

Introducción a la Genómica

UNAL nov 2017

Alejandro Caceres
ISGlobal, Barcelona

October 14, 2017

datos de SNPs

Cada programa tiene un formato diferente y es importante saber cambiar de formato

- ▶ PLINK: Es un programa compilado, corre por la linea de comandos y es muy rapido. Es particularmente util para manejar las bases de datos en si, exlcuir sujetos, seleccionar SNPs. No tiene la versatilidad de R para explorar graficos, crear nuevas funciones o hacer graficos, pero es muy utilizado y con experiencia en computacion facil de hacer pipelines.
- ▶ snpStats (bioconductor): Tiene varias funciones para ver la estructura de los datos (linakage-disequilibrium, pca, Fst), y hace analisis de asociacion en base de datos grandes, pero no prueba diferentes modelo de herencia. Usa un fromato especial (raw data).
- ▶ snpAssoc (r-cran): versatil para probar diferentes modelos de herencia, pero las funciones no estan optimizadas para menejar matrices muy grandes.
- ▶ tabix : un programa para gestionar datos en formato VCF usado por los 1000 genomas

Es un programa por linea de comandos desarrollado por Chrostopher Chang.

The screenshot shows the PLINK 1.90 beta page. The browser's address bar shows the URL <https://www.cog-genomics.org/plink2>. The page has a navigation bar with links: PLINK 1.9 home, plink2-users, GitHub, File formats, PLINK 1.9 index, and PLINK 2.0. On the left is a sidebar with a table of contents including sections like Introduction, downloads, General usage, Standard data input, and Input filtering. The main content area is titled 'PLINK 1.90 beta' and contains a paragraph about the update, a 'Binary downloads' section with a table, and a list of unsupported flags.

PLINK 1.90 beta

This is a comprehensive update to Shaun Purcell's [PLINK](#) command-line program, developed by [Christopher Chang](#) with support from the [NIH-NIDDK's](#) Laboratory of Biological Modeling, the [Purcell Lab](#) at Mount Sinai School of Medicine, and others. ([What's new?](#)) ([Credits.](#)) ([Methods paper.](#))

Binary downloads

Operating system ¹	Build		
	Stable (beta 4.6, 15 Aug)	Development (6 Sep)	Old ² (v1.07)
Linux 64-bit	download	download	download
Linux 32-bit	download	download	download
OS X (64-bit)	download	download	download
Windows 64-bit	download	download	download
Windows 32-bit	download	download	download

¹ Solaris is no longer explicitly supported, but it should be able to run the Linux binaries.
² These are just mirrors of the binaries posted at <http://zzz.bwh.harvard.edu/plink/download.shtml>.

Source code, compilation instructions, and the like are on the [developer page](#).

The following documented PLINK 1.07 flags are not supported by 1.90 beta 4:

- `--qual-geno-scores3`
- `--segment4`
- `--dflam`
- `--tucc`

Tiene una documentación muy completa

PLINK

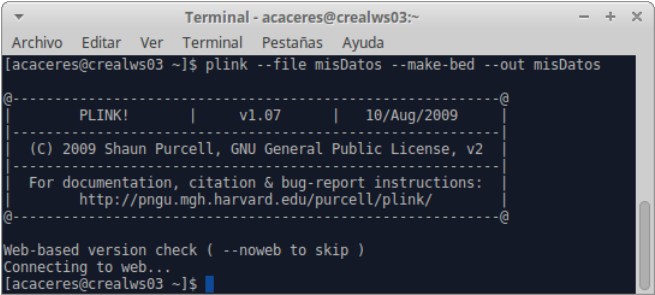
PLINK tiene dos formatos

- ▶ `.bed`, `.bim`, `.fam`: es el mas usado y separa la información en tres archivos genotipos (`.bed`), anotacion de SNPs (`.bim`), fenotipos (`.fam`)
- ▶ `.ped`, `.map`: `.ped` son los `.fam` en las primeras columnas y `.map` es una versión con menos info que `.bim`

