gwascat: structuring and querying the NHGRI GWAS catalog

VJ Carey[*]

April 30, 2020

Contents

[*]Generous support of Robert Gentleman and the Computational Biology Group of Genentech, Inc. is gratefully acknowledged

1 Introduction

NHGRI maintains and routinely updates a database of selected genome-wide association studies. This document describes R/Bioconductor facilities for working with contents of this database.

1.1 Installation

The package can be installed using Bioconductor's *BiocManager* package, with the sequence

```
library(BiocManager)
BiocManager::install("gwascat")
```

1.2 Attachment and access to documentation

Once the package has been installed, use `library(gwascat)` to obtain interactive access to all the facilities. After executing this command, use `help(package="gwascat")` to obtain an overview. The current version of this vignette can always be accessed at www.bioconductor.org, or by suitably navigating the web pages generated with `help.start()`.

1.3 Illustrations: computing

Available functions are:

```
> library(gwascat)
> objects("package:gwascat")
```

```
 [1] "bindcadd_snv"          "chklocs"                 "get_cached_gwascat"
 [4] "getRsids"              "getTraits"               "gwcex2gviz"
 [7] "ldtagr"                "locs4trait"              "makeCurrentGwascat"
[10] "obo2graphNEL"          "process_gwas_dataframe" "riskyAlleleCount"
[13] "subsetByChromosome"    "subsetByTraits"          "topTraits"
[16] "traitsManh"
```

The extended GRanges instance with all SNP-disease associations is obtained as follows, using the GRCh38 genome build.

```
> data(ebicat_2020_04_30)
```

To determine the most frequently occurring traits:

```
> topTraits(ebicat_2020_04_30)
```

```
                       Blood protein levels
                                       6517
                  Heel bone mineral density
                                       4503
                            Body mass index
                                       4405
                                     Height
                                       4301
                          Metabolite levels
                                       2385
                    Systolic blood pressure
                                       2223
Educational attainment (years of education)
                                       2126
                              Schizophrenia
                                       2051
                            Type 2 diabetes
                                       1675
         Post bronchodilator FEV1/FVC ratio
                                       1631
```

For a given trait, obtain a GRanges with all recorded associations; here only three associations are shown:

```
> subsetByTraits(ebicat_2020_04_30, tr="LDL cholesterol")[1:3]
```

```
gwasloc instance with 3 records and 38 attributes per record.
Extracted:  2020-04-30 23:24:51
metadata()$badpos includes records for which no unique locus was given.
Genome:  GRCh38
Excerpt:
GRanges object with 3 ranges and 3 metadata columns:
      seqnames    ranges strand |   DISEASE/TRAIT        SNPS   P-VALUE
         <Rle> <IRanges>  <Rle> |     <character> <character> <numeric>
  [1]       19  19678719      * | LDL cholesterol   rs2304130     3e-06
  [2]        7  21567734      * | LDL cholesterol  rs12670798     6e-09
  [3]       11  61829740      * | LDL cholesterol    rs174570     4e-13
  -------
  seqinfo: 24 sequences from GRCh38 genome
```

