R como herramienta para el análisis de datos genómicos

Pilar Cacheiro

Grupo de Medicina Xenómica
Universidade de Santiago de Compostela

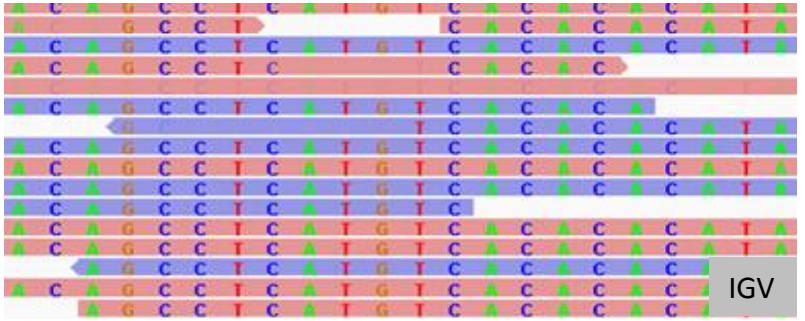
¿Qué hacemos?

- Genética de enfermedades raras y complejas
- Farmacogenómica
- Genética poblacional y evolutiva

……

¿Cómo lo hacemos?

Estudiando la secuencia de ADN de individuos con una determinada patología o fenotipo e identificando variaciones respecto a una secuencia de referencia y/o población control



