

# Introducción a la Genómica

## UNAL nov 2017

Alejandro Cáceres  
ISGlobal, Barcelona

November 8, 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 línea de comandos y es muy rápido. Es particularmente útil para manejar las bases de datos en sí, excluir sujetos, seleccionar SNPs. No tiene la versatilidad de R para explorar gráficos, crear nuevas funciones o hacer gráficos, pero es muy utilizado y con experiencia en computación fácil de hacer pipelines.
- ▶ snpStats (bioconductor): Tiene varias funciones para ver la estructura de los datos (linkage-disequilibrium, pca, Fst), y hace análisis de asociación en base de datos grandes, pero no prueba diferentes modelos de herencia. Usa un formato especial (raw data).
- ▶ snpAssoc (r-cran): versátil para probar diferentes modelos de herencia, pero las funciones no están optimizadas para manejar 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 1.9 website. The browser title is "PLINK 1.9 - Mozilla Firefox". The address bar shows "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".

The main content area is titled "PLINK 1.90 beta". It contains the following text:

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.](#))

Below this is a section for "Binary downloads" with a table:

Operating system <sup>1</sup>	Build		
	Stable (beta 4.6, 15 Aug)	Development (6 Sep)	Old <sup>2</sup> (v1.07)
Linux 64-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
Linux 32-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
OS X (64-bit)	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
Windows 64-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>
Windows 32-bit	<a href="#">download</a>	<a href="#">download</a>	<a href="#">download</a>

Footnotes:

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-scores`<sup>3</sup>
- `--segment`<sup>4</sup>
- `--dfam`
- `--tucc`

A sidebar on the left contains a list of links: Introduction, downloads, Recent version history, What's new?, Future development, Limitations, Note to testers, (Jump to search box), General usage, Citation instructions, Standard data input, PLINK 1 binary (.bed), Autosconversion behavior, PLINK text (.ped, .tped...), VCF (.vcf.gz), .bct, Oxford (.genl.gz), .bgem, 23andMe text, 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, and Stat.manhattan.txt.

Tiene una documentación muy completa

# PLINK

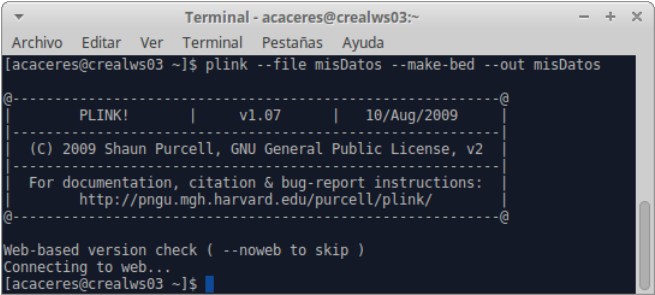
PLINK tiene dos formatos

- ▶ `.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

Después del preprocesamiento de los datos, los datos que se obtienen es de un genotipo por individuo. Si tenemos 1 millón de SNPs y 1000 individuos, esto es típicamente 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

	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 típicos 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

# PLINK

en <https://www.cog-genomics.org/plink/1.9/resources> hay datos de prueba para aprender a usar PLINK

Si PLINK esta instalado

- ▶ bajar `1kg_phase1_chr22.tar.gz`
- ▶ descomprimir
- ▶ ejecutar

```
plink --bfile 1kg_phase1_chr22 --make-bed --chr 22  
--out mydata --to-mb 20
```

Esto selecciona unos datos del cromosoma 22 hasta 20 Mb y los guarda en `mydata.bed`, `mydata.fam` y `mydata.bim`

# SNPstats

Es un programa en R (Bioconductor)

The screenshot shows the Bioconductor website for the **snpStats** package. The browser window is titled "Bioconductor - snpStats - Mozilla Firefox". The address bar shows the URL: <https://www.bioconductor.org/packages/release/bioc/html/snpStats.html>. The page features a teal header with the Bioconductor logo and navigation links: Home, Install, Help, Developers, and About. Below the header, the package name **snpStats** is displayed. A summary bar shows statistics: platforms (all), downloads (top 5%), bugs (0), in BioC (6.5 years), build (all), commits (0.55), and test coverage (unknown). The DOI is 10.18129/BIC.BIC.snpStats. The main content area is titled "SnpMatrix and XSnpmatrix classes and methods". It describes the package as "Bioconductor version: Release (3.5)" and "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 <dc208@cam.ac.uk> and the maintainer is David Clayton <dc208@cam.ac.uk>. A citation is provided: Clayton D (2015). snpStats: SnpMatrix and XSnpmatrix classes and methods. R package version 1.25.0. The installation section instructs to start R and enter: 

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

 The documentation section instructs to view documentation for the version installed in your system, start R and enter: 

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

 On the right side, there are sections for "Documentation" and "Support". The "Documentation" section lists links for Package vignettes and manuals, WebPages for learning and use, Course and conference material, and Videos. The "Support" section asks users to read the posting guide and post questions about Bioconductor to one of the following locations: Support site, Bioconductor packages, and Bioconductor mailing list for package developers.