PLINK

Para cambiar los formatos de misDatos.ped y misDatos.map a misDatos.bed, misDatos.bim y misDatos.fam

```
plink --file misDatos --make-bed --out misDatos
```



```
Terminal - acaceres@crealws03:~
Archivo  Editar  Ver    Terminal  Pestañas  Ayuda
[acaceres@crealws03 ~]$ plink --file misDatos --make-bed --out misDatos
@-----@
|          PLINK!          |          v1.07          |          10/Aug/2009          |
|-----|-----|-----|
| (C) 2009 Shaun Purcell, GNU General Public License, v2 |
|-----|-----|-----|
| For documentation, citation & bug-report instructions: |
| http://pngu.mgh.harvard.edu/purcell/plink/              |
|-----|-----|-----|
@-----@
Web-based version check ( --noweb to skip )
Connecting to web...
[acaceres@crealws03 ~]$
```

Datos de SNPs

Despues del preprocesamiento de los datos, los datos que se obtienen es de un gentipo por individuo. Si tenemos 1 millon de SNPs y 1000 individuos, esto es tipicamente una matriz de $10^3 \times 10^6$. Hay diferentes formas de organizar estos datos

	rs33	rs36	rs43	
NA090	A/C	G/G	T/A	...
NA091	A/A	G/G	T/A	...
NA092	A/A	G/C	T/A	...
NA093	A/C	C/C	A/A	...
...				

Datos de SNPs

Hay diferentes formas de organizar estos datos

	rs33	rs36	rs43	
NA090	A/C	G/G	T/A	...
NA091	A/A	G/G	T/A	...
NA092	A/A	G/C	T/A	...
NA093	A/C	C/C	A/A	...
...				

Una forma eficiente es llamar 0:homocigoto, 1:heterocigoto y 2:heterocigoto variante.

- ▶ para SNP=rs33 el alelo mas frecuente es A y el menos frecuente es C.

Entonces: A/A=0, A/C=1, CC=2

- ▶ para SNP=rs36 el alelo mas frecuente es G y el menos frecuente es C.

Entonces: G/G=0, G/C=1, CC=2

Datos típicos de SNPs (PLINK) formato bed

- Datos de los genotipos (datos.bed)

	rs33	rs36	rs43	
NA090	1	0	1	...
NA091	0	0	1	...
NA092	0	1	1	...
NA093	1	2	2	...
...				

Datos tipicos de SNPs (PLINK) formato bed

- Datos con la anotacion de SNPs (datos.bim)

chr	snp	mor	pos	allele1	allele2
1	rs33	0	1034	A	C
1	rs36	0	2000	G	C
1	rs43	0	10056	T	A
...					

Datos típicos de SNPs (PLINK) formato bed

- Datos con los fenotipos (datos.fam)

ID	FAMID	sex	asthma	BMI-z
NA090	1	1	1	1.2
NA091	1	1	0	1.5
NA092	2	0	0	0.9
NA093	2	0	1	1

SNPstats

Es un programa en R (bioconductor)

The screenshot shows the Bioconductor website for the SNPstats package. The browser window has multiple tabs open, including 'Adagio', 'El polémico', 'The Use of', 'Bioconductor', 'Genetic var', 'gwas catal', 'gwas-gr', 'Técnicas g', and 'haplotype'. The address bar shows the URL: <https://www.bioconductor.org/packages/release/bioc/html/snpStats.html>. The Bioconductor logo is at the top left, with the tagline 'OPEN SOURCE SOFTWARE FOR BIOINFORMATICS'. The navigation bar includes links for Home, Install, Help, Developers, and About. The main content area is titled 'snpStats' and shows package statistics: platforms (all), downloads (top 5%), bugs (0), in BioC (6.5 years), last (link), commits (0.3), test coverage (unknown), and DOI (10.18129/B2.2019.snpStats). Below this, it says 'Snpmatrix and XSnpmatrix classes and methods'. The Bioconductor version is Release (3.5). The description states: 'Classes and statistical methods for large SNP association studies. This extends the earlier snpMatrix package, allowing for uncertainty in genotypes.' The author is David Clayton <dc209@cam.ac.uk> and the maintainer is David Clayton <dc209@cam.ac.uk>. The citation is: 'Clayton (from within R, enter `citation("snpStats")`): Clayton D (2015). snpStats: Snpmatrix and XSnpmatrix classes and methods. R package version 1.25.0.' The installation section says: 'To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("snpStats")
```

'. The documentation section says: 'To view documentation for the version of the package installed in your system, start R and enter:

