G2PML: tutorial on basic usage of the tool

Juan A. Botía

25/04/2019

Introduction

G2PML is set of datasets and software to study monogenic diseases. We study the diseases through the genes already discovered to be associated to the disease in the Mendelian way (i.e. one mutation, one case). These genes form what is known as a panel of genes.

Our approach to the disease is by describing genes in terms of (a) genetic constraint related properties, (b) genomic related properties and (c) transcriptomics.

With G2PML and your panel P={g1,g2,. . . ,gn} of choice, where gi is the i-th gene, you can

- Perform an analysis on which features are relevant to distinguish those genes from the rest
- Create prediction models to predict new genes from the set of P genes and annotate and score the predictions.
- Enlarge your DDBB of properties to incorporate new ones to enrich your analysis

In our studies, we have been using the PanelApp database of gene panels, from Genomics England, accesible at http://panelapp.genomicsengland.co.uk. In the package we incorporta functions to navigate the panels and access the genes. For example, in

```
library(G2PML)
panels = getPanelsFromPanelApp()
names(panels)
```

```
##  [1] "Name"                "Number_of_Regions"  "CurrentCreated"
##  [4] "Status"              "Number_of_STRs"     "Relevant_disorders"
##  [7] "CurrentVersion"      "DiseaseSubGroup"    "DiseaseGroup"
## [10] "Panel_Id"            "Number_of_Genes"    "PanelTypes"
```

```
dim(panels)
```

```
## [1] 173  12
```

```
panels$Name[1:10]
```

```
##  [1] "Adult solid tumours for rare disease"
##  [2] "Amelogenesis imperfecta"
##  [3] "Amyotrophic lateral sclerosis/motor neuron disease"
##  [4] "Anophthalmia or microphthalmia"
##  [5] "Arrhythmogenic cardiomyopathy"
##  [6] "Arthrogryposis"
##  [7] "Atypical haemolytic uraemic syndrome"
##  [8] "Auditory Neuropathy Spectrum Disorder"
##  [9] "Autosomal recessive congenital ichthyosis"
## [10] "Beckwith-Wiedemann syndrome (BWS) and other congenital overgrowth disorders"
```

We can access 173 gene panels and their genes. For example, we can get the genes for monogenic forms of PD as follows

```r
genes = getGenesFromPanelApp (disorder="Neurology and neurodevelopmental disorders",
                    panel="Parkinson Disease and Complex Parkinsonism",
                    color="green")
genes
```

```
##  [1] "PRKN"     "ATP13A2"  "ATP1A3"   "C19orf12" "CSF1R"    "DCTN1"
##  [7] "DNAJC6"   "FBXO7"    "FTL"      "GBA"      "GCH1"     "GRN"
## [13] "LRRK2"    "LYST"     "MAPT"     "OPA3"     "PANK2"    "PARK7"
## [19] "PINK1"    "PLA2G6"   "PRKRA"    "RAB39B"   "SLC30A10" "SLC39A14"
## [25] "SLC6A3"   "SNCA"     "SPG11"    "SPR"      "SYNJ1"    "TH"
## [31] "TUBB4A"   "VPS13A"   "VPS35"    "WDR45"
```

We'll take those genes as an example on how to use GP2ML to study the phenotype.

Relevant features on your gene panel

At this point, we have ready our gene list to study. And we want to know what makes it different this list of genes from the rest of genes in the genome (note we only consider for now protein coding genes as many of the features we use to desdribe them just only make sense for that gene biotype).

Note that we have many different properties to describe our genes. To quickly get a hint on how these genes are described, we can do

```r
genedata = G2PML::fromGenes2MLData(genes=genes,which.controls="allgenome")
```

```
## ALLGENOME mode: We have to remove 34 disease genes from the all genome controls
## We have to remove 52 genes RTF2 MTREX TUT7 RAB5IF PPP1R2C  because we don't know them
## Removing columns for attributes DPI and DSI and ESTcount and constitutiveexons
```

```r
dim(genedata)
```

```
## [1] 17636   161
```

