Capture the Genes and Variants Related to a Genetic Disease from Public Databases

Zongfu Cao

National Research Institute for Family Planning

Feb. 1, 2016

Contents

1. Introduction

The rapid development of Genomic technology, especially DNA sequencing, provide a powerful tool to understand the human genome more and more conveniently over the last decade. Human Genome Project[1], the International HapMap Project[2,3], Human Genome Diversity Project[4]，HGDP）and 1000 Genomes Project[5]）have been completed. The rapid progress of the Cancer Genome Atlas[6], International Cancer Genome Consortium(ICGC[7]）and some reports on genetic basis of the common diseases, rare diseases based on the Next-generate sequencing technology, make us understand that there is genetic basis for more and more diseases, which is helpful to predict the risk and early diagnosis of a disease, especially a genetic disease, even response and adverse to a drug[8,9,10].

The development of human genomics, disease genomics and pharmacogenomics is bringing a medical revolution, which will provide a new pattern of health management, prevention, diagnosis and treatment of the diseases, and enter an era of personalized or precision medicine eventually. Many diseases such as cancers, Mendel diseases, are caused by the mutations on the according genes. Genetic testing may be necessary in the study or medical practice of a disease. Targeted sequencing on the candidate genes is one of the feasible methods balancing the need and cost. Genes-variants-phenotype database for a special genetic disease or phenotype needs to be compiled. Firstly, in the design stage, target regions and pathogenic variants need to be determined. Too large regions will increase the sequencing cost and increase the false positive rate, while too small regions will reduce the sensitivity. On the other hand, the result interpretation also needs the clinical evidence of a variant. The information of phenotype and variants relationship is erupting every day by the high throughout technology. So the timely update of the gene-variant-phenotype database is essential. However, it is inefficient and troublesome to capture the information from the literature directly. How to capture the information rapidly and accurately is an issue to be solved.

Fortunately, many integration databases focusing on the relationships among human genes/variants and phenotypes have been public and can be freely accessible, which include HPO, MedGen, GeneReview, OMIM, ClinVar, Orphanet, Uniprot, and COSMIC etc. It may be a solution to capture the genes and variants related to a disease based on the public databases. However, Manual parsing and searching the information from these databases one by one is time-consuming and error-prone. The R package, VarfromPDB, was developed to further ease the use of this vast amount of genomic data. It can be very valuable for R programmers or anyone who is interested in disease-related genes/variants in precision medicine based on the target sequencing strategy using automated scripting.

2. Getting started

The VarfromPDB package provides the following functionalities to capture the genes and variants related to a genetic disease especially Mendel disease from public databases.

1) **Localize the public databases**: *localPDB()* performs the localization of the necessary files in several databases, including HPO, Orphanet, OMIM, ClinVar and Uniprot.
2) **Getting the alias of a genetic disease**: VarfromDB gets the alias of a genetic disease from HPO database for a given keyword, which can be a disease name or a clinical feature.
3) **Capturing the genes related to a genetic disease/phenotype**: The information of relationships among genes and phenotypes are extracted from several public databases including HPO, OMIM, ClinVar, Orphanet, Uniprot respectively. The gene names are transformed into the approved symbols based on HGNC database, and then gotten the union of the relationship pairs by gene symbols/locus and phenotypes.
4) **Capture the variants on the genes related to a known genetic disorder**: The variants on the candidate genes, which may be the interested genes or all the integrated genes from different databases, are extracted from OMIM, ClinVar, and Uniprot database respectively. The according phenotypes are checked whether they are related to the interested disease or clinical feature.
5) **Capture the genes and variants from PubMed**
 What's more, the information can be captured from disease-related abstracts

based on a equery in PubMed by the function extract_pubmed. In the text mining process, the abstracts are separated into multiple words or phrases by the separators such as blank space, prepositions, conjunctions, or articles in the first. Then gene symbols and gene alias in HGNC can be captured, and mutation nomenclature recommendations at the DNA level and protein-level by HGVS are searched by regular expression and the names set of amino acid. The phenotype information can be identified in titles and conclusions by anchoring the high frequently words such as ‘syndrome’, ‘with’, ‘Y-linked’, ‘autosomal dominant’, ‘cause*’, ‘associated’, etc. The mining order of priority is TITLE, CONCLUSION(S), RESULTS, METHODS, and BACKGROUD. The gene names are checked and transformed into approved symbols. The protein change are transformed into 3-character abbreviations of amino acid.