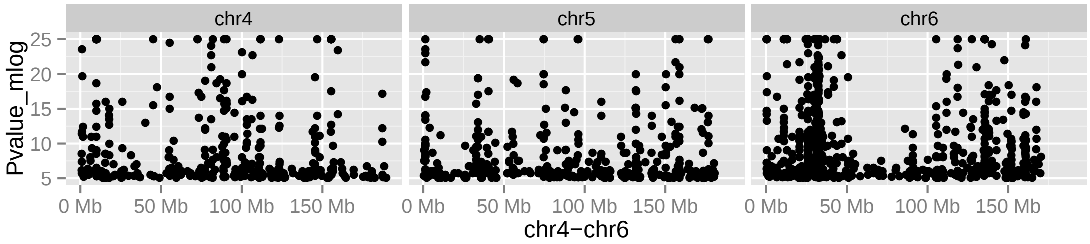
2 Some visualizations

2.1 Basic Manhattan plot

A basic Manhattan plot is easily constructed with the ggbio package facilities. Here we confine attention to chromosomes 4:6. First, we create a version of the catalog with $-log_{10}p$ truncated at a maximum value of 25.

```
> gwtrunc = ebicat_2020_04_30
> requireNamespace("S4Vectors")
> mcols = S4Vectors::mcols
> mlpv = mcols(ebicat_2020_04_30)$PVALUE_MLOG
> mlpv = ifelse(mlpv > 25, 25, mlpv)
> S4Vectors::mcols(gwtrunc)$PVALUE_MLOG = mlpv
> library(GenomeInfoDb)
> seqlevelsStyle(gwtrunc) = "UCSC"
> gwlit = gwtrunc[ which(as.character(seqnames(gwtrunc)) %in% c("chr4", "chr5", "chr6
> library(ggbio)
> mlpv = mcols(gwlit)$PVALUE_MLOG
> mlpv = ifelse(mlpv > 25, 25, mlpv)
> S4Vectors::mcols(gwlit)$PVALUE_MLOG = mlpv

> methods:::bind_activation(FALSE)
> autoplot(gwlit, geom="point", aes(y=PVALUE_MLOG), xlab="chr4-chr6")
```



2.2 Annotated Manhattan plot

A simple call permits visualization of GWAS results for a small number of traits. Note the defaults in this call.

```
> traitsManh(gwtrunc)
```

2.3 Integrative view of potential genetic determinants

The following chunk uses GFF3 data on eQTL and related phenomena distributed at the GBrowse instance at eqtl.uchicago.edu. A request for all information at 43-45 Mb was made on 2 June 2012, yielding the GFF3 referenced below. Of interest are locations and scores of genetic associations with DNaseI hypersensitivity (scores identifying dsQTL, see Degner et al 2012).

```
> gffpath = system.file("gff3/chr17_43000000_45000000.gff3", package="gwascat")
> library(rtracklayer)
> c17tg = import(gffpath)
```

We make a Gviz DataTrack of the dsQTL scores.

```
> c17td = c17tg[ which(S4Vectors::mcols(c17tg)$type == "Degner_dsQTL") ]
> library(Gviz)
> dsqs = DataTrack( c17td, chrom="chr17", genome="hg19", data="score",
+   name="dsQTL")
```

We start the construction of the graph here.

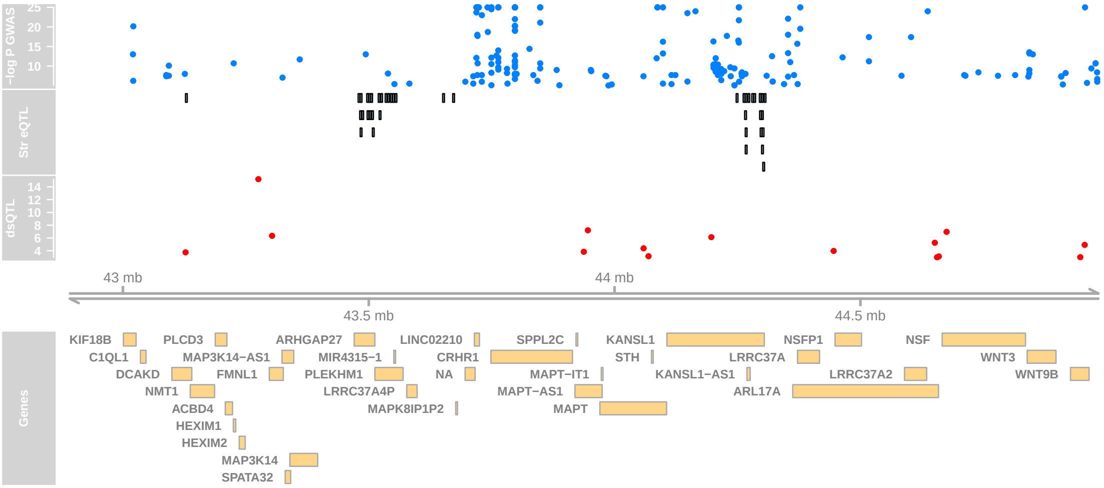
```
> g2 = GRanges(seqnames="chr17", IRanges(start=4.3e7, width=2e6))
> seqlevelsStyle(ebicat_2020_04_30) = "UCSC"
> basic = gwcex2gviz(basegr = ebicat_2020_04_30, contextGR=g2, plot.it=FALSE)
```

We also collect locations of eQTL in the Stranger 2007 multipopulation eQTL study.

```
> c17ts = c17tg[ which(S4Vectors::mcols(c17tg)$type == "Stranger_eqtl") ]
> eqloc = AnnotationTrack(c17ts,  chrom="chr17", genome="hg19", name="Str eQTL")
> displayPars(eqloc)$col = "black"
> displayPars(dsqs)$col = "red"
> integ = list(basic[[1]], eqloc, dsqs, basic[[2]], basic[[3]])
```

Now use Gviz.

```
> plotTracks(integ)
```



3 SNP sets and trait sets

3.1 SNPs by name

We can regard the content of a SNP chip as a set of SNP, referenced by name. The pd.genomewidesnp.6 package describes the Affymetrix SNP 6.0 chip. We can determine which traits are associated with loci interrogated by the chip as follows. We work with a subset of the 1 million loci for illustration.

