Una solución integrada con R para el Análisis de Interacciones entre genes con datos de Supervivencia en un estudio GWAS

Jesús Herranz, Antoni Picornell, María L. Calle, **Núria Malats**

Epidemiología Genética y Molecular – CNIO

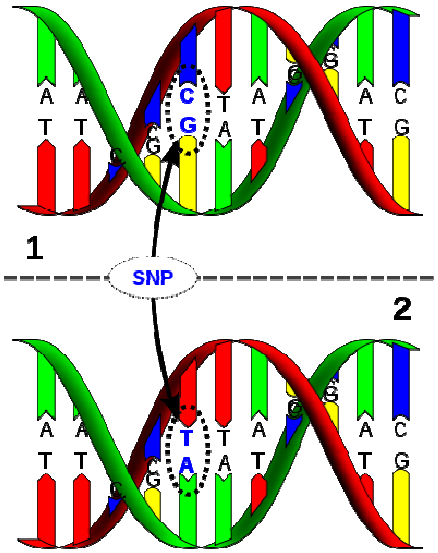
III Jornadas de Usuarios de R - 17 Noviembre 2011

cnio
Centro Nacional de Investigaciones Oncológicas

Introducción

- Estudiar **variantes genéticas** implicadas en la **progresión** de enfermedades complejas
  - Factores pronóstico – Técnicas de **análisis de supervivencia**

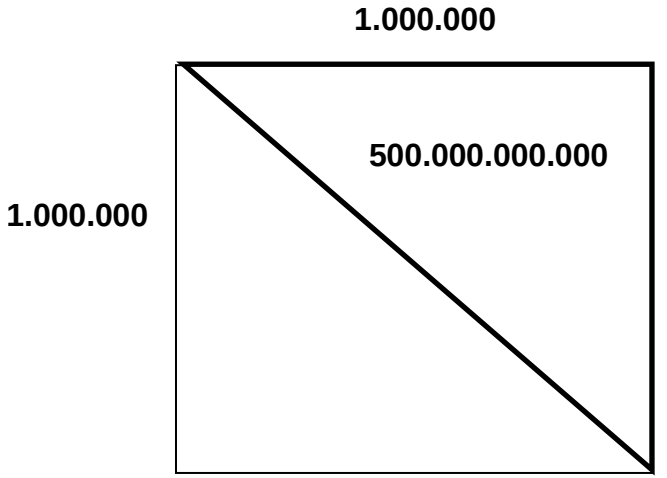
- **Polimorfismos** – SNPs (Single Nucleotide Polymorphism)
  - Cambios en el genoma de 1 sola base
  - Genotipos:
    - Homocigotos comunes
    - Heterocigotos
    - Homocigotos variantes



- **GWAS** (Genome-wide Association Studies)
  - Estudios pangenómicos (genoma completo)
  - 100.000 - 1 Millón de SNPs

Introducción

- **Interacciones gen-gen (pares)**
  - Alto coste computacional
  - Miles de millones de interacciones



- **Número de Publicaciones - GWAS**

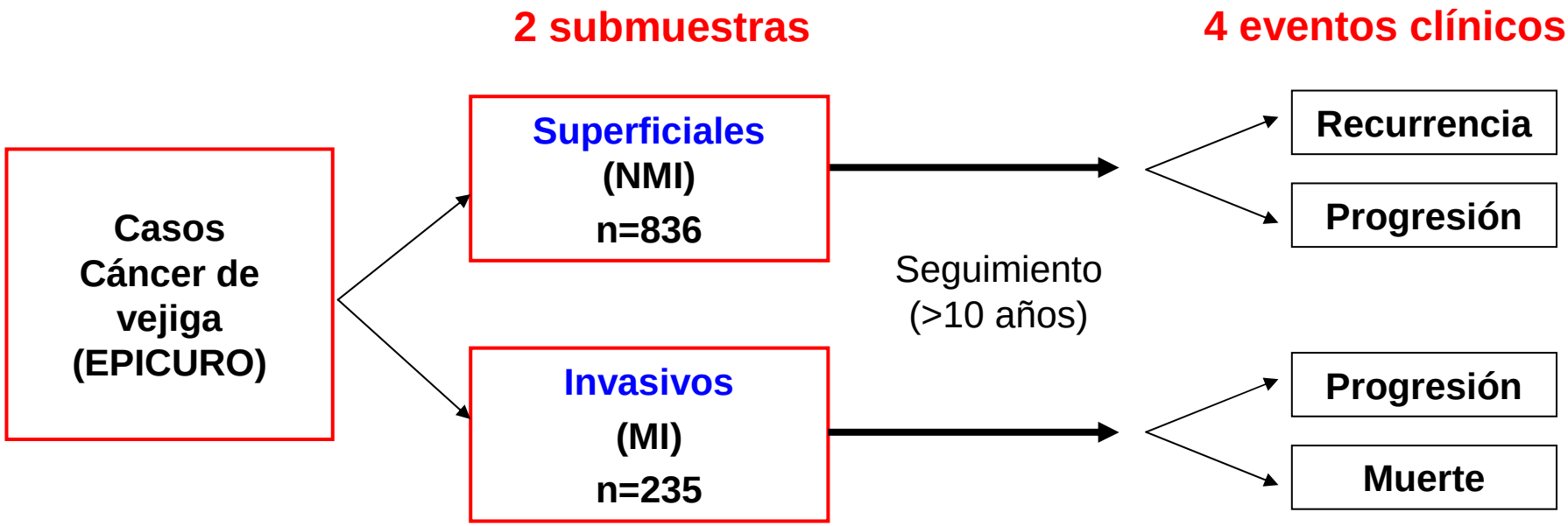
	Univariantes	Interacciones
Fac. Riesgo	> 1000	< 10
Fac. Pronóstico	< 10	**0**