Investigadora en Grupo de Medina Xenómica USC

BSc Biología

MSc Estadística

Estudiante doctorado Estadística

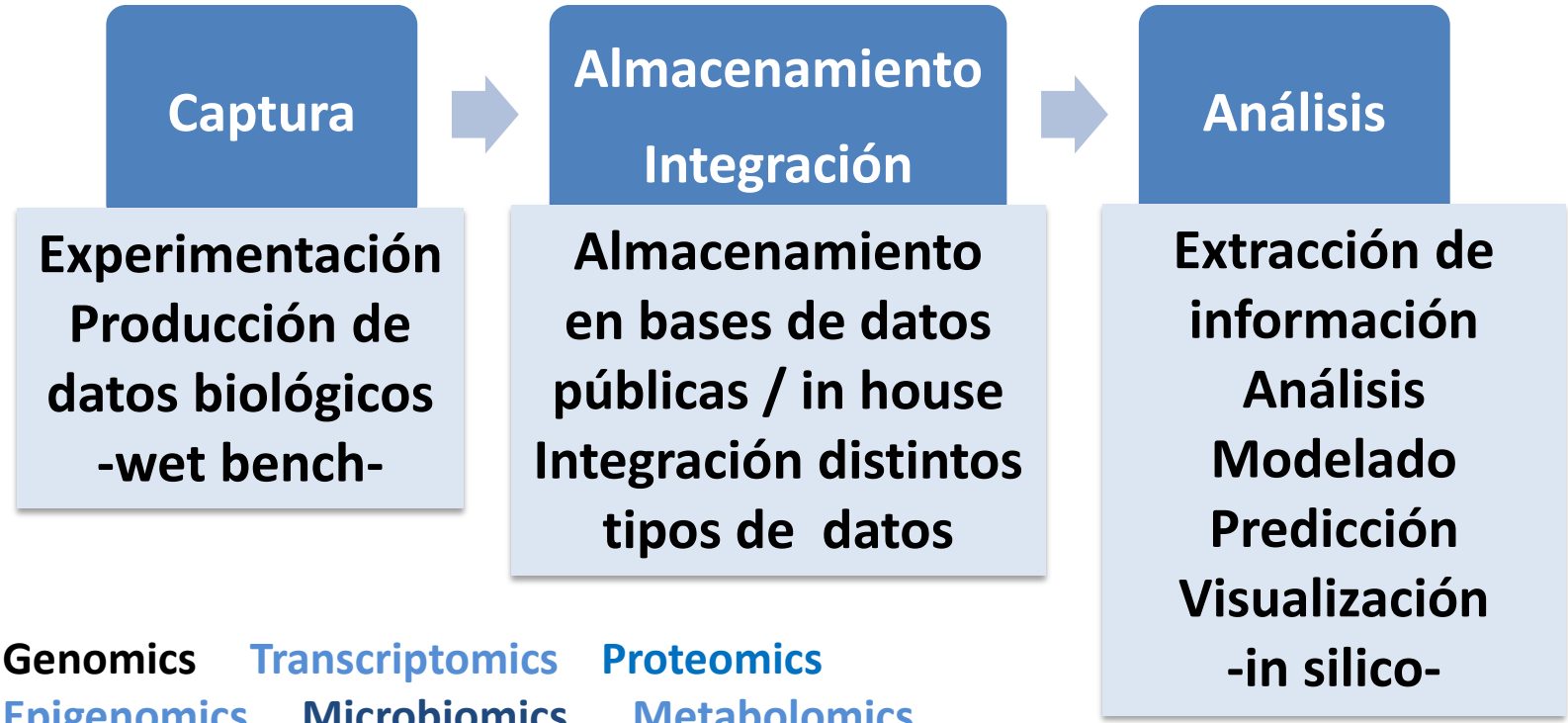
Áreas de investigación

- ≥ Métodos de selección de variables
- ≥ Técnicas de clasificación
- ≥ Análisis network

Perspectiva >>>> **Genómica computacional**

Biología computacional
Bioinformática
Estadística genómica

Aplicación de métodos computacionales y estadísticos para extraer conocimiento biológico de secuencias genómicas



Genomics **Transcriptomics** **Proteomics**
Epigenomics **Microbiomics** **Metabolomics**
Pharmacogenomics **Metagenomics** **Toxigenomics**

R/Bioconductor

- Proyecto de software libre que proporciona herramientas para el análisis de datos genómicos
- Basado principalmente en R
- Bioconductor 3.4 (Octubre, 2016)
  - 1294 paquetes software
  - 309 paquetes datos experimentos
  - 933 paquetes anotaciones
- Funcionalidades:
  - Métodos estadísticos y gráficos para el análisis de datos genómicos con formatos particulares (BAM, VCF, BED)
  - Paquetes muy documentados con vignettes y workflows para facilitar reproducibilidad (para arrays específicos)
  - Anotaciones: conexión a bases de datos para facilitar la inclusión de metadatos en el análisis de datos genómicos (para diferentes especies)

DNA-seq: detección de variantes : SNVs, indels, CNVs.
RNA-seq: estudios de expresión diferencial
ChIP-seq: identificación de sitios de unión de factores de transcripción

M Morgan. Sequences, Genomes, and Genes in R / Bioconductor

Análisis de datos genómicos en R

Manipulación Integración datos

Formatos específicos de datos genómicos

Data frames de grandes dimensiones y gran número de data frames

Integración de diferentes datasets

Anotación

Bases de datos de anotaciones biológicas

Cadenas de caracteres

Análisis

Estudios de asociación

Predicción: risk scores

Enrichment analysis

Específicos para datos genéticos: haplotipos/datación

……

Paquetes R. Mis imprescindibles

- Anotaciones
  - **biomaRt** (R/Bioconductor)
  - **AnnotationDbi** (R/Bioconductor) S Durinck
    - **org.Hs.eg.db** H Pages , Carlson ,S Falcon and N Li. AnnotationDbi: Annotation Database Interface
    - **GO.db** M Carlson .GO.db: A set of annotation maps describing the entire Gene Ontology
    - **KEGG.db** M Carlson. KEGG.db: A set of annotation maps for KEGG
    - **Category** MRGentleman with contributions from S Falcon and D Sarkar