The locon6 data frame has information on 10000 probes, acquired through the following code (not executed here to reduce dependence on the pd.genomewidesnp.6 package, which is very large.

```
> library(pd.genomewidesnp.6)
> con = pd.genomewidesnp.6@getdb()
> locon6 = dbGetQuery(con,
+    "select dbsnp_rs_id, chrom, physical_pos from featureSet limit 10000")
```

Instead use the serialized information:

```
> data(locon6)
> rson6 = as.character(locon6[[1]])
> rson6[1:5]
```

```
[1] "rs2887286"  "rs1496555"  "rs41477744" "rs3890745"  "rs10492936"
```

We subset the GWAS ranges structure with rsids that are common to both the chip and the GWAS catalog. We then tabulate the diseases associated with the common loci.

```
> intr = ebicat_2020_04_30[ intersect(getRsids(ebicat_2020_04_30), rson6) ]
> sort(table(getTraits(intr)), decreasing=TRUE)[1:10]
```

```
            Adolescent idiopathic scoliosis
                                         29
                                     Height
                                         11
                          Metabolite levels
                                          6
                       Blood protein levels
                                          4
Educational attainment (years of education)
                                          4
                                     Asthma
                                          3
                            Body mass index
                                          3
                              Breast cancer
                                          3
                  C-reactive protein levels
                                          3
                  Heel bone mineral density
                                          3
```

3.2 Traits by genomic location

We will assemble genomic coordinates for SNP on the Affymetrix 6.0 chip and show the effects of identifying the trait-associated loci with regions of width 1000bp instead of 1bp.

The following code retrieves coordinates for SNP interrogated on 10000 probes (to save time) on the 6.0 chip, and stores the results in a GRanges instance.

```
> gr6.0 = GRanges(seqnames=ifelse(is.na(locon6$chrom),0,locon6$chrom),
+        IRanges(ifelse(is.na(locon6$phys),1,locon6$phys), width=1))
> S4Vectors::mcols(gr6.0)$rsid = as.character(locon6$dbsnp_rs_id)
> seqlevels(gr6.0) = paste("chr", seqlevels(gr6.0), sep="")
```

Here we compute overlaps with both the raw disease-associated locus addresses, and with the locus address ± 500bp.

```
> ag = function(x) as(x, "GRanges")
> ovraw = suppressWarnings(subsetByOverlaps(ag(ebicat_2020_04_30), gr6.0))
> length(ovraw)
```

```
[1] 1
```

```
> ovaug = suppressWarnings(subsetByOverlaps(ag(ebicat_2020_04_30+500), gr6.0))
> length(ovaug)
```

```
[1] 517
```

To acquire the subset of the catalog to which 6.0 probes are within 500bp, use:

```
> rawrs = mcols(ovraw)$SNPS
> augrs = mcols(ovaug)$SNPS
> ebicat_2020_04_30[augrs]
```

```
gwasloc instance with 517 records and 38 attributes per record.
Extracted:  2020-04-30 23:24:51
metadata()$badpos includes records for which no unique locus was given.
Genome:  GRCh38
Excerpt:
GRanges object with 5 ranges and 3 metadata columns:
      seqnames    ranges strand |       DISEASE/TRAIT        SNPS   P-VALUE
         <Rle> <IRanges>  <Rle> |         <character> <character> <numeric>
  [1]    chr10  58153390      * |     Crohn's disease   rs1819658     9e-17
  [2]     chr1  67240275      * |     Crohn's disease  rs11209026     1e-64
  [3]    chr10  55657201      * | Cardiac hypertrophy   rs1916521     5e-07
  [4]     chr1 207478831      * |     Type 2 diabetes  rs17045328     7e-06
  [5]     chr1 165439858      * |                AIDS  rs10800098     4e-06
  -------
  seqinfo: 24 sequences from GRCh38 genome
```