```
browseVignettes("snpStats")
```

'. On the right, there are sections for 'Documentation' (Bioconductor package vignettes and manuals, Bioconductor for learning and use, Conferences and conference material, Videos, Community resources and journals, R / CRAN packages and documentation) and 'Support' (Please read the posting guide. Post questions about Bioconductor to one of the following locations: Support lists - for questions about Bioconductor packages, Discourse mailing list - for package developers).

tiene la ventaja de que esta en ambiente R y se pueden usar otros paquetes de bioconductor

SNPstats

se instalala como desde R por medio de los comandos

```
source("https://bioconductor.org/biocLite.R")  
biocLite("snpStats")
```

SNPstats

se carga con

```
library("snpStats")  
  
## Loading required package: survival  
## Loading required package: Matrix
```

puede leer datos de PLINK (formato .bed) mediante la función

```
snp<-read.plink(misDatos)
```

se pueden guardar como binarios de R `snp.RData`

```
save(snp, file="snp.RData")
```

también se pueden guardar datos de `snpStats` en PLINK con `write.plink`

SNPstats

se pueden cargar los binarios snp.RData

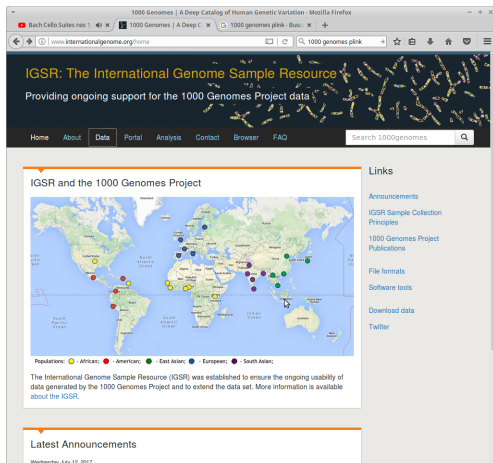
```
load("datos/snp.RData")
```