Interacciones gen-gen
GWAS
Pronóstico - Supervivencia

- **Objetivo: Estrategia** para analizar **todos los pares de interacciones** de un **estudio pangenómico** que incluye **1 millón de SNPs** con **datos de supervivencia**.

Estudio Español Cáncer Vejiga / EPICURO

- Estudios multicéntrico realizado en 18 hospitales
  - 1219 pacientes de **cáncer de vejiga** (reclutamiento: 1998 – 2001)
  - 1271 controles – Estudio casos y controles – Factores riesgo
  - **1071 casos de cáncer de vejiga con información clínico-patológica, información genética y seguimiento**



Una solución integrada con R

- **Control de las 2 submuestras**
  - Control de individuos y variables en cada análisis
  - **No duplicar información**

- **R scripts únicos**
  - **Análisis simultáneo de los 4 estudios**
  - Parámetros de entrada
  - Ficheros de salida con los resultados

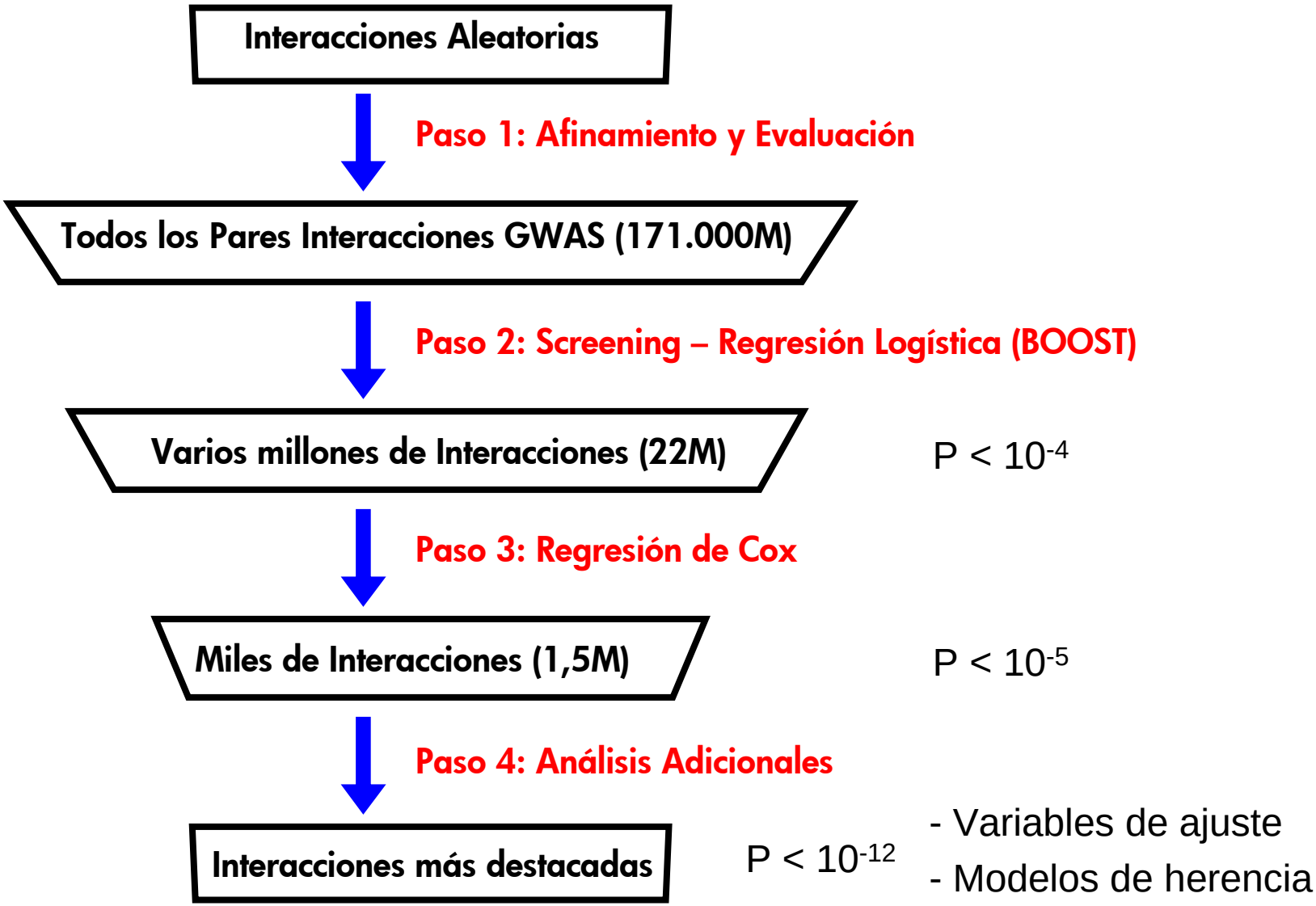
- **Funciones con los análisis estadísticos**
  - Incorporar nuevas técnicas estadísticas o modificaciones