Relaxing the intersection criterion in this limited case leads to a larger set of traits.

```
> setdiff( getTraits(ebicat_2020_04_30[augrs]), getTraits(ebicat_2020_04_30[rawrs]) )
```

```
  [1] "Crohn's disease"
  [2] "Cardiac hypertrophy"
  [3] "Type 2 diabetes"
  [4] "AIDS"
  [5] "Waist-hip ratio"
  [6] "Speech perception in dyslexia"
  [7] "Diabetic retinopathy"
  [8] "Two-hour glucose challenge"
  [9] "Malaria"
 [10] "Metabolite levels"
 [11] "Cognitive performance"
 [12] "Mean corpuscular hemoglobin concentration"
 [13] "Melanoma"
 [14] "Metabolite levels (MHPG)"
 [15] "Axial length"
```

```
[16] "IgG glycosylation"
[17] "Basal cell carcinoma"
[18] "Corneal structure"
[19] "RR interval (heart rate)"
[20] "Platelet count"
[21] "Obesity"
[22] "Testicular germ cell tumor"
[23] "Bipolar disorder"
[24] "Pit-and-Fissure caries"
[25] "Tonometry"
[26] "Epilepsy"
[27] "Hypertension"
[28] "Formal thought disorder in schizophrenia"
[29] "Orofacial clefts"
[30] "Urate levels"
[31] "Palmitic acid (16:0) levels"
[32] "Obesity-related traits"
[33] "Alzheimer's disease (cognitive decline)"
[34] "Pediatric non-alcoholic fatty liver disease activity score"
[35] "Alzheimer's disease (late onset)"
[36] "Periodontitis (Mean PAL)"
[37] "Intraocular pressure"
[38] "Serum alkaline phosphatase levels"
[39] "Venous thromboembolism"
[40] "Height"
[41] "Diabetic retinopathy (all NPDR and PDR)"
[42] "Vertical cup-disc ratio (adjusted for vertical disc diameter)"
[43] "Brain region volumes"
[44] "Appendicular lean mass"
[45] "Eosinophil counts"
[46] "General risk tolerance (MTAG)"
[47] "Heel bone mineral density"
[48] "Mean corpuscular hemoglobin"
[49] "Red blood cell count"
[50] "Neuroticism"
[51] "Blond vs. brown/black hair color"
[52] "Male-pattern baldness"
[53] "3-month functional outcome in ischaemic stroke (modified Rankin score)"
[54] "Adventurousness"
[55] "Myopia (age of diagnosis)"
[56] "Adolescent idiopathic scoliosis"
[57] "Red cell distribution width"
```

```
[58] "Spherical equivalent or myopia (age of diagnosis)"
[59] "Age at menopause"
[60] "Age spots"
[61] "Cardiovascular disease"
[62] "General cognitive ability"
[63] "Lung function (FEV1/FVC)"
[64] "White blood cell count"
[65] "Chlamydia trachomatis seropositivity"
[66] "Respiratory diseases"
[67] "Eczema"
[68] "Glaucoma (primary open-angle)"
[69] "Hair color"
[70] "Household income (MTAG)"
[71] "Hay fever and/or eczema"
[72] "Urinary calcium excretion"
[73] "White matter microstructure (axial diusivities)"
[74] "Systolic blood pressure"
[75] "Cognitive ability, years of educational attainment or schizophrenia (pleiotropy)
[76] "White matter microstructure (mode of anisotropy)"
[77] "Glucose homeostasis traits"
[78] "Coronary artery calcification"
[79] "Anticoagulant levels"
[80] "Paget's disease"
[81] "Dental caries"
[82] "Gaucher disease severity"
[83] "Plasma omega-3 polyunsaturated fatty acid level (eicosapentaenoic acid)"
[84] "Colorectal cancer (diet interaction)"
[85] "Migraine without aura"
[86] "Migraine with aura"
[87] "Serum metabolite levels"
[88] "Fibrinogen"
[89] "Serum thyroid-stimulating hormone levels"
[90] "Asthma (sex interaction)"
[91] "Blood metabolite ratios"
[92] "Response to TNF inhibitor in rheumatoid arthritis (erythrocyte sedimentation rat
[93] "Response to TNF inhibitor in rheumatoid arthritis (change in swollen 28-joint co
[94] "Birth weight"
[95] "Shingles"
[96] "Lung function (FVC)"
[97] "Nonalcoholic fatty liver disease"
[98] "C-reactive protein levels"
[99] "Bone mineral density (hip)"
```

```
[100] "Diverticular disease"
[101] "Glomerular filtration rate in diabetes"
[102] "Coffee consumption"
[103] "Tea consumption"
[104] "Depression"
[105] "Diastolic blood pressure"
[106] "Body mass index"
[107] "Cerebrospinal fluid p-tau levels"
[108] "Blood protein levels"
[109] "Erectile dysfunction"
[110] "Lifetime smoking index"
[111] "Bipolar disorder or body mass index"
[112] "Non-lobar intracerebral hemorrhage (MTAG)"
[113] "Amblyopia"
[114] "Metabolic syndrome"
[115] "Waist-to-hip ratio adjusted for BMI (additive genetic model)"
[116] "LDL cholesterol"
[117] "Total cholesterol levels"
[118] "Worry/vulnerability (special factor of neuroticism)"
[119] "Visceral fat"
[120] "Cortical brain region measurements (area, volume and thickness)"