```
snp
```

```
## A SnpMatrix with 1500 rows and 439 columns  
## Row names: 1 ... 1500  
## Col names: 1 ... 439
```

1000 Genomes

Repositorio de datos de los 1000 genomas donde se pueden descargar los datos de 2504 individuos



The screenshot shows the IGSR website in a Mozilla Firefox browser. The page title is "IGSR: The International Genome Sample Resource". The main content area features a world map titled "IGSR and the 1000 Genomes Project" with colored dots representing different populations. A legend below the map identifies the populations: Africans (yellow), Americans (red), East Asians (green), Europeans (blue), and South Asians (purple). Text below the map states: "The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and to extend the data set. More information is available about the IGSR." The sidebar on the right contains a "Links" section with the following items: Announcements, IGSR Sample Collection Principles, 1000 Genomes Project Publications, File formats, Software tools, Download data, and Twitter. The "Latest Announcements" section at the bottom shows the date "Wednesday, July 12, 2017".

Hay un servidor ftp para descargar datos Los archivos son enormes, pero se pueden leer por regiones con Tabix

1000 Genomes

También hay un browser para bajar datos de regiones

Mozilla Firefox

http://phase3browser.1000genomes.org/index.html

This website has been archived. The preferred way to access 1000 Genomes data is via the [Ensembl GRCh37](#) genome browser.

1000 Genomes


A Deep Catalog of Human Genetic Variation

Tools | Help

Search 1000 Genomes

e.g. gene BRCA2 or Chromosome 6:133090746-133108748

Start Browsing 1000 Genomes data

 [Browse Human](#) -- GRCh37

[Protein variations](#) -- View the consequences of sequence variation at the level of each protein in the genome.

[Individual genotypes](#) -- Show different individual's genotype, for a variant.

Browser update October 2014

This release is based on [Ensembl 80](#) and contains the phase 3 integrated release for 2504 individuals. The data can be found on [the ftp site](#).

Please see [www.1000genomes.org](#) for more information about the data presented here and instructions for downloading the complete data set.

- [View sample data](#)


The 1000 Genomes Browser


1000 Genomes Browser based on [Ensembl v80 GRCh37](#)


As the Phase 3 1000 Genomes variants are in the process of being archived at dbSNP and DGVs, we have created a version of the Ensembl databases which contain the phase3 autosomal variants. This is presented here alongside the v80 GRCh37 Ensembl core and regulatory databases. This release represents more than 80M short variants with genotypes for 2504 individuals across 26 populations.

[Read more about this browser's features.](#)

Links

 [1000 Genomes](#) -- More information about the 1000 Genomes Project on the 1000 genomes main site.


 [Phase1 browser](#) -- This browser is based on Ensembl release 73 and represents the variant set analysed as part of [An integrated map of genetic variation from 1,092 human genomes](#), Nature 491, 56-65.

 [Tutorial](#) -- The 1000 Genomes Browser Tutorial.

The 1000 Genomes Project is an international collaborative project described at [www.1000genomes.org](#).

The 1000 Genomes Browser is based on Ensembl web code.

[Ensembl](#) is a joint project of [EMBL-EBI](#) and the [Wellcome Trust](#)

[Sanger Institute](#) 

www.1000genomes.org

1000 Genomes

obtemos datos para *MAF*

The screenshot shows the 1000 Genomes browser interface in a Mozilla Firefox browser window. The address bar shows the URL: `phase3browser:1000genomes.org/Homo_sapiens/Location/ViewHdb-core.g-ENSGG0C`. The page title is "A Deep Catalog of Human Genetic Variation".

The interface has a sidebar on the left with a "Location" menu and a "Tools" menu. The "Tools" menu is expanded, showing options like "Custom Data", "Manage Configurations", "Online Tools", and "Data Slicer". The "Data Slicer" option is selected.

The main content area is titled "Data Slicer:" and contains the following text:

When slicing a VCF or BAM file, both the data file and its index file should be present on the web server and named correctly.
The VCF file should have a ".vcf.gz" extension, and the index file should have a ".vcf.gz.tbi" extension,
E.g: MyData.vcf.gz, MyData.vcf.gz.tbi
The BAM file should have a ".bam" extension, and the index file should have a ".bam.bai" extension,
E.g: MyData.bam, MyData.bam.bai

Click [here](#) for more extensive documentation.

Below this text is the "Upload files" section, which includes a "VCF File URL:" field with the value `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr17.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz`. There is a "Clear box" link below the field.

Below the URL field is a "Region:" field with the value `17:43921017-43972966`. Below this is a "BAM options (this doesn't apply to VCF files):" section with a checkbox for "Generate .bai file *".

Below the BAM options is a note: "(please note that the generation of .bai file may take approximately 30 seconds)".

Below the note is a "VCF filters (this doesn't apply to BAM files):" section with two radio buttons: "No filtering" (selected) and "By individual".

At the bottom of the page, there is a "Drag Select" bar and a "Chromosome bands" section.

1000 Genomes

Formato VCF

1000 Genomes browser: Homo sapiens - Region in detail - Chromosome 17:43,921,017-43,972,966 - Mozilla Firefox

Bach Cello Suites nos 1 | 1000 Genomes browser: H | Haploview | Broad Institute

phase3browser: 1000genomes.org/homo_sapiens/location/View?db=core;g=ENSG00C | haploview

This website has been archived. The preferred way to access 1000 Genomes data is via the [Ensembl GRCh37](#) genome browser.

Configure Region Image | Configure Overview Image | Configure Chromosome Image | Personal Data

1000 Genomes browser: A De

Custom Data
Add your data
Attach DAG
Manage Data
Features on Karyotype

Manage Configurations
Configurations for this page
All configurations
Configuration sets

Online Tools
Variant Effect Predictor
Assembly Converter
ID History Converter
Data Slicer
Variation Pattern Finder
VCF to PED converter
Forge Analysis (v1.0)
Allele Frequency

Location: 17:43921017-43972966
Gene:

Thank you - your VCF file
[17:43921017-43972966.ALL.chr17.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz](#) [Size: 367480] has been generated.
Right click on the file name and choose "Save link as ..." from the menu

Preview

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	H000096
17	43921068		rs555347111	C	T	100	PASS	AC=2;AF=	
17	43921106		rs573543994	C	G	100	PASS	AC=1;AF=	
17	43921126		rs542617372	C	A	100	PASS	AC=1;AF=	
17	43921130		rs562398147	C	T	100	PASS	AC=1;AF=	
17	43921131		rs576107214	G	A	100	PASS	AC=1;AF=	

Location: 17:43921017-43972966 Go

Gene: Go

Formato VCF



Paquete Variant Annotation para leer datos VCF

The screenshot shows the Bioconductor website for the VariantAnnotation package. The browser window is titled "Bioconductor - VariantAnnotation - Mozilla Firefox". The address bar shows the URL "bioconductor.org/packages/release/bioc/html/VariantAnnotation.html". The page has a teal header with the Bioconductor logo and navigation links: Home, Install, Help, Developers, and About. Below the header, the breadcrumb trail is "Home > Bioconductor 3.5 > Software Packages > VariantAnnotation". The main heading is "VariantAnnotation" in green. Below this, there are statistics: "platforms: all", "downloads: top 5%", "posts: 15 / 17274", "in Bioc: 6 years", "build: ok", "current: 2.67", and "test coverage: 73%". The DOI is "10.18129/B3.bioc.VariantAnnotation". The section "Annotation of Genetic Variants" describes the package's purpose: "Bioconductor version: Release (3.5). Annotate variants, compute amino acid coding changes, predict coding outcomes." It lists the author as Valerie Obenchain [aut, cre], the maintainer as Valerie Obenchain, and provides a citation. The "Installation" section shows the R command to install the package. The "Documentation" section shows the R command to browse vignettes. On the right side, there are sections for "Documentation" and "Support".

Bioconductor - VariantAnnotation - Mozilla Firefox

bioconductor.org/packages/release/bioc/html/VariantAnnotation.html

Search:

Home Install Help Developers About

Home > Bioconductor 3.5 > Software Packages > VariantAnnotation

VariantAnnotation

platforms: all downloads: top 5% posts: 15 / 17274 in Bioc: 6 years
build: ok current: 2.67 test coverage: 73%

DOI: 10.18129/B3.bioc.VariantAnnotation

Annotation of Genetic Variants

Bioconductor version: Release (3.5)
Annotate variants, compute amino acid coding changes, predict coding outcomes.
Author: Valerie Obenchain [aut, cre], Martin Morgan [aut], Michael Lawrence [aut], Stephanie Gogarten [ctb]
Maintainer: Valerie Obenchain <maintainer at bioconductor.org>
Citation (from within R, enter `citation("VariantAnnotation")`):
Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P and Morgan M (2014). "VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants." *Bioinformatics*, 30(14), pp. 2076-2078. doi: 10.1093/bioinformatics/btu146

Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("VariantAnnotation")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("VariantAnnotation")
```