By simply doing that, we have created an annotation dataset for our genes (and also for the rest of genes in the genome), i.e. 17635 genes including the PD ones, with 161 features of interest. A quick look at the features we use . . .

```r
str(genedata[,1:30])
```

```
## 'data.frame':    17636 obs. of  30 variables:
##  $ gene                                : chr  "PRKN" "ATP13A2" "ATP1A3" "C19orf12" ...
##  $ GeneLength                          : num  1380351 25970 30915 16903 60081 ...
##  $ TranscriptCount                     : num  11 15 13 9 9 ...
##  $ CountsOverlap                       : int  3 1 0 0 2 2 3 0 0 1 ...
##  $ NumJunctions                        : int  65 145 120 13 65 287 76 51 3 66 ...
##  $ IntronicLength                      : int  1375858 20744 25870 11424 54953 20527 160766 15642 703
##  $ GCcontent                           : num  40.4 57.2 56.1 50.4 48.1 ...
##  $ String                              : num  1.61 5 3 0 8 ...
##  $ LoFTool                             : num  0.46028 0.0289 0.00296 0.041 0.15 ...
##  $ EvoTol                              : num  50.31 1.16 15.59 11.61 19.18 ...
##  $ RVIS                                : num  1.035 0.975 0.558 0.986 0.942 ...
##  $ pAD                                 : num  0.436 0.26 1 0.995 0.787 ...
##  $ pAR                                 : num  0.564 0.74 0 0.005 0.213 ...
##  $ gnomadpLI                           : num  0.356 0.125 1 0.466 0.356 ...
##  $ gnomadpRec                          : num  0.6376 1 0.0333 0.6525 0.6376 ...
##  $ gnomadpNull                         : num  3.13e-01 4.48e-02 5.15e-05 2.01e-01 3.13e-01 ...
##  $ gnomadpMiss                           : num  1.067 1.48 6.589 0.531 1.595 ...
##  $ gnomadOELoF                           : num  0.6256 0.4392 0.0414 0.4902 0.6256 ...
##  $ gnomadOEMiss                          : num  0.908 0.918 0.292 1.009 0.922 ...
##  $ RankedMMSpecificRankMMAdiposeSub      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMAdiposeVisceral: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMAdrenalGland    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMArteryAorta     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMArteryCoronary : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMArteryTibial    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMAmygdala        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMAntCingCortex  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMCaudate         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMCerebHemisphere: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RankedMMSpecificRankMMCerebellum      : int  0 0 0 0 0 0 0 0 0 0 ...
```

And we see we have the gene name, and many other features of interest that we describe elsewhere. But basically, between `GeneLength` and `GCcontent` we have genomic properties, `String` gives an indication on how active (or sticky) is the gene product (i.e. the protein). Then we have from `LoFTool` to `gnomadOEMiss` attributes that try to describe how resistant or fragile the gene is to mutation. The rest of properties are related with the expression and co-expression in GTEx 47 tissues.

Finally, we have

```
table(genedata$condition)
```

```
##
##    Disease Nondisease
##         34      17602
```

that helps us distinguish between our genes and the rest.

Now that we have this, we can perform an analysis on what are the most relevant features to distinguish, from a machine learning perspective, between your genes and the rest. For that we can perform a Caret based feature selection analysis. But we have to take into account that highly inbalance in the dataset, i.e. in the case of PD genes, we have 517 non panel genes for each of the genes in the panel. Also note that we expect that some of those genes not in the panel should also be there. And that is precissely what we aim for: to discover such genes.

We can perform a feature selection analysis simply doing this (it will take a while):

```
fspd = featureSelection(genes=genes,k=10,repeats=40,controls="allgenome")
```

By using Caret and recursive feature subset eliminination strategy http://topepo.github.io/caret/recursive-feature-elimination.html averaged many times on perfectly balanced subproblems (i.e. 50% PD genes, 50% rest of protein coding genes) we will get a result set. Let us suppose we have it saved and we load it now

```
fspd = readRDS("~/Dropbox/KCL/workspace/G2PML/inst/g2pml/Parkinson_Disease_and_Complex_Parkinsonism.fs.
```

```
## Warning: replacing previous import 'ggplot2::empty' by 'plyr::empty' when
## loading 'caret'
```

```
metadata = G2PML::getVarsMetaDataFromFS(fspd,r=0.4)
```

```
## Working now with Unknown
```

```
eff = order(as.numeric(metadata$meaneffects[,"meaneffect"]))
metadata$meaneffects[,"meaneffect"][eff]
```

