

There was little evidence prop_consequent_genes influenced the results so these were removed from the plots to make it clearer.

Just included plots for alcohol consumption per day because they're fairly representative of the other traits and have enough genes to actually look at gene overlap.

For each figure, the simulation results (1000 permutations) are plotted and the empirical OR is represented by the black horizontal line.

Thoughts from these results

Gene overlap results suggest that actually the original EWAS may be identifying causal genes. Taken together with the result that the overlap between the GWAS and EWAS genes is not more than expected by chance, it suggests causal genes are being identified, but 1) This is expected to happen by chance and 2) the causal genes being identified are different to those in GWAS.

The KEGG results suggest the simulations aren't actually representative of what is happening empirically.

The GO results suggest something similar to the KEGG results, but the empirical OR is lower than the simulations… Overall this suggests that not only are the simulations not appropriate, but also that even though the EWAS are detecting causal genes, it's not necessarily detecting causal pathways…

Very confused by the contrast in these results!!! Any help would be great!

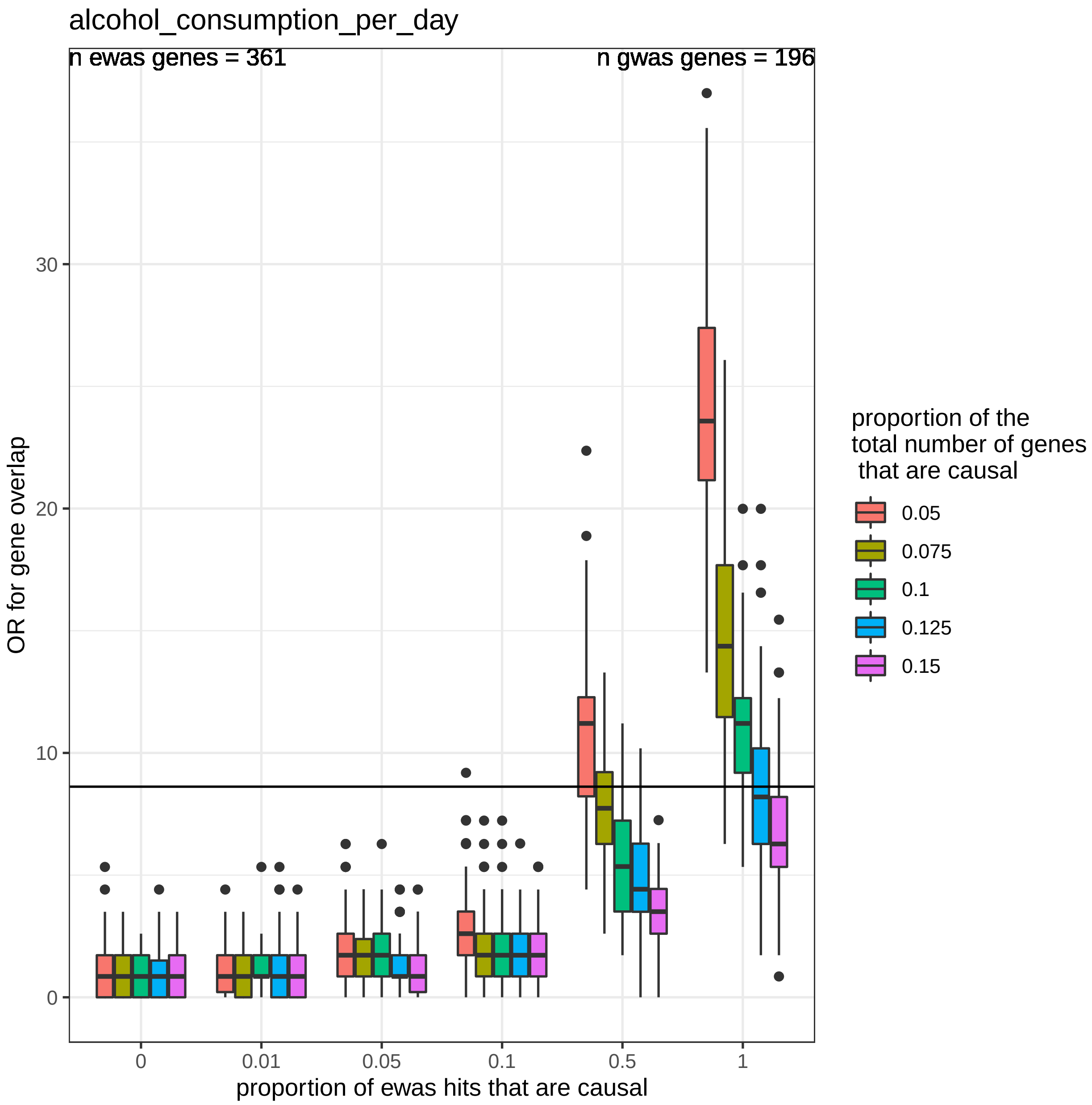

Figure 1: Gene overlap

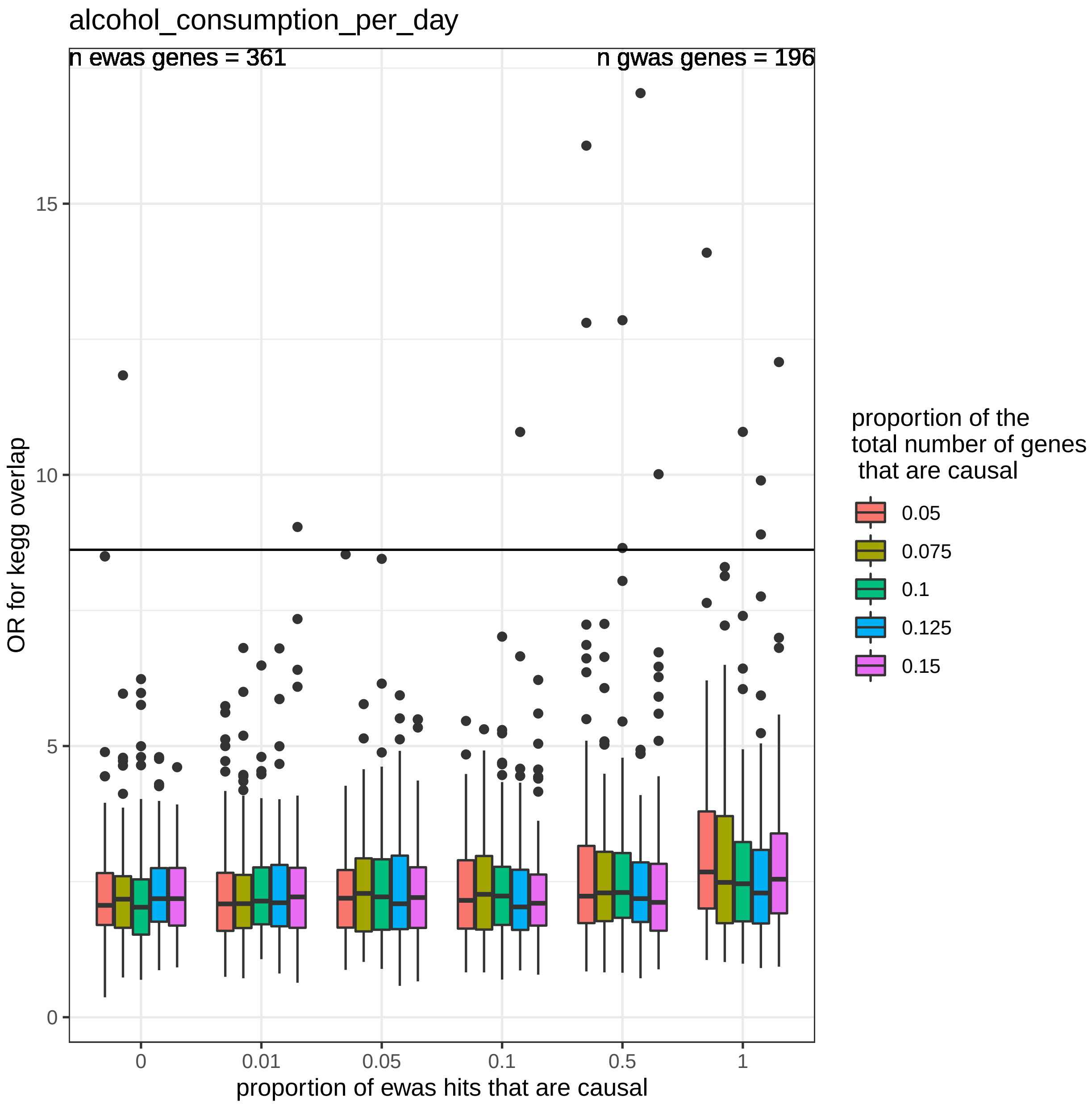

Figure 2: KEGG overlap

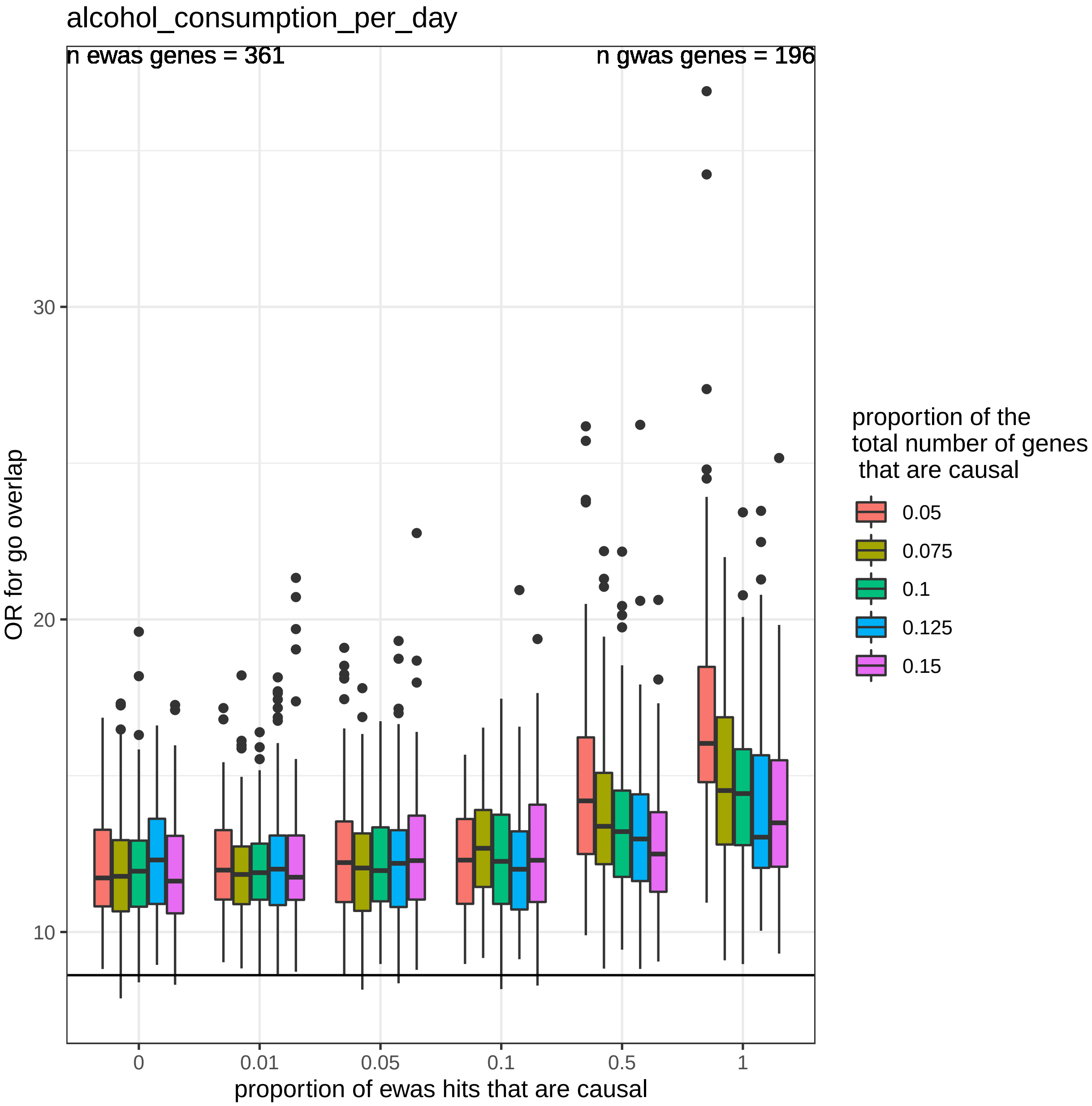

Figure 3: GO overlap

simulation functions for interest

```
perform_enrichment <- function(ewas_vars, gwas_vars, total_vars)
{
  q <- sum(ewas_vars %in% gwas_vars) # overlap between ewas and gwas pathways
  m <- length(gwas_vars) - q # number of pathways identified by GWAS but not EWAS
  k <- length(ewas_vars) - q # number of pathways identified by EWAS but not by GWAS