Documentation

Bioconductor

- Package vignettes and manuals.
- Webinars for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

Support

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [Bioconductor](#) mailing list - for package developers

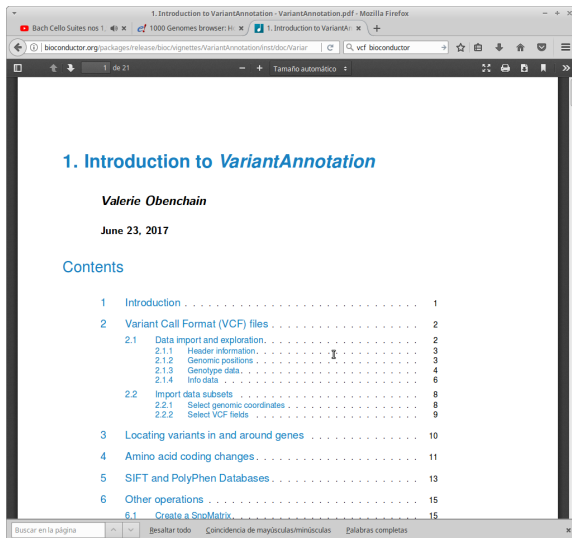
VCF in R

se pueden cargar los binarios snp.RData

```
source("http://bioconductor.org/biocLite.R")  
biocLite("VariantAnnotation")
```

Bioconductor

Variant annotation



The screenshot shows a Mozilla Firefox browser window with the address bar displaying the URL: `bioconductor.org/packages/release/bioc/vignettes/VariantAnnotation/inst/doc/VariantAnnotation.pdf`. The search bar contains the text "vcf bioconductor". The main content area displays the title "1. Introduction to VariantAnnotation" in blue, followed by the author "Valerie Obenchain" and the date "June 23, 2017". Below this is a "Contents" section with a table of contents listing chapters and page numbers. The table of contents includes: 1 Introduction (1), 2 Variant Call Format (VCF) files (2), 2.1 Data import and exploration (2), 2.1.1 Header information (3), 2.1.2 Genomic positions (3), 2.1.3 Genotype data (4), 2.1.4 Info data (6), 2.2 Import data subsets (8), 2.2.1 Select genomic coordinates (8), 2.2.2 Select VCF fields (9), 3 Locating variants in and around genes (10), 4 Amino acid coding changes (11), 5 SIFT and PolyPhen Databases (13), 6 Other operations (15), and 6.1 Create a SnpMatrix (15). The browser's status bar at the bottom shows search options: "Buscar en la página", "Resaltar todo", "Coincidencia de mayúsculas/minúsculas", and "Palabras completas".