- Manejo de datos
  - **data.table** M Dowle, A Srinivasan, T Short, S Lianoglou with contributions from R Saporta and E Antonyan
  - **dplyr** H Wickham, R Francois
  - **plyr** H Wickham
  - **reshape/ reshape2** H Wickham
  - **stringr** H Wickham
  - **splitstackshape** A Mahto

Datos genómicos

- ❑ Diferentes tipo de datos: **DNA**, RNA, Metilación, ChIP,..
- ❑ Diferentes tecnologías:

Array

Se buscan cambios en la secuencia de ADN en posiciones específicas

Next Generation Sequencing

Se buscan cambios en la secuencia de ADN a lo largo de todas las posiciones (genoma/ exoma / genes)

- Elección de tecnología depende de diversos factores: objetivo del estudio limitaciones de coste, tiempo, preparado de muestras, capacidad para analizar los datos…
- Diferencias en el formato de los datos de salida

  Array: filas (muestras) x columnas (marcadores)(ej ~ 50000 x 5000000)

  NGS: filas (variantes) x columnas (anotaciones) (ej ~50000 x 100) ***

  *** 1 archivo por muestra

Arrays

id	rs10399749	rs11260616	rs4648633	rs6659552	rs7550396	rs12239794	rs6688969	rs10753357	rs1495243
Sample001	CC	AA	TT	GG	GG	GG	CC	AC	GG
Sample002	CC	AT	CT	CG	GG	GG	CT	AA	AG
Sample003	CC	AA	TT	CG	GG	GG	CT	AA	AA
Sample004	CC	AT	TT	GG	GG	<NA>	CC	AC	GG
Sample005	CC	AA	CT	GG	GG	GG	CT	CC	GG
Sample006	CC	AA	TT	CG	GG	GG	CC	CC	AG
Sample007	CC	AA	TT	CG	GG	GG	CT	CC	AG
Sample008	CC	AA	CT	CG	GG	GG	CT	CC	AG
Sample009	CC	AT	CT	CG	GG	GG	CC	CC	AA
Sample010	CC	AT	<NA>	GG	GG	GG	CC	AC	AG
Sample011	CC	AA	CT	CG	GG	GG	CC	CC	AG
Sample012	CC	AT	CT	CG	GG	GG	CC	CC	GG
Sample013	CC	AT	TT	CG	GG	GG	TT	CC	AA
Sample014	CC	AT	TT	GG	GG	GG	CT	CC	GG
Sample015	CC	TT	CT	CG	GG	GG	CT	AC	GG
Sample016	CC	AA	CT	CG	GG	GG	CT	CC	AG
Sample017	CC	AT	TT	CG	GG	<NA>	TT	AC	AG
Sample018	CC	AA	CT	CG	GG	GG	CT	CC	AG

miles filas x millones columnas (marcadores genotipados + imputados)

Dataset modificado a partir del ejemplo incluido en el paquete de R **snpAssoc** -Juan R González, Lluís Armengol, Elisabet Guinó, Xavier Solé and Víctor Moreno (2014). SNPassoc: SNPs-based whole genome association studies. R package version 1.9-2 - a partir de datos de HapMap http://www.hapmap.org

Arrays

sample	rs7909677	rs7093061	rs12773042	rs7475011	rs11253563	rs4881551	rs4881552	rs4880750	rs10904596
Sample001	0	0	2	1	2	1	1	2	0
Sample002	0	0	2	2	2	2	0	2	0
Sample003	0	0	2	2	2	2	0	2	0
Sample004	0	1	2	1	1	2	1	2	1
Sample005	0	1	2	0	1	1	2	1	1
Sample006	0	0	2	0	2	1	1	1	0
Sample007	0	0	2	1	2	1	1	1	0
Sample008	1	0	1	1	2	1	1	2	0
Sample009	0	1	2	1	2	1	1	1	0
Sample010	0	1	2	0	1	1	2	1	1