[121] "Smoking cessation (MTAG)"
[122] "Age of smoking initiation (MTAG)"
[123] "Red blood cell traits"
[124] "Coenzyme Q10 levels"
[125] "Vitamin D levels"
[126] "Systemic sclerosis"
[127] "Response to bronchodilator in chronic obstructive pulmonary disease (change in F
[128] "Total body bone mineral density (age over 60)"
[129] "Subjective well-being (MTAG)"
[130] "Squamous cell lung carcinoma"
[131] "Fear of minor pain"
[132] "Triglycerides"
[133] "Hepatitis A"
[134] "Lung cancer in never smokers"
[135] "Measles"
[136] "Lung cancer"
[137] "QRS duration"
[138] "6-month creatinine clearance change response to tenofovir treatment in HIV infec
[139] "Facial morphology (factor 13, vertical position of alar curvature relative to up
[140] "Body fat mass"
[141] "Waist-to-hip ratio adjusted for BMI"
```

```
[142] "Lean body mass"
[143] "Facial morphology (factor 20)"
[144] "Psoriasis"
[145] "Corneal astigmatism"
[146] "Feeling worry"
[147] "Fractional shortening"
[148] "Calcium levels"
[149] "Rheumatoid arthritis (ACPA-positive)"
[150] "Prothrombin time"
[151] "Estimated glomerular filtration rate"
[152] "Cognitive ability (MTAG)"
[153] "Benign prostatic hyperplasia and/or lower urinary tract symptoms"
[154] "Common carotid intima-media thickness in HIV negative individuals"
[155] "Dupuytren's disease"
[156] "Blood urea nitrogen levels"
[157] "Nonsyndromic cleft lip"
[158] "Intake of total sugars"
[159] "Estimated glomerular filtration rate in non-diabetics"
[160] "Psoriasis vulgaris"
[161] "Obstructive sleep apnea trait (apnea hypopnea index)"
[162] "Post bronchodilator FEV1/FVC ratio"
[163] "Attention function in attention deficit hyperactive disorder"
[164] "Pediatric autoimmune diseases"
[165] "Platelet thrombus formation"
[166] "Mild influenza (H1N1) infection"
[167] "Gut microbiome composition (summer)"
[168] "Inflammatory bowel disease"
[169] "Non-alcoholic fatty liver disease histology (other)"
[170] "C-reactive protein"
[171] "Carotid plaque burden (smoking interaction)"
[172] "Schizophrenia"
[173] "Post bronchodilator FEV1"
[174] "Gestational age at birth in premature rupture of membrane-initiated deliveries (
[175] "Post bronchodilator FEV1/FVC ratio in COPD"
[176] "Metabolite levels (small molecules and protein measures)"
[177] "Extraversion"
[178] "Pelvic organ prolapse (moderate/severe)"
[179] "Parental extreme longevity (95 years and older)"
[180] "Mean corpuscular volume"
[181] "Granulocyte percentage of myeloid white cells"
[182] "Monocyte percentage of white cells"
[183] "Educational attainment (college completion)"
```

```
[184] "High light scatter reticulocyte count"
[185] "Interferon gamma levels"
[186] "Photic sneeze reflex"
[187] "Resting heart rate"
[188] "Daytime sleep phenotypes"
[189] "Night sleep phenotypes"
[190] "Pulse pressure"
[191] "Neutrophil percentage of white cells"
[192] "Myeloid white cell count"
[193] "Chronic inflammatory diseases (ankylosing spondylitis, Crohn's disease, psoriasi
[194] "Nose size"
[195] "Urate levels in overweight individuals"
[196] "Breast cancer"
[197] "Food allergy"
[198] "Bipolar disorder (body mass index interaction)"
[199] "Vein graft stenosis in coronary artery bypass grafting"
[200] "Anxiety disorder"
[201] "Vogt-Koyanagi-Harada syndrome"
[202] "Macrophage inflammatory protein 1b levels"
[203] "Stem cell factor levels"
[204] "Colorectal adenoma (advanced)"
[205] "Colorectal cancer"
[206] "Glomerular filtration rate (creatinine)"
[207] "Asthma"
[208] "Alcohol consumption (drinks per week) (MTAG)"
[209] "Smoking initiation (ever regular vs never regular) (MTAG)"
[210] "Systemic lupus erythematosus"
[211] "Monoclonal gammopathy of undetermined significance"
[212] "Medication use (agents acting on the renin-angiotensin system)"
[213] "Cognitive empathy"
[214] "Metastasis in stage I-III microsatellite instability low/stable colorectal cance
[215] "Strabismus"
[216] "Uterine fibroids"
[217] "Response to cognitive-behavioural therapy in anxiety disorder"
[218] "Positive urgency"
[219] "Educational attainment (years of education)"
[220] "Bone mineral density (femoral neck)"
[221] "Educational attainment (MTAG)"
[222] "Self-reported math ability (MTAG)"
[223] "Highest math class taken (MTAG)"
[224] "Self-reported math ability"
```