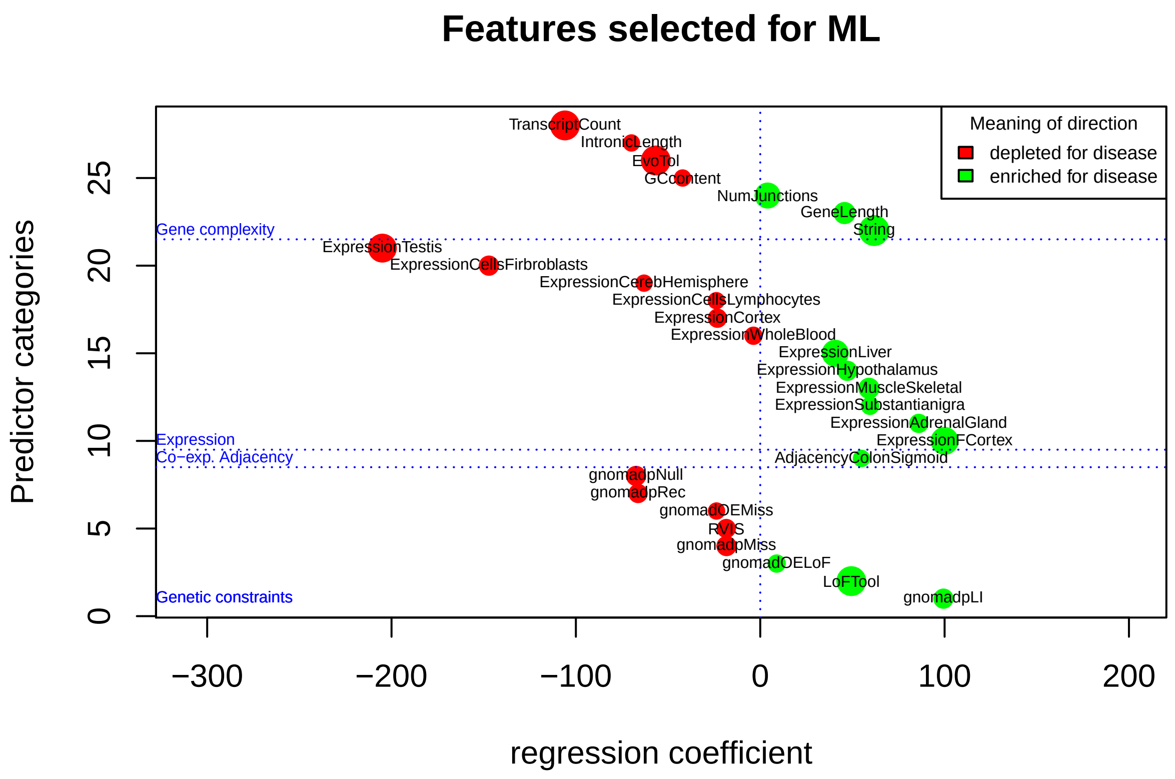
```
##          ExprSpecificFCortex                     gnomadpLI
##          "-0.300328945335776"          "-0.298463167912373"
##      ExprSpecificAdrenalGland                        String
##          "-0.258243695475316"          "-0.185526933091989"
##   ExprSpecificSubstantianigra    ExprSpecificMuscleSkeletal
##          "-0.178792394506871"          "-0.177154291999409"
##    AdjSpecificAdjColonSigmoid                       LoFTool
##          "-0.164954367083234"          "-0.148451430862964"
##      ExprSpecificHypothalamus                    GeneLength
##          "-0.141897873033249"          "-0.137344584422352"
##             ExprSpecificLiver                   gnomadOELoF
##          "-0.122230605877312"         "-0.0267538849693904"
##                  NumJunctions        ExprSpecificWholeBlood
##         "-0.0120114332543787"          "0.0110227654979671"
##                   gnomadpMiss                          RVIS
##           "0.054961938240615"          "0.0552493714660945"
##            ExprSpecificCortex                  gnomadOEMiss
##          "0.0697060202594493"          "0.0712383162283031"
##  ExprSpecificCellsLymphocytes                     GCcontent
##          "0.0714680772394238"           "0.126443651737662"
##                        EvoTol   ExprSpecificCerebHemisphere
##           "0.170115227227838"           "0.189256662730315"
##                    gnomadpRec                   gnomadpNull
##           "0.199017866281845"           "0.202321949288084"
##                IntronicLength               TranscriptCount
##           "0.209814755444196"           "0.318129228300929"
## ExprSpecificCellsFirbroblasts            ExprSpecificTestis
##           "0.441980705334164"           "0.615581923246568"
```

And from left to right we see the attributes that genes in the panel are enriched for higher values (left) and attributes (right) for which genes in the panel show lower varlues. Only those that show some statistical relevance are shown. We see that in terms of tissue and expression, Frontal cortex, adrenal gland, substantia nigra and scheletal muscle are relevant tissues. We also see that these genes lead to proteins with high interaction coefficients as found in String database, etc.

We can get a nice plot out of it as well

```
G2PML::featureSelectionPlot(fspd,r=0.4)
```

```
## Working now with FALSE
```


Features selected for ML
Predictor categories
regression coefficient
Meaning of direction
🟥 depleted for disease
🟩 enriched for disease
Gene complexity
Expression
Co−exp. Adjacency
Genetic constraints
TranscriptCount
IntronicLength
EvoTol
GCcontent
NumJunctions
GeneLength
String
ExpressionTestis
ExpressionCellsFirbroblasts
ExpressionCerebHemisphere
ExpressionCellsLymphocytes
ExpressionCortex
ExpressionWholeBlood
ExpressionLiver
ExpressionHypothalamus
ExpressionMuscleSkeletal
ExpressionSubstantianigra
ExpressionAdrenalGland
ExpressionFCortex
AdjacencyColonSigmoid
gnomadpNull
gnomadpRec
gnomadOEMiss
RVIS
gnomadpMiss
gnomadOELoF
LoFTool
gnomadpLI
0
5
10
15
20
25
−300
−200
−100
0
100
200