6) Compile the genes and variants

Function genes.compile compiles the union of the different gene sets by the approved gene symbols, and then the according phenotypes in different databases and gene positions are annotated. Function variants.compile compiles the variants from different databases. The variants are compared between ClinVar and other databases. Firstly, the captured variants from OMIM is compared with ClinVar by the OMIM variants ID, and that from UniProt is compared with ClinVar by the protein changes. For PubMed information, both cDNA and protein changes are compared with that in ClinVar. Finally a union of gene-variant-phenotype relationships with the ClinVar-like format can be gotten consisting of the additional variants and the set from ClinVar. Function genes_add_pubmed compares the genes from PubMed abstracts with that from the public databases, then the additional gene-phenotype pairs are added.

2.1 Start up

Assuming that you have installed VarfromPDB package, you first need to load it:

```
> library(VarfromPDB)
```

2.2 Creating a local database

We strongly recommend that you download the files from HPO, MedGen, HGNC, GeneReview, ClinVar, OMIM, Ophanet and Uniprot before you start a job for the first time, which maybe more efficient. All the databases can be free accessed except the OMIM, so you should apply for the FTP URL and API key from OMIM before your first job. Suppose you already have the OMIM FTP URL and API key, omim.url and omim.api. You can just type

```
> localPDB(omim.url= omim.url)
```

Database	File	url
HPO	phenotype_annotation.tab	http://compbio.charite.de/hudson/job/hpo.annotations/lastStableBuild/artifact/misc/phenotype_annotation.tab
	diseases_to_genes.txt	http://compbio.charite.de/hudson/job/hpo.annotations.monthly/lastStableBuild/artifact/annotation/diseases_to_genes.txt
MedGen	NAMES.csv.gz	ftp://ftp.ncbi.nlm.nih.gov/pub/medgen/csv/NAMES.csv.gz
HGNC	hgnc_complete_set.txt.gz	ftp://ftp.ebi.ac.uk/pub/databases/genenames/hgnc_complete_set.txt.gz
GeneReview	GRtitle_shortname_NBKid.txt	ftp://ftp.ncbi.nlm.nih.gov/pub/GeneReviews/GRtitle_shortname_NBKid.txt
Cli nVar	variant_summary.txt.gz	ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz
	gene_condition_source_id	ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/gene_condition_source_id
OMIM	morbidmap	ftp://ftp.omim.org/OMIM/morbidmap
	API	api.omim.org
Uniprot	humsavar.txt	ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/variants/humsavar.txt
Orphanet	en_product6.xml	http://www.orphadata.org/data/xml/en_product6.xml
UCSC	refFlat.txt	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/refFlat.txt.gz

3. Capture the genes and variants related to a genetic disease

Suppose we are interested in the ***retinoblastoma***

```
>keywords  = "retinoblastoma"
```

3.1 Extract the phenotypes and genes from HPO

```
> HPO.phenotype = pheno_extract_HPO(keywords)

> print(dim(HPO.phenotype))

> phenoID.hpo.omim = as.character(unique(HPO.phenotype[grep("OMIM",HPO.phenotype[,1]),1]))

> phenoID.hpo.orphanet = as.character(unique(HPO.phenotype[grep("ORPHANET",HPO.phenotype[,1]),1]))

> genes.hpo = as.character(unique(HPO.phenotype[,3]))

> genes.hpo = genes.hpo[genes.hpo!=""]
```

3.2 Extract genes from Orphanet

```
> orphanet.phenotype = extract_genes_orphanet(keyword = keywords,

                        HPO.disease = phenoID.hpo.orphanet)

> print(dim(orphanet.phenotype))

> genes.orphanet = orphanet.phenotype[,"GeneSymbol"]

> genes.merge = union(genes.orphanet, genes.hpo)
```

3.3 Extract genes and variants from OMIM

Suppose you have a OMIM API key, omim.api

```
> omim.phenotype = extract_omim(keyword= keywords,
                    omim.apiKey = omim.api,
                    HPO.disease = phenoID.hpo.omim, genelist = genes.merge)
> genes.omim = omim.phenotype [[1]]
> dim(genes.omim)
> variants.omim = omim.phenotype [[2]]
> dim(variants.omim)
```

3.4 Extract the genes and variants from ClinVar

```
>clinvar.phenotype = extract_clinvar(keyword= keywords,
                  HPO.disease = phenoID.hpo.omim, genelist = genes.merge)
> genes.clinvar = clinvar.phenotype [[1]]
> dim(genes.clinvar)
> variants.clinvar = clinvar.phenotype [[2]]
> dim(variants.clinvar)
```