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

# SNPstats

se instalala como desde R por medio de los comandos

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

Todos los paquetes de Bioconductor tienen manuales de usuarios (vineta)

# SNPstats

la librería se carga con

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

cargemos los datos

```
snp<-read.plink("datos/mydata")  
  
## Warning: non-unique value when setting 'row.names': '.'  
## Error in 'row.names<-.data.frame'('*tmp*', value = value):  
duplicate 'row.names' are not allowed
```

Un error típico de cuando algunos SNPs no están anotados

# SNPstats

Veamos cuales son los SNPs que estan anotados

```
bim<-read.table("datos/mydata.bim",as.is=TRUE)

rs<-bim[,2]
tb<-table(rs)
dup<-tb[tb>1]
selrs<-rs[!rs%in%names(dup)]

head(rs)

## [1] "rs149201999" "rs146752890" "rs139377059" "rs188945759" "rs65183
## [6] "rs62224609"
```

# SNPstats

Lamos los SNPs anotados usando la opción `select.snps`

```
snp<-read.plink("datos/mydata", select.snps = selrs)  
names(snp)
```

```
## [1] "genotypes" "fam"          "map"
```



# SNPstats

```
snp$genotypes
```

```
## A SnpMatrix with 1092 rows and 43067 columns  
## Row names: HG00096 ... NA20828  
## Col names: rs149201999 ... rs145875228
```

```
head(snp$fam)
```

```
##          pedigree  member father mother sex affected  
## HG00096         NA HG00096   <NA>   <NA>   1         NA  
## HG00097         NA HG00097   <NA>   <NA>   2         NA  
## HG00099         NA HG00099   <NA>   <NA>   2         NA  
## HG00100         NA HG00100   <NA>   <NA>   2         NA  
## HG00101         NA HG00101   <NA>   <NA>   1         NA  
## HG00102         NA HG00102   <NA>   <NA>   2         NA
```

```
head(snp$map)
```

```
##          chromosome      snp.name cM position allele.1 allele.2  
## rs149201999         22 rs149201999 NA 16050408         C         T  
## rs146752890         22 rs146752890 NA 16050612         G
```

# SNPstats

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

```
save(snp, file="datos/mydata.RData")
```