  n <- length(total_vars) - q - m - k # all pathways minus pathways identified by EWAS and GWAS
  tab <- matrix(c(q, m, k, n), 2, 2)
  out <- fisher.test(tab, alternative = "greater")
  return(out)
}
sim3 <- function(prop_causal_genes, n_ewas_genes, n_gwas_genes, percent_ewas_causal, prop_consequent_gen
{
    # 1. sample causal and consequential genes from total number of genes
    # 2. sample gwas and ewas genes and do gene overlap tests
    # 3. link these to pathways and do pathway overlap tests
    # 4. write out results

    # 1.
    ca_genes <- sample_n(all_genes, prop_causal_genes * nrow(all_genes))
    con_genes <- sample_n(all_genes, prop_consequent_genes * nrow(all_genes))

    # 2.
    gwasg <- sample(ca_genes$ensembl_gene_id, n_gwas_genes)
    ewas_cag <- sample(ca_genes$ensembl_gene_id, n_ewas_genes * percent_ewas_causal)
    ewasg <- unique(c(ewas_cag, sample(con_genes$ensembl_gene_id, n_ewas_genes - length(ewas_cag))))

    outg <- perform_enrichment(ewasg, gwasg, all_genes$ensembl_gene_id)

    # 3.
    database <- c("go", "kegg")
    outp <- lapply(1:2, function(x) {
        pathway_dat <- get(paste0(database[x], "_terms"))
        gwasp <- pathway_dat %>%
            dplyr::filter(ensembl_gene_id %in% gwasg) %>%
            pull(pathway_id) %>%
            unique
        ewasp <- pathway_dat %>%
            dplyr::filter(ensembl_gene_id %in% ewasg) %>%
            pull(pathway_id) %>%
            unique
        out_res <- perform_enrichment(gwasp, ewasp, unique(pathway_dat$pathway_id))
        overlap <- sum(ewasp %in% gwasp)
        return(list(res = out_res, overlap = overlap))
    })
    names(outp) <- database

    # 4.
    out <- list(gene = outg,
                gene_overlap = sum(ewasg %in% gwasg),
                go = outp$go$res,
                go_overlap = outp$go$overlap,
                kegg = outp$kegg$res,
```

```r
                kegg_overlap = outp$kegg$overlap)
    return(out)
}
```