1. Introduction to *VariantAnnotation*

Valerie Obenchain

June 23, 2017

Contents

1	Introduction	1
2	Variant Call Format (VCF) files	2
2.1	Data import and exploration	2
2.1.1	Header information	3
2.1.2	Genomic positions	3
2.1.3	Genotype data	4
2.1.4	Info data	6
2.2	Import data subsets	8
2.2.1	Select genomic coordinates	8
2.2.2	Select VCF fields	9
3	Locating variants in and around genes	10
4	Amino acid coding changes	11
5	SIFT and PolyPhen Databases	13
6	Other operations	15
6.1	Create a SnpMatrix	15

VCF in R

```
library(VariantAnnotation)
fl<-"datos/17.43921017-43972966.ALL.chr17.phase3_shapeit2_mvncall_integ
vcf <- readVcf(fl, "hg19")
vcf

## class: CollapsedVCF
## dim: 1863 2504
## rowRanges(vcf):
##   GRanges with 5 metadata columns: paramRangeID, REF, ALT, QUAL, FILTER
## info(vcf):
##   DataFrame with 27 columns: CIEND, CIPOS, CS, END, IMPRECISE, MC, MEND, MLEN
## info(header(vcf)):
##           Number Type      Description
##   CIEND      2      Integer Confidence interval around END for i
##   CIPOS      2      Integer Confidence interval around POS for i
##   CS         1      String   Source call set.
##   END        1      Integer End coordinate of this variant
##   IMPRECISE   0      Flag     Imprecise structural variation
##   MC         .      String   Merged calls.
##   MEINFO     4      String   Mobile element info of the form NAME
##   MEND       1      Integer Mitochondrial end coordinate of inse
##   MLEN       1      Integer Estimated length of mitochondrial i
```

VCF in R

```
genos<-geno(vcf)
```

```
names(genos)
```

```
## [1] "GT"
```

```
dim(genos$GT)
```

```
## [1] 1863 2504
```

```
genos$GT[1:5,1:5]
```

```
##           HG00096 HG00097 HG00099 HG00100 HG00101
## rs555347111 "0|0"  "0|0"  "0|0"  "0|0"  "0|0"
## rs573543994 "0|0"  "0|0"  "0|0"  "0|0"  "0|0"
## rs542617372 "0|0"  "0|0"  "0|0"  "0|0"  "0|0"
## rs562398147 "0|0"  "0|0"  "0|0"  "0|0"  "0|0"
## rs576107214 "0|0"  "0|0"  "0|0"  "0|0"  "0|0"
```

Si el archivo es grande readVcf permite leer sólo regiones de interes

VCF in R

Los genotipos en formato 0,1,2 pueden ser encontrados en *genos\$DS*.

Si no se puede entonces se puede calcular asi

```
snps<-genos$GT
snps[snps=="0|0"]<-0
snps[snps=="1|1"]<-2
snps[snps!=0 & snps !=2]<-1
snps[1:5,1:5]
```

##		HG00096	HG00097	HG00099	HG00100	HG00101
##	rs555347111	"0"	"0"	"0"	"0"	"0"
##	rs573543994	"0"	"0"	"0"	"0"	"0"
##	rs542617372	"0"	"0"	"0"	"0"	"0"
##	rs562398147	"0"	"0"	"0"	"0"	"0"
##	rs576107214	"0"	"0"	"0"	"0"	"0"

```
save(snps, file="snpsMAPT.RData")
```

