Using the NHLBI GRASP repository of GWAS test results with Bioconductor

Vincent J. Carey

October 9, 2014

Contents

1 Introduction **1**

2 Installation **2**

3 Demonstration **2**

3.1 Attachment and messaging . . . 2

3.2 Indexed p-value bins . . . 2

3.3 Tabulations . . . 3

3.4 Inspecting some relatively weak associations in asthma . . . 3

4 Quick view of the basic interfaces **4**

4.1 dplyr-oriented . . . 4

4.2 RSQLite-oriented . . . 4

5 Some QC (Consistency check): Are NHGRI GWAS catalog loci included? **5**

1 Introduction

GRASP (Genome-Wide Repository of Associations Between SNPs and Phenotypes) v2.0 was released in September 2014. The primary GRASP web resource includes links to a web-based query interface. This document describes a Bioconductor package that replicates information in the v2.0 textual release.

The primary reference for version 2 is: Eicher JD, Landowski C, Stackhouse B, Sloan A, Chen W, Jensen N, Lien J-P, Leslie R, Johnson AD (2014) GRASP v 2.0: an update to the genome-wide repository of associations between SNPs and phenotypes. Nucl Acids Res, published online Nov 26, 2014 PMID 25428361

From the main web page:

GRASP includes all available genetic association results from papers, their supplements and web-based content meeting the following guidelines:

- All associations with P<0.05 from GWAS defined as >= 25,000 markers tested for 1 or more traits.
- Study exclusion criteria: CNV-only studies, replication/follow-up studies testing <25K markers, non-human only studies, article not in English, gene-environment or gene-gene GWAS where single SNP main effects are not given, linkage only studies, aCGH/LOH only studies, heterozygosity/homozygosity (genome-wide or long run) studies, studies only presenting gene-based or pathway-based results, simulation-only studies, studies which we judge as redundant with prior studies since they do not provide significant inclusion of new samples or exposure of new results (e.g., many methodological papers on the WTCCC and FHS GWAS).
- More detailed methods and resources used in constructing the catalog are described at Methods & Resources.
- Terms of Use for GRASP data: http://apps.nhlbi.nih.gov/Grasp/Terms.aspx
- Medical disclaimer: [http://apps.nhlbi.nih.gov/Grasp/Overview.aspx] (http://apps.nhlbi.nih.gov/Grasp/Overview.aspx)

2 Installation

Install the package in Bioconductor version 3.1 or later using `BiocInstaller::biocLite("grasp2db")`.

The first time the grasp2 data base is referenced, via the `GRASP2()` function described below, a large (5.3Gb) file is downloaded to a local cache using *AnnotationHub*. This may take a considerable length of time (10's of minutes, perhaps an hour or more depending on internet connections). Subsequent uses refer to the locally cached file, and are fast.

3 Demonstration

3.1 Attachment and messaging

Attach the package.

```
library(grasp2db)
```

Workflows start with a reference to the data base.

```
grasp2 <- GRASP2()
```

```
## snapshotDate(): 2015-05-22
```

```
grasp2
```

```
## src:  sqlite 3.8.6 [AnnotationHub()[["AH21414"]]]
## tbls: count, study, variant
```

There are three tables. The 'variant' table summarizes 8864671 variants. The 'study' table contains 2044 citations from which the data are extracted; the 'variant' and 'study' tables are related by PMID identifier. The 'count' table contains 19793480 records summarizing SNPs observed in Discovery and Replication samples from 12 distinct Populations; theh 'variant' and 'count' tables are related by NHLBIkey.

3.2 Indexed p-value bins

The database has been indexed on a number of fields, and an integer rounding of $-\log_{10}$ of the quantity recorded as the Pvalue of the association is available.

```
variant <- tbl(grasp2, "variant")
q1 = (variant %>% select(Pvalue, NegativeLog10PBin) %>%
       filter(NegativeLog10PBin > 8) %>%
       summarize(maxp = max(Pvalue), n=n()))
q1
```

```
## Source: sqlite 3.8.6 [AnnotationHub()[["AH21414"]]]
## From: <derived table> [?? x 2]
##
##           maxp      n
## 1  3.16181e-09 322794
## ..         ...    ...
```

This shows that the quantities in NegativeLog10PBin are upper bounds on the exponents of the p-values in the integer-labeled bins defined by this quantity.