4 Counting alleles associated with traits

We can use `riskyAlleleCount` to count risky alleles enumerated in the GWAS catalog. This particular function assumes that we have genotyped at the catalogued loci. Below we will discuss how to impute from non-catalogued loci to those enumerated in the catalog.

```
> data(gg17N) # translated from GGdata chr 17 calls using ABmat2nuc
> gg17N[1:5,1:5]

        rs6565733 rs1106175 rs17054921 rs8064924 rs8070440
NA06985 "G/G"     "A/G"     "C/C"      "G/G"     "G/G"
NA06991 "G/G"     "A/A"     "C/C"      "G/G"     "G/G"
NA06993 "G/G"     "A/A"     "C/C"      "G/G"     "G/G"
NA06994 "A/G"     "A/G"     "C/C"      "A/G"     "G/G"
NA07000 "G/G"     "A/A"     "C/C"      "G/G"     "G/G"
```

This function can use genotype information in the A/B format, assuming that B denotes the alphabetically later nucleotide. Because we have direct nucleotide coding in our matrix, we set the `matIsAB` parameter to false in this call.

```
> h17 = riskyAlleleCount(gg17N, matIsAB=FALSE, chr="ch17",
+  gwwl = ebicat_2020_04_30)
> h17[1:5,1:5]

        rs7217319 rs2034088 rs741677 rs9907102 rs12938449
NA06985         0         0        0         0          0
NA06991         0         1        0         0          1
NA06993         0         1        0         0          0
NA06994         0         2        0         0          0
NA07000         0         1        0         0          0

> table(as.numeric(h17))

    0     1     2
92390 34509 25561
```

It is of interest to bind the counts back to the catalog data.

```
> gwr = ebicat_2020_04_30
> gwr = gwr[colnames(h17),]
> S4Vectors::mcols(gwr) = cbind(mcols(gwr), DataFrame(t(h17)))
> sn = rownames(h17)
> gwr[,c("DISEASE/TRAIT", sn[1:4])]
```

```
gwasloc instance with 1694 records and 5 attributes per record.
Extracted:  2020-04-30 23:24:51
metadata()$badpos includes records for which no unique locus was given.
Genome:  GRCh38
Excerpt:
GRanges object with 5 ranges and 5 metadata columns:
      seqnames    ranges strand |
         <Rle> <IRanges>  <Rle> |
  [1]    chr17    189133      * |
  [2]    chr17    519811      * |
  [3]    chr17    560603      * |
  [4]    chr17    561592      * |
  [5]    chr17    583581      * |


  [1]
  [2]                                                                   Hip circumferen
  [3] Waist circumference adjusted for BMI (joint analysis main effects and physical ac
  [4]
  [5]
        NA06985   NA06991   NA06993   NA06994
      <integer> <integer> <integer> <integer>
  [1]         0         0         0         0
  [2]         0         1         1         2
  [3]         0         0         0         0
  [4]         0         0         0         0
  [5]         0         1         0         0
  -------
  seqinfo: 24 sequences from GRCh38 genome
```

Now by programming on the metadata columns, we can identify individuals with particular risk profiles.