Dataset modificado a partir del ejemplo incluido en el paquete de R/Bioconductor **snpStats** - David Clayton (2015). snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.22.0 - a partir de bases de datos públicas:
http://www.hapmap.org
http://www.affymetrix.com/support/technical/sample_data/500k_hapmap_genotype_data.affx

- Archivos auxiliares
  - Integración datos clínicos (info muestra)
  - Integración información SNPs (info variante)

rs	chr	position	A1	A2
rs7909677	10	101955	A	G
rs7093061	10	112109	C	T
rs12773042	10	117636	C	G

ID	status	sexo	var_1	var_2
Sample001	control	h	1524	5
Sample002	control	m	962	4
Sample003	control	h	450	3
Sample004	caso	h		1

nature genetics

LETTERS

Array

~8 millones variantes

13708 casos

95282 controles

2014

Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease

Mike A Nalls[1,46], Nathan Pankratz[2,46], Christina M Lill[3,4], Chuong B Do[5], Dena G Hernandez[1,6], Mohamad Saad[7–9], Anita L DeStefano[10–12], Eleanna Kara[13], Jose Bras[13], Manu Sharma[14,15], Claudia Schulte[15], Margaux F Keller[1], Sampath Arepalli[1], Christopher Letson[1], Connor Edsall[1], Hreinn Stefansson[16], Xinmin Liu[17], Hannah Pliner[1], Joseph H Lee[18], Rong Cheng[18], International Parkinson's Disease Genomics Consortium (IPDGC)[19], Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI)[19], 23andMe[19], GenePD[19], NeuroGenetics Research Consortium (NGRC)[19], Hussman Institute of Human Genomics (HIHG)[19], The Ashkenazi Jewish Dataset Investigator[19], Cohorts for Health and Aging Research in Genetic Epidemiology (CHARGE)[19], North American Brain Expression Consortium (NABEC)[19], United Kingdom Brain Expression Consortium (UKBEC)[19], Greek Parkinson's Disease Consortium[19], Alzheimer Genetic Analysis Group[19], M Arfan Ikram[20–22], John P A Ioannidis[23], Georgios M Hadjigeorgiou[24], Joshua C Bis[25], Maria Martinez[8,9], Joel S Perlmutter[26–28], Alison Goate[26,28–30], Karen Marder[18,31–33], Brian Fiske[34], Margaret Sutherland[35], Georgia Xiromerisiou[24,36], Richard H Myers[10], Lorraine N Clark[17,18], Kari Stefansson[16], John A Hardy[6], Peter Heutink[37], Honglei Chen[38], Nicholas W Wood[13], Henry Houlden[13], Haydeh Payami[39], Alexis Brice[40–42], William K Scott[43], Thomas Gasser[15], Lars Bertram[3,44], Nicholas Eriksson[5], Tatiana Foroud[45] & Andrew B Singleton[1]

We conducted a meta-analysis of Parkinson's disease genome-wide association studies using a common set of 7,893,274 variants across 13,708 cases and 95,282 controls.

both confirmed the central Parkinson's disease and impl cascade[18]. These data have a

ARTICLE

doi:10.1038/nature18642

Array + secuenciación

26.7 millones de variantes

90000 individuos

2016

The genetic architecture of type 2 diabetes

A list of authors and affiliations appears in the online version of the paper

The genetic architecture of common traits, including the number, frequency, and effect sizes of inherited variants that contribute to individual risk, has been long debated. Genome-wide association studies have identified scores of common variants associated with type 2 diabetes, but in aggregate, these explain only a fraction of the heritability of this disease. Here, to test the hypothesis that lower-frequency variants explain much of the remainder, the GoT2D and T2D-GENES consortia performed whole-genome sequencing in 2,657 European individuals with and without diabetes, and exome sequencing in 12,940 individuals from five ancestry groups. To increase statistical power, we expanded the sample size via genotyping and imputation in a further 111,548 subjects. Variants associated with type 2 diabetes after sequencing were overwhelmingly common and most fell within regions previously identified by genome-wide association studies. Comprehensive enumeration of sequence variation is necessary to identify functional alleles that provide important clues to disease pathophysiology, but large-scale sequencing does not support the idea that lower-frequency variants have a major role in predisposition to type 2 diabetes.