3.5 Extract the genes and variants from Uniprot

```
> uniprot.phenotype = extract_uniprot(keyword= keywords,
              HPO.disease = phenoID.hpo.omim, genelist = genes.merge)
> genes.uniprot = uniprot.phenotype [[1]]
> dim(genes.uniprot)
> variants.uniprot = uniprot.phenotype [[2]]
> dim(variants.uniprot)
```

3.6 Compile the gene and variant database for a disease

Eventually, we can compile the gene-disease and variants-disease relationships as the following steps

```
> genesPDB <- genes_compile(HPO = HPO.phenotype, orphanet = orphanet.phenotype,
                  omim = genes.omim,
                  clinvar = genes.clinvar,
                  uniprot = genes.uniprot)
> variantsPDB <- variants_compile(omim = variants.omim,
                  clinvar = variants.clinvar, uniprot = variants.uniprot)
```

3.7 Capture the genes and variants from PubMed abstracts

The information of phenotypes, genes, variants, article title, PubMed ID, public years will be captured.

```
> pubmed.phenotype <- extract_pubmed(query = "retinoblastoma [Title\\/Abstract] AND (gene[Title\\/Abstract] OR genes[Title\\/Abstract] OR mutation[Title\\/Abstract] OR mutations[Title\\/Abstract] OR polymorphisms[Title\\/Abstract] OR genotype[Title\\/Abstract] OR SNP[Title\\/Abstract] OR SNPs[Title\\/Abstract] OR associated[Title\\/Abstract] OR translocation[Title\\/Abstract])", keyword=keywords)

> genes.pubmed <- pubmed.phenotype[[1]]
```

3.8 Add the additional gene-phenotype pairs

The gene-variant-phenotype relationships are filtered by the gene status: 1) have a approved symbol; 2) physical position; 3) with definite mutations on the gene. Then the phenotypes are checked manually one by one.

```
>geneAll <- genes_add_pubmed(genepdb= genesPDB, pubmed=genes.pubmed)
```

However, because the recoding of some variants are not always to follow the nomenclature by Human Genome Variation Society(HGVS), it is a very challenge to compile and integrate the variants automatically. We strongly recommend that all the genes and variants especially additional variants should be checked manually one by one.

References

[1] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct 21;431(7011):931-45. PubMed PMID: 15496913.

[2] The International HapMap Consortium. A Haplotype Map of the Human Genome. Nature. 2005 Oct 27;437(7063):1299-320. PubMed PMID: 16255080; PubMed Central PMCID: PMC1880871.

[3] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct 18;449(7164):851-61. PubMed PMID: 17943122; PubMed Central PMCID: PMC2689609.

[4] Li, J. Z.; Absher, D. M.; Tang, H.; Southwick, A. M.; Casto, A. M.; Ramachandran, S.; Cann, H. M.; Barsh, G. S. et al. (2008). "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation". Science 319 (5866): 1100–4. doi:10.1126/science.1153717. PMID 18292342.

[5] McVean, G. A.; Abecasis, D. M.; Auton, R. M.; Brooks, G. A. R.; Depristo, D. R.; Durbin, A.; Handsaker, A. G.; Kang, P.; Marth, E. E.; McVean, P.; Gabriel, S. B.; Gibbs, R. A.; Green, E. D.; Hurles, M. E.; Knoppers, B. M.; Korbel, J. O.; Lander, E. S.; Lee, C.; Lehrach, H.; Mardis, E. R.; Marth, G. T.; McVean, G. A.; Nickerson, D. A.; Schmidt, J. P.; Sherry, S. T.; Wang, J.; Wilson, R. K.; Gibbs (Principal Investigator), R. A.; Dinh, H.; Kovar, C. (2012). "An integrated map of genetic variation from 1,092 human genomes". Nature 491 (7422): 56–65. doi:10.1038/nature11632. PMC 3498066. PMID 23128226.

[6] https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp

[7] https://icgc.org/icgc

[8] Wilson BJ, Nicholls SG. The Human Genome Project, and recent advances in personalized genomics. Risk Manag Healthc Policy. 2015 Feb 16;8:9-20. doi: 10.2147/RMHP.S58728. eCollection 2015. Review. PubMed PMID: 25733939; PubMed Central PMCID: PMC4337712.

[9] Williams MS. Realizing precision medicine. Manag Care. 2013 May;22(5):42-8.PubMed PMID: 23757833.

[10] Nelson HD, Pappas M, Zakher B, Mitchell JP, Okinaka-Hu L, Fu R. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: a systematic review to update the U.S. Preventive Services Task Force recommendation. Ann Intern Med. 2014 Feb 18;160(4):255-66. Review. PubMed PMID: 24366442.