5 Formal management of trait vocabularies

5.1 Diseases: Disease Ontology

The Disease Ontology project **?** formalizes a vocabulary for human diseases. Bioconductor's DO.db package is a curated representation.

```
> library(DO.db)
> DO()
```

```
Quality control information for DO:


This package has the following mappings:

DOANCESTOR has 6569 mapped keys (of 6570 keys)
DOCHILDREN has 1811 mapped keys (of 6570 keys)
DOOBSOLETE has 2374 mapped keys (of 2374 keys)
DOOFFSPRING has 1811 mapped keys (of 6570 keys)
DOPARENTS has 6569 mapped keys (of 6570 keys)
DOTERM has 6570 mapped keys (of 6570 keys)


Additional Information about this package:

DB schema: DO_DB
DB schema version: 1.0
```

All tokens of the ontology are acquired via:

```
> alltob = unlist(mget(mappedkeys(DOTERM), DOTERM))
> allt = sapply(alltob, Term)
> allt[1:5]
```

```
                  DOID:0001816                      DOID:0002116
                "angiosarcoma"                       "pterygium"
                  DOID:0014667                      DOID:0050004
         "disease of metabolism" "seminal vesicle acute gonorrhea"
                  DOID:0050012
                 "chikungunya"
```

Direct mapping from disease trait tokens in the catalog to this vocabulary succeeds for a modest proportion of records.

```
> cattra = mcols(ebicat_2020_04_30)$`DISEASE/TRAIT`
> mat = match(tolower(cattra), tolower(allt))
> catDO = names(allt)[mat]
> na.omit(catDO)[1:50]
```

```
 [1] "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778"
 [7] "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778"
[13] "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778"
[19] "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778"
[25] "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778"
```

```
[31] "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778"
[37] "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778"
[43] "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778" "DOID:8778"
[49] "DOID:8778" "DOID:8778"
```

```
> mean(is.na(catDO))
```

```
[1] 0.9054001
```

Approximate matching of unmatched tokens can proceed by various routes. Some traits are not diseases, and will not be mappable using Disease Ontology. However, consider

```
> unique(cattra[is.na(catDO)])[1:20]
```

```
 [1] "Congenital heart malformation"
 [2] "Body mass in chronic obstructive pulmonary disease"
 [3] "Alcohol consumption (transferrin glycosylation)"
 [4] "Sudden cardiac arrest"
 [5] "Orofacial clefts (maternal alcohol consumption interaction)"
 [6] "Height"
 [7] "Mean forced vital capacity from 2 exams"
 [8] "Pulmonary function"
 [9] "Glioma"
[10] "Bone mineral density (hip)"
[11] "Bone mineral density (spine)"
[12] "Idiopathic dilated cardiomyopathy"
[13] "Osteoporosis-related phenotypes"
[14] "Waist circumference"
[15] "Cutaneous nevi"
[16] "Primary biliary cholangitis"
[17] "Cardiac hypertrophy"
[18] "Adiposity"
[19] "Uric acid levels"
[20] "Prostate-specific antigen levels"
```

```
> nomatch = cattra[is.na(catDO)]
> unique(nomatch)[1:5]
```

```
[1] "Congenital heart malformation"
[2] "Body mass in chronic obstructive pulmonary disease"
[3] "Alcohol consumption (transferrin glycosylation)"
[4] "Sudden cardiac arrest"
[5] "Orofacial clefts (maternal alcohol consumption interaction)"
```

Manual searching shows that a number of these have very close matches.

5.2 Other phenotypic traits: Human Phenotype Ontology

Bioconductor does not possess an annotation package for phenotype ontology, but the standardized OBO format can be parsed and modeled into a graph.

```
> hpobo = gzfile(dir(system.file("obo", package="gwascat"), pattern="hpo", full=TRUE)
> HPOgraph = obo2graphNEL(hpobo)
> close(hpobo)
```

The phenotypic terms are obtained via:

```
> requireNamespace("graph")
> hpoterms = unlist(graph::nodeData(HPOgraph, graph::nodes(HPOgraph), "name"))
> hpoterms[1:10]
```

```
                                HP:0000001 
                                     "All" 
                                HP:0000002 
              "Abnormality of body height" 
                                HP:0000003 
            "Multicystic kidney dysplasia" 
                                HP:0000004 
               "Onset and clinical course" 
                                HP:0000005 
                     "Mode of inheritance" 
                                HP:0000006 
          "Autosomal dominant inheritance" 
                                HP:0000007 
         "Autosomal recessive inheritance" 
                                HP:0000008 
"Abnormality of female internal genitalia" 
                                HP:0000009 
   "Functional abnormality of the bladder" 
                                HP:0000010 
      "Recurrent urinary tract infections" 
```