VCF in snpStats

snpStats usa formato 1,2,3 para genotipos y el 0 para missing

```
library(snpStats)
snpsnew<-t(snps)
snpsnew[snps=="0"] <- 1
snpsnew[snps=="1"] <- 2
snpsnew[snps=="2"] <- 3

snpsSNPstats <- new("SnpMatrix", snpsnew)

## coercing object of mode character to SnpMatrix

print(as(snpsSNPstats[1:5,1:5], 'character'))

##          rs555347111 rs573543994 rs542617372 rs562398147 rs576107214
## HG00096 "A/A"         "A/A"         "A/A"         "A/A"         "A/A"
## HG00097 "A/A"         "A/A"         "A/A"         "A/A"         "A/A"
## HG00099 "A/A"         "A/A"         "A/A"         "A/A"         "A/A"
## HG00100 "A/A"         "A/A"         "A/A"         "A/A"         "A/A"
## HG00101 "A/A"         "A/A"         "A/A"         "A/A"         "A/A"

save(snpsSNPstats, file="snpsSNPstats.RData")
```

1000 Genomes

Los datos de los 1000 genomas (y HapMap) también están en formato PLINK por cromosomas

Resources - PLINK 1.9 - Mozilla Firefox

Chopin - Complete Noci | Resources - PLINK 1.9

https://www.cog-genomics.org/plink/1.9/resources

PLINK 1000 genomes

Limitations
Note to testers
[Jump to search box]
Citation instructions

General usage

Standard data input
PLINK 1 binary (.bed)
Autocorrelation behavior
PLINK test (.pos, .bed...) VCF (.vcf.gz), .bed
Oxford (.gen(.gz), .igen)
23andMe test
Generate random
Unusual chromosome IDs
Recombination map
Phenotypes
Covariates
Clusters of samples
Variant sets
Binary distance matrix
IBD report (.genome)

Input filtering

Sample ID file
Variant ID file
Cluster membership
Set membership
Attribute-based
Chromosomes
SNPs only
Simple variant window
Multiple variant ranges
Sample/variant thinning
Covariates (-filter)
Missing genotypes
Missing phenotypes
Minor allele frequencies
Hardy-Weinberg
Mendel errors
Quality scores
Relationships

Main functions

Data management
--make-bed
--recode
--output chr
--zero-cluster
--split a--merge a
--set-missing
--fix-missing a2
--set-missing-var-ids
--update-map...
--update-ids...
--r2
--p-ican
--use-allele-freqs

Genotype data

1000 Genomes phase 1 (hosted by [GigaDB](#), Aspera download available there)

- Entire dataset as a single .tar.gz (1.11 GB)
- Split by chromosome:
 - chr1 (93.1 MB)
 - chr2 (102 MB)
 - chr3 (83.8 MB)
 - chr4 (84.2 MB)
 - chr5 (77.2 MB)
 - chr6 (76.1 MB)
 - chr7 (70.9 MB)
 - chr8 (66.7 MB)
 - chr9 (53.5 MB)
 - chr10 (59.6 MB)
 - chr11 (59.9 MB)
 - chr12 (57.6 MB)
 - chr13 (43.1 MB)
 - chr14 (39.9 MB)
 - chr15 (37.4 MB)
 - chr16 (40.0 MB)
 - chr17 (35.0 MB)
 - chr18 (34.9 MB)
 - chr19 (28.9 MB)
 - chr20 (27.4 MB)
 - chr21 (17.2 MB)
 - chr22 (17.4 MB)
 - chrX, not including pseudoautosomal region (38.6 MB)
 - chrY (802 KB)
 - Pseudoautosomal region (2.7 MB)
 - chrMT (45.8 KB)

Refer to the 1000 Genomes website for [additional sample information](#), [data usage rules](#), and [citation instructions](#).

HapMap phase 2

See the [PLINK 1.07 resources page](#).

Teaching materials and example dataset

ds

Buscar todo Coincidencia de mayúsculas/minúsculas Palabras completas 2 de 11 aciertos

PLINK a VCF

- ▶ los comandos PLINK pueden usar formato VCF
- ▶ también se puede convertir .bed .bim .fam a formato a VCF y vice-versa

```
$ plink --bfile [filename prefix] --recode vcf --out [VCF prefix]
```

```
$ plink --vcf [VCF filename] --out [.bed/.bim/.fam prefix]
```

Ejercicio

- ▶ Descargar datos de los 1000 Genomas en PLINK
- ▶ leerlos en snpStats
- ▶ si PLINK está instalado convertirlos en VCF
- ▶ leerlos en snpStats