A useful utility from dplyr is the query explain method:

```
explain(q1)
```

```
## <SQL>
## SELECT "maxp", "n"
## FROM (SELECT MAX("Pvalue") AS "maxp", COUNT() AS "n"
## FROM "variant"
## WHERE "NegativeLog10PBin" > 8.0) AS "_W1"
##
##
## <PLAN>
##   selectid order from
## 1        0     0    0
##                                                                              detail
## 1 SEARCH TABLE variant USING INDEX NegativeLog10PBin_index (NegativeLog10PBin>?)
```

3.3 Tabulations

This query select variants with large effect from the 'variant' table, and joins them to their published phenotypic effect in the 'study' table.

```
study <- tbl(grasp2, "study")
large_effect <-
    variant %>% select(PMID, SNPid_dbSNP134, NegativeLog10PBin) %>%
        filter(NegativeLog10PBin > 5)
phenotype <-
    left_join(large_effect,
              study %>% select(PMID, PaperPhenotypeDescription))
```

```
## Joining by: "PMID"
```

```
phenotype
```

```
## Source: sqlite 3.8.6 [AnnotationHub()[["AH21414"]]]
## From: <derived table> [?? x 4]
##
##        PMID SNPid_dbSNP134 NegativeLog10PBin    PaperPhenotypeDescription
## 1  20502693            253                 6 Gene expression in monocytes
## 2  19913121            255                 6     Lipid level measurements
## 3  19913121            256                 6     Lipid level measurements
## 4  19913121            263                 6     Lipid level measurements
## 5  19913121            264                 6     Lipid level measurements
## 6  19913121            271                 6     Lipid level measurements
## 7  19913121            285                 6     Lipid level measurements
## 8  19913121            301                 6     Lipid level measurements
## 9  20502693            326                 6 Gene expression in monocytes
## 10 20502693            327                 6 Gene expression in monocytes
## ..      ...            ...               ...                          ...
```

3.4 Inspecting some relatively weak associations in asthma

```
lkaw <- semi_join(
    variant %>%
        filter(NegativeLog10PBin <= 4) %>%
        select(PMID, chr_hg19, SNPid_dbSNP134, PolyPhen2),
    study %>% filter(PaperPhenotypeDescription == "Asthma"))
```

```
## Joining by: "PMID"
```

We materialize the filtered table into a data.frame and check how many PolyPhen2 notations including a substring of ‘Damaging’:

```
lkaw %>% filter(PolyPhen2 %like% "%amaging%")
```

```
## Source: sqlite 3.8.6 [AnnotationHub()[["AH21414"]]]
## From: <derived table> [?? x 4]
## Filter: PolyPhen2 %like% "%amaging%"
##
##        PMID chr_hg19 SNPid_dbSNP134
## 1  20860503        1           4762
## 2  20860503        1         880633
## 3  20860503        1         880633
## 4  20860503        3        1053338
## 5  20860503       19        1054940
## 6  20860503        1        1060622
## 7  20860503        1        1156281
## 8  20860503       13        1536207
## 9  20860503       10        1799853
## 10 20860503       11        2186797
## ..      ...      ...            ...
## Variables not shown: PolyPhen2 (chr)
```

4 Quick view of the basic interfaces

4.1 dplyr-oriented

```
grasp2
```

```
## src:  sqlite 3.8.6 [AnnotationHub()[["AH21414"]]]
## tbls: count, study, variant
```

4.2 RSQLite-oriented

```
gcon = grasp2$con
library(RSQLite)
gcon
```

```
## <SQLiteConnection>
```

```
dbListTables(gcon)
```

```
## [1] "count"   "study"   "variant"
```

Note that the package opens the SQLite data base in ‘read-only’ mode, but updates are possible (e.g., directly opening a connection to the data base without restricting access). There may be implicit control if the user does not have write access to the file.