Exact hits to unmatched GWAS catalog traits exist:

```
> intersect(tolower(nomatch), tolower(hpoterms))
```

```
 [1] "glioma"                         "stroke"                        
 [3] "autism"                         "glioblastoma"                  
 [5] "knee osteoarthritis"            "coronary artery calcification" 
 [7] "hearing impairment"             "nephropathy"                   
 [9] "hypertriglyceridemia"          "cirrhosis"
[11] "insomnia"                      "depression"
[13] "eczema"                        "nasal polyps"
[15] "cognitive impairment"          "eating disorders"
[17] "ischemic stroke"               "iga nephropathy"
[19] "ketonuria"                     "hematuria"
[21] "neurofibrillary tangles"       "retinal arteriolar tortuosity"
[23] "hashimoto thyroiditis"         "selective iga deficiency"
[25] "headache"                      "calcific aortic valve stenosis"
[27] "thrombosis"                    "febrile seizures"
[29] "dysphagia"                     "excessive daytime sleepiness"
[31] "sagittal craniosynostosis"     "benign prostatic hyperplasia"
[33] "bicuspid aortic valve"         "orthostatic hypotension"
[35] "freckling"                     "impulsivity"
```

More work on formalization of trait terms is underway.

6 CADD scores

Kircher et al. (**?**) define combined annotation-dependent depletion scores measuring variant pathogenicity in an integrative way. Small requests to bind scores for SNV to GRanges can be resolved through HTTP; large requests can be carried out on a local tabix-indexed selection from their archive.

```
> g3 = as(ebicat_2020_04_30, "GRanges")
> bg3 = bindcadd_snv( g3[which(seqnames(g3)=="chr3")][1:20] )
> inds = ncol(mcols(bg3))
> bg3[, (inds-3):inds]
```

This requires cooperation of network interface and server, so we don't evaluate in vignette build but on 1 Apr 2014 the response was:

```
GRanges with 20 ranges and 4 metadata columns:
       seqnames                 ranges strand   |         Ref         Alt
          <Rle>              <IRanges>  <Rle>   | <character> <character>
   [1]        3 [109789570, 109789570]      *   |           A           G
   [2]        3 [ 25922285,  25922285]      *   |           G           A
   [3]        3 [109529550, 109529550]      *   |           T           C
   [4]        3 [175055759, 175055759]      *   |           T           G
   [5]        3 [191912870, 191912870]      *   |           C           T
   ...      ...                    ...    ... ...         ...         ...
  [16]        3 [187716886, 187716886]      *   |           A           G
  [17]        3 [160820524, 160820524]      *   |           G           C
  [18]        3 [169518455, 169518455]      *   |          T          C
  [19]        3 [179172979, 179172979]      *   |          G          T
  [20]        3 [171785168, 171785168]      *   |          G          C
          CScore     PHRED
       <numeric> <numeric>
   [1] -0.182763     3.110
   [2] -0.289708     2.616
   [3]  0.225373     5.216
   [4] -0.205689     3.003
   [5] -0.172189     3.161
   ...       ...       ...
  [16] -0.019710     3.913
  [17] -0.375183     2.235
  [18] -0.695270     0.987
  [19] -0.441673     1.949
  [20]  0.231972     5.252
  ---
  seqlengths:
           1         2         3         4 ...        21        22         X
   249250621 243199373 198022430 191154276 ...  48129895  51304566 155270560
```

7 Appendix: Adequacy of location annotation

A basic question concerning the use of archived SNP identifiers is durability of the association between asserted location and SNP identifier. The chklocs function uses a current Bioconductor SNPlocs package to check this.

For example, to verify that locations asserted on chromosome 20 agree between the Bioconductor dbSNP image and the gwas catalog,

```
> if ("SNPlocs.Hsapiens.dbSNP144.GRCh37" %in% installed.packages()[,1]) {
+  library(SNPlocs.Hsapiens.dbSNP144.GRCh37)
+  chklocs("20", ebicat_2020_04_30)
+ }
```

This is not a fast procedure.