Etapas del análisis

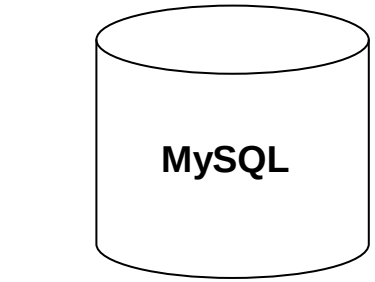
- **Preparación de los datos.** Etapas previas
  1. Almacenamiento de los datos. Objetos ff
  2. Imputación de missings
  3. Criterios de inclusión de SNPs. Control de calidad
  4. Reducción del número de variables para analizar

- **Estrategia analítica. Análisis Estadístico**
  - No se pueden analizar 500.000 millones de interacciones con Regresión de Cox
  - Etapa de screening con Regresión Logística (BOOST - C)
  - Análisis posteriores con Regresión de Cox
  - Alternativas: Survival - MDR (Multifactor Dimensionality Reduction)

Descripción de la Estrategia Analítica



Etapa 1: Almacenar los datos del GWAS (ff)



AA\tAT\tGG\tAA ……
GG\tAA\tTG\tTT ……
……
……

998.349 x 1
Cada fila, 1 SNP

R

Creación de objetos con los datos de los SNPs

0 Hom. común
1 Heterocigoto
2 Hom. Variante

Se mantienen los NAs

0 0 1 0 NA 0 1 2 0 ……
0 1 NA 2 0 0 0 1 2 ……
… NA ….. NA ……..

CONTROLES

1095 x 998.349

0 0 NA 0 2 1 0 0 0 ……
1 0 1 0 2 0 NA 0 0 ……
…. NA ……NA ……

CASOS

1134 x 998.349

- **indGWAS**
- NameSNP
- Alleles

Summary SNPs (df)

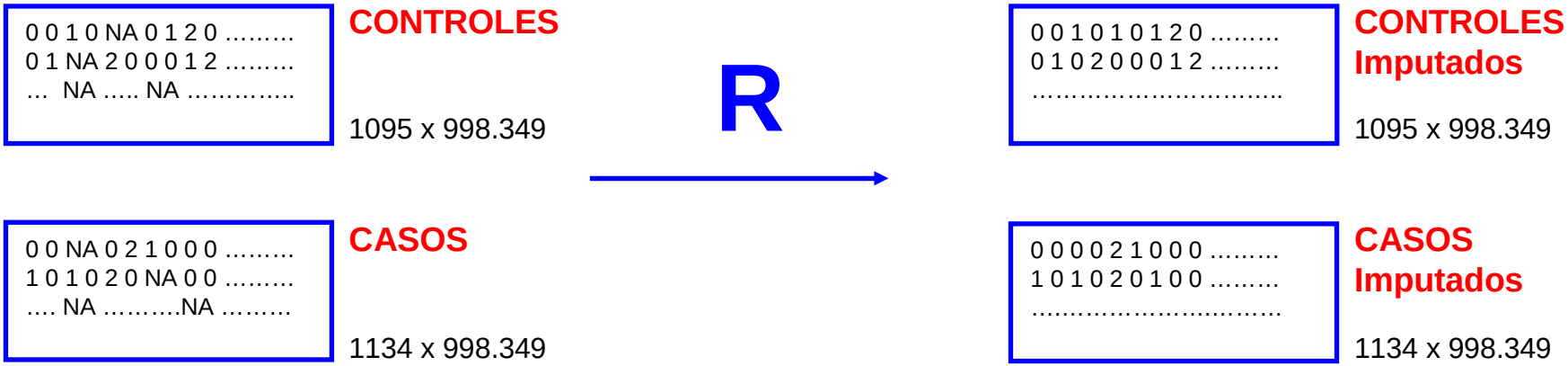
998.349

- **id**

Identificadores pacientes (vector)