5 Some QC (Consistency check): Are NHGRI GWAS catalog loci included?

We have an image of the NHGRI GWAS catalog inheriting from GRanges.

```
library(gwascat)
data(gwrngs19)  # hg19 addresses; NHGRI ships hg38
gwrngs19
```

```
## gwasloc instance with 17254 records and 35 attributes per record.
## Extracted:  Mon Sep  8 13:08:13 2014
## Genome:  hg19
## Excerpt:
## GRanges object with 5 ranges and 3 metadata columns:
##     seqnames                 ranges strand |        Disease.Trait
##        <Rle>              <IRanges>  <Rle> |          <character>
##   1    chr19 [  7739177,   7739177]      * |     Resistin levels
##   2     chr6 [ 32626601,  32626601]      * | Asthma and hay fever
##   3     chr4 [ 38799710,  38799710]      * | Asthma and hay fever
##   4     chr5 [110467499, 110467499]      * | Asthma and hay fever
##   5     chr2 [102966549, 102966549]      * | Asthma and hay fever
##            SNPs   p.Value
##     <character> <numeric>
##   1   rs1423096     1e-07
##   2   rs9273373     4e-14
##   3   rs4833095     5e-12
##   4   rs1438673     3e-11
##   5  rs10197862     4e-11
##   -------
##   seqinfo: 23 sequences from hg19 genome
```

We would like to verify that the majority of the variants enumerated in the NHGRI catalog are also present in GRASP 2.0. We supply a function called checkAnti which obtains the anti-join between a chromosome-specific slice of the NHGRI catalogue and the slice of GRASP2 for the same chromosome. We compute for chromosome 22 the fraction of NHGRI records present in GRASP2.

```
gr22 = variant %>% filter(chr_hg19 == "22")
abs22 = checkAnti("22")
1 - (abs22 %>% nrow())/(gr22 %>% nrow())
```

```
## [1] 0.9982036
```

The absent records can be seen to be relatively recent additions to the NHGRI catalog.

```
abs22
```

```
## Source: local data frame [217 x 42]
##
##    seqnames    start      end width strand Date.Added.to.Catalog PUBMEDID
## 1        22 37545505 37545505     1      *            08/05/2014 24390342
## 2        22 39747671 39747671     1      *            08/05/2014 24390342
## 3        22 21979096 21979096     1      *            08/05/2014 24390342
## 4        22 48923459 48923459     1      *            07/28/2014 24322204
## 5        22 36125264 36125264     1      *            07/29/2014 24348519
## 6        22 21431054 21431054     1      *            07/23/2014 24324551
## 7        22 35144411 35144411     1      *            07/23/2014 24324551
## 8        22 22047969 22047969     1      *            07/23/2014 24324551
## 9        22 37586792 37586792     1      *            07/23/2014 24324551
## 10       22 34386473 34386473     1      *            04/26/2014 24165912
## ..      ...      ...      ...   ...    ...                   ...      ...
## Variables not shown: First.Author (chr), Date (chr), Journal (chr), Link
##   (chr), Study (chr), Disease.Trait (chr), Initial.Sample.Size (chr),
##   Replication.Sample.Size (chr), Region (chr), Chr_id (chr), Chr_pos.hg38
##   (dbl), Reported.Gene.s. (chr), Mapped_gene (chr), Upstream_gene_id
##   (chr), Downstream_gene_id (chr), Snp_gene_ids (chr),
##   Upstream_gene_distance (chr), Downstream_gene_distance (chr),
##   Strongest.SNP.Risk.Allele (chr), SNPs (chr), Merged (chr),
##   Snp_id_current (chr), Context (chr), Intergenic (chr),
##   Risk.Allele.Frequency (chr), p.Value (dbl), Pvalue_mlog (dbl),
##   p.Value..text. (chr), OR.or.beta (dbl), X95..CI..text. (chr),
##   Platform..SNPs.passing.QC. (chr), CNV (chr), num.Risk.Allele.Frequency
##   (dbl), chr_hg19 (chr), pos_hg19 (int)
```