Arrays : GWAS (Genome wide association studies)

- Aproximación empleada para el estudio de rasgos/enfermedades comunes/complejas
- Paquetes en R/Bionconductor
  - postgwas (GitHub) — M Hiersche, F Rühle, M Stoll
  - GenABEL (R) — GenABEL project developers
  - SNPassoc (R) — JR González, L Armengol, E Guinó, X Solé and V Moreno
  - snpStats (Bioconductor) — D Clayton
  - GWASTools (Bionconductor) — SM Gogarten, T Bhangale , MP Conomos , CA Laurie, CP McHugh , I Painter , X Zheng, DR Crosslin, D Levine, T Lumley , SC Nelson , K Rice , J Shen, R Swarnkar , BS Weir and CC Laurie.
- Workflow habitual:
  - Integración datos
  - Control de calidad
  - Control estructura poblacional (PCA, otros)
  - Estudio de asociación
  - Visualización de resultados
- Excelente tutorial R User Conference 2016: Andrea S. Foulkes. Genome-Wide Association Analysis and Post-Analytic Interrogation with R. http://user2016.org/tutorials/14.html

Arrays : GWAS (Genome wide association studies)

- Resultado: asociación a nivel marcador (nivel de significación para todo el genoma)

- Otros análisis complementarios:
  - Genetic risk scores
  - Enrichment analysis

- Se pueden emplear otras aproximaciones:
  - Regresión penalizada para selección de marcadores

- Enrichment analysis:

```
library(GO.db) # annotations
library(org.Hs.eg.db) # annotations
library(Category) # annotations
library(Gostats) #  over or under-representation hypergeometric test
library(multtest) # multiple testing correction

# selected_genes # character vector of genes of interest
# universe_genes # universe of genes

param.GO.BP <- new("GOHyperGParams", geneIds = selected_genes,
  universeGeneIds = universe_genes, ontology="BP",
  annotation="org.Hs.eg.db",
  testDirection="over", pvalueCutoff=1,conditional=T)
hyperGTest(param.GO.BP)
results.param.GO.BP <- as.data.frame(summary(hyperGTest(param.GO.BP)))
pvalue.GO.BP <- results.param.GO.BP$Pvalue
pvalue.adjusted.GO.BP <- mt.rawp2adjp(pvalue.GO.BP,proc=c("BH"),
   alpha = 0.05, na.rm = FALSE)
pvaladj <- as.data.frame(pvalue.adjusted.GO.BP$adjp)
resultados.GO.BP <- cbind(results.param.GO.BP,pvaladj)
```

- Regresión penalizada para selección de marcadores

```
library(scrime) # recodificar SNPs, útil para otras variables
library(glmnet) # regresión penalizada
p.otros <- read.table("p.imput.snps.txt",header=T,sep="\t")

set.seed(1000)
folds<-5
indices<-matrix(c(sample(rownames(p.otros))),ncol=folds,byrow=TRUE)
testset<-indices[,5]
trainingset<-as.vector(indices[,c(1,2,3,4)])
test<-p.otros[testset,]
learning<-p.otros[trainingset,]
snps.learning<-learning[,4:71]
snps.test<-test[,4:71]

snps.learning.aditivo<- recodeSNPs(snps.learning, first.ref =
FALSE, geno = 0:2, snp.in.col = TRUE)
snps.test.aditivo<- recodeSNPs(snps.test, first.ref = FALSE, geno =
0:2, snp.in.col = TRUE)
xlearning<-as.matrix(snps.learning.aditivo)
ylearning<-learning$caso.control
xtest<-as.matrix(snps.test.aditivo)
```

- Regresión penalizada para selección de marcadores

```
set.seed(1000)

cvglmnet.1<-cv.glmnet(xlearning,ylearning,family="binomial",
  alpha=1,nfolds=4,type.measure="class")

predicciones.cv.1se.1<-predict(cvglmnet.1,newx=xtest,s="lambda.1se",
  type="response")
coef(cvglmnet.1,s=0.01367659)
predicciones.cv.min.1<-predict(cvglmnet.1,newx=xtest,s="lambda.min",
  type="response")
coef(cvglmnet.1,s=0.03805592)
```

Next Generation Sequencing

Análisis primario

- Producción de reads y quality scores

Análisis secundario

- Filtrado reads mala calidad
- Alineamiento frente a secuencia de referencia
- Identificación de variantes (variant calling)
- Conversión de formatos de archivos

Análisis terciario

- Análisis de variante: anotación / filtrado
- Inferencia estadística, estudios de asociación
- Interpretación biológica

Oliver et al. Bioinformatics for Clinical Next Generation Sequencing.
Morgan M. Sequences, Genomes, and Genes in R / Bioconductor.
Moorthie et al. Informatics and clinical genome sequencing: opening the black blox.
http://blog.goldenhelix.com/grudy/a-hitchhikers-guide-to-next-generation-sequencing-part-2/

NGS

Análisis primario

Análisis secundario

Análisis terciario

- Capacidad computacional muy elevada
- Archivos específicos para este tipo de datos
  - (FASTAQ) – secuencias no alineadas
  - (.bam) – secuencias alineadas
  - (.vcf) – variantes indexadas
  - (.bed) – posiciones
- Análisis primario /secundario con programas específicos
- Parte del análisis secundario se puede hacer con programas externos o con R/Bioconductor: desde la importación de archivos bam

- Visualización
- Anotación adicional (bases de datos)
- Filtrado variantes
- Análisis estadístico

- Diferentes etapas no son compartimentos separados, pueden integrarse y modificarse en función del objetivo del estudio, las necesidades del investigador…
- Análisis secundario y terciario se puede hacer exclusivamente con R/Bioconductor

NGS

- ❑ Diferentes formatos de datos específicos (bam, vcf, bed)
- ❑ Archivo final con variantes anotadas (texto plano):

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
ID1	chrX	100547933	rs5951328	874	99	77708	nonsynonymous	c.101T>C	0.005	0
ID1	chr21	43522349	rs220109	629	99	72807	synonymous	c.1044T>C	0.46	0.42
ID1	chr20	45204266	rs1880898	129	99	72221	synonymous	c.1185C>T	0.45	0.44
ID1	chr20	44501458	rs6130959	7	NA	71209	nonsynonymous	c.131A>G	0.48	0.46
ID1	chr21	33368188	rs2070371	254	99	72622	synonymous	c.1413T>C	0.3	0.33
ID1	chr20	57290347	rs6026468	195	99	71343	nonsynonymous	c.1453C>G	NA	NA
ID1	chrX	12924826	rs3764880	274	99	75430	nonsynonymous	c.1A>G	0.46	0.19
ID1	chr21	45970993	rs233239	354	99	73399	nonsynonymous	c.349G>C	0.23	0.37
ID1	chrX	154456747	rs572013	856	99	76878	nonsynonymous	c.367A>G	0	0
ID1	chr20	35381262	rs73105200	470	99	72038	nonsense	c.370C>T	NA	NA
ID1	chr20	25434139	rs17857107	290	99	71842	nonsynonymous	c.470G>A	0.07	0.001
ID1	chr20	25434139	rs17857107	290	99	71842	nonsynonymous	c.470G>A	0.07	0.001
ID1	chr22	21998280	rs73166641	231	99	73668	nonsynonymous	c.482G>A	0.005	0

- 30000 – 50000 filas (variantes detectadas en 1 exoma)
- > 100 columnas de anotaciones (~20-30 interés)
- 1 data frame por muestra (se pueden combinar múltiples muestras en un data frame)

Dataset modificado a partir del ejemplo incluido en la vignette del paquete de R/Bioconductor **VarianfFiltering** - Elurbe, D.M., Mila, M. and Castelo, R. VariantFiltering: filtering of coding and non-coding genetic variants 2014
Multisample VCF file from CEU individuals of 1000G project http://www.internationalgenome.org/

NGS manipular múltiples data frames

```
library(plyr)

archivos <- list.files(pattern="1000g")
lista_archivos <- llply(archivos,read.delim,stringsAsFactors=F,
   na.strings="")

dimensiones <- ldply(lista_archivos, dim)

nombres <- ldply(lista_archivos, names)

for (i in 1:length(lista_archivos)){
  lista_archivos[[i]]$varID <-  paste(lista_archivos[[i]]$Chr,
  lista_archivos[[i]]$Start, lista_archivos[[i]]$End,
  lista_archivos[[i]]$Ref,lista_archivos[[i]]$Alt,sep=":")
}

for (i in 1:length(lista_archivos)){
  lista_archivos[[i]]$sampleID <- rep(archivos[[i]],
  dim(lista_archivos[[i]])[1])
}
```

NGS manipular múltiples data frames

```
for (i in 1:length(lista_archivos)){
  lista_archivos[[i]] <- mutate(lista_archivos[[i]],sampleID =
  sapply(strsplit(lista_archivos[[i]]$sampleID,split=".",
  fixed=TRUE),function(x)(x[2])))
  }

lista_filtrada_1 <- llply(lista_archivos,
  subset,!is.na(OMIM_disorder))
lista_filtrada_2 <- llply(lista_archivos,subset,!is.na(OMIM_disorder)
  & Zygosity=="hom")

for (i in seq_along(lista_filtrada_2)){
  write.table(lista_filtrada_2[[i]],paste(archivos[i],
  "filtro2.txt",sep="_"),quote=F,sep="\t",row.names=F)
  }

archivo_único <- do.call("rbind",lista_archivos)
```

NGS anotaciones

```
library(data.table)
library(plyr)
library(biomaRt)

tg <- fread("annotation_data.txt")
tg_genes <- unique(as.character(tg$Genes))

mymart <-useMart(biomart="ensembl", dataset="hsapiens_gene_ensembl")

tl0 <- getBM(attributes=c('hgnc_symbol','ensembl_gene_id',
  'ensembl_transcript_id','transcript_length'),filters='hgnc_symbol',
   values = tg_genes, mart = mymart)

tl1 <- tl0[,c(1,4)]
names(tl1) <- c("gene","transcript_length")

tl_ave <- ddply(tl1,.(gene),summarize,transcript_length_ave =
round(mean(transcript_length),0))
```

NGS anotaciones

```
# manipular strings
# regular expressions


library(splitstackshape)

as.data.frame(cSplit(data[,c(1,8)],"V2",","))

subset(data,grepl(paste(c("value_A","value_B"),collapse="|"),V1))

subset(data,!grepl("value_C",V1))

pattern_gene <- paste("(^|[[:punct:]])",genes,"([[:punct:]]|$)",
   collapse="|",sep="")

subset(data,grepl(pattern_gene,V2))
```

Conclusiones

- Datos de alta dimensión
- Recurso importante R/Bioconductor
- Algunos formatos exclusivos de datos genómicos
- Pipelines de análisis específicas para datos genómicos
- Manejo de datos, manipulación de cadenas de caracteres
- Anotaciones a partir de bases datos biológicas
- Análisis estadísticos específicos / no

Capacidad generar datos >>>>>> capacidad para analizar e interpretar esos datos

Muchas Gracias

Pilar Cacheiro

Grupo de Medicina Xenómica
Universidade de Santiago de Compostela