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

# SNPstats

Estos son solo genotipos guardados en binario `snp.RData`

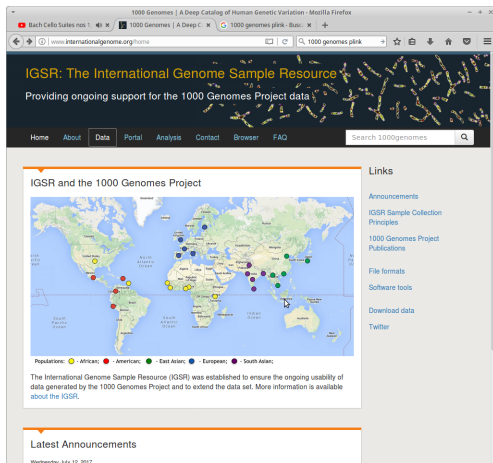
```
load("datos/snp.RData")
```

```
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 homepage of the International Genome Sample Resource (IGSR). The header includes the title "IGSR: The International Genome Sample Resource" and the subtitle "Providing ongoing support for the 1000 Genomes Project data". Below the header is a navigation bar with links: Home, About, Data, Portal, Analysis, Contact, Browser, and FAQ. A search bar is also present. The main content area features a world map titled "IGSR and the 1000 Genomes Project" showing the distribution of the 2504 individuals. The map is color-coded by population group: African (yellow), American (red), East Asian (green), European (blue), and South Asian (purple). A legend below the map identifies these groups. To the right of the map is a sidebar with a "Links" section containing links to Announcements, IGSR Sample Collection Principles, 1000 Genomes Project Publications, File formats, Software tools, Download data, and Twitter. Below the map, there is a section for "Latest Announcements" with a date of Wednesday, July 12, 2017.

IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

Home About Data Portal Analysis Contact Browser FAQ

Search 1000genomes

IGSR and the 1000 Genomes Project

Populations: ● - African; ● - American; ● - East Asian; ● - European; ● - South Asian;

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.

Latest Announcements

Wednesday, July 12, 2017

Links

- Announcements
- IGSR Sample Collection Principles
- 1000 Genomes Project Publications
- File formats
- Software tools
- Download data
- Twitter

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

obtenemos 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 main content area is titled "Configure Region Image" and "Configure Overview Image". It contains a "Data Slicer" section with the following text:

**Data Slicer:**

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.

**Upload files**

**VCF File URL:** `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr17.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz`

[Clear box](#)

e.g. `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr17.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz`

**Region:** `17:43921017-43972966`

e.g. `1:1-50000`

**BAM options (this doesn't apply to VCF files):** ☐ Generate .bai file \*

\*(please note that the generation of .bai file may take approximately 30 seconds)

**VCF filters (this doesn't apply to BAM files):** ☐ No filtering ☐ By individual

The bottom of the page shows 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=ENSG0C 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

# 1000 Genomes

## Formato VCF

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

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

phase3browser: 1000genomes.org/homo\_sapiens/location/View?db=core;g=ENSG0C 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

2010 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

Thank you - your VCF file  
[17:43921017-43972966.ALL.chr17\\_phase3\\_shapeit2\\_mvncall\\_integrated\\_v5a.20130502.genotypes.vcf.gz](#) [Size: 3674602] has been generated.  
Right click on the file name and choose "Save link as ..." from the menu

Preview

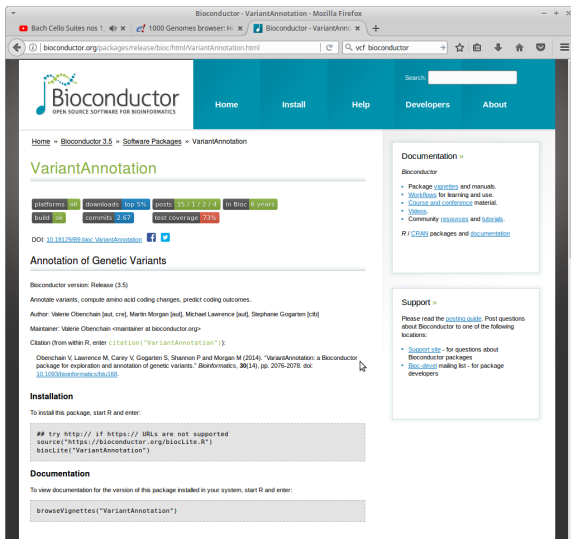
HG00119	HG00120	HG00121	HG00122	HG00123	HG00125	HG00126	HG00127	HG00128	HG00129
0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0

Location: 17:43921017-43972966 Go

Gene: Go



## 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. A search bar is also present.

The main content area is titled "VariantAnnotation" and includes a breadcrumb trail: Home > Bioconductor 3.5 > Software Packages > VariantAnnotation. Below the title, there are statistics: platforms: all, downloads: top 5%, posts: 15 / 17274, in Bioc: 6 years, build: ok, commits: 2/67, and test coverage: 73%.

The "Annotation of Genetic Variants" section 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 for the package.

The "Installation" section provides instructions for installing the package in R, including a code block for trying the package from a remote source.

The "Documentation" section provides instructions for viewing the documentation for the installed package, including a code block for browsing vignettes.

On the right side, there are two sections: "Documentation" and "Support". The "Documentation" section lists links to package vignettes, manuals, and other resources. The "Support" section provides information on how to get help, including a link to the support site and a link to the mailing list.

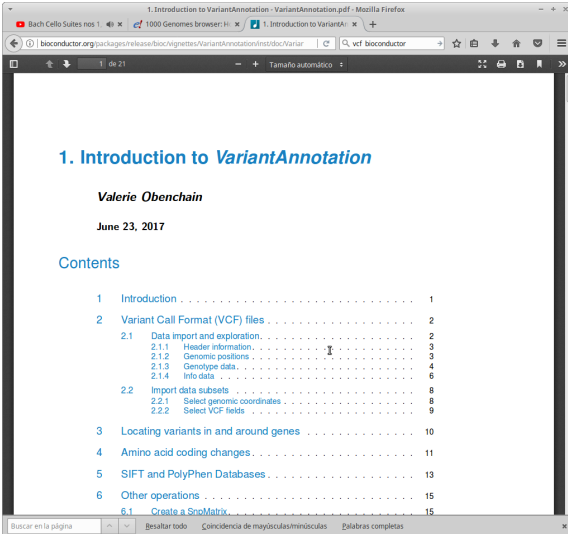
# VCF in R

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]  
General usage  
Citation instructions

**Standard data input**  
PLINK 1 binary (.bed)  
Autocorrelation behavior  
PLINK text (.ped, .bed...) VCF (.vcf.gz), .bcl  
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 s  
--set-missing  
--fit missing a2  
--set-missing var-ids  
--update map...  
--update ids...  
--rpl  
--rpl-ican

**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 (78.1 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 formato PLINK
- ▶ leerlos en snpStats
- ▶ si PLINK está instalado convertirlos en VCF
- ▶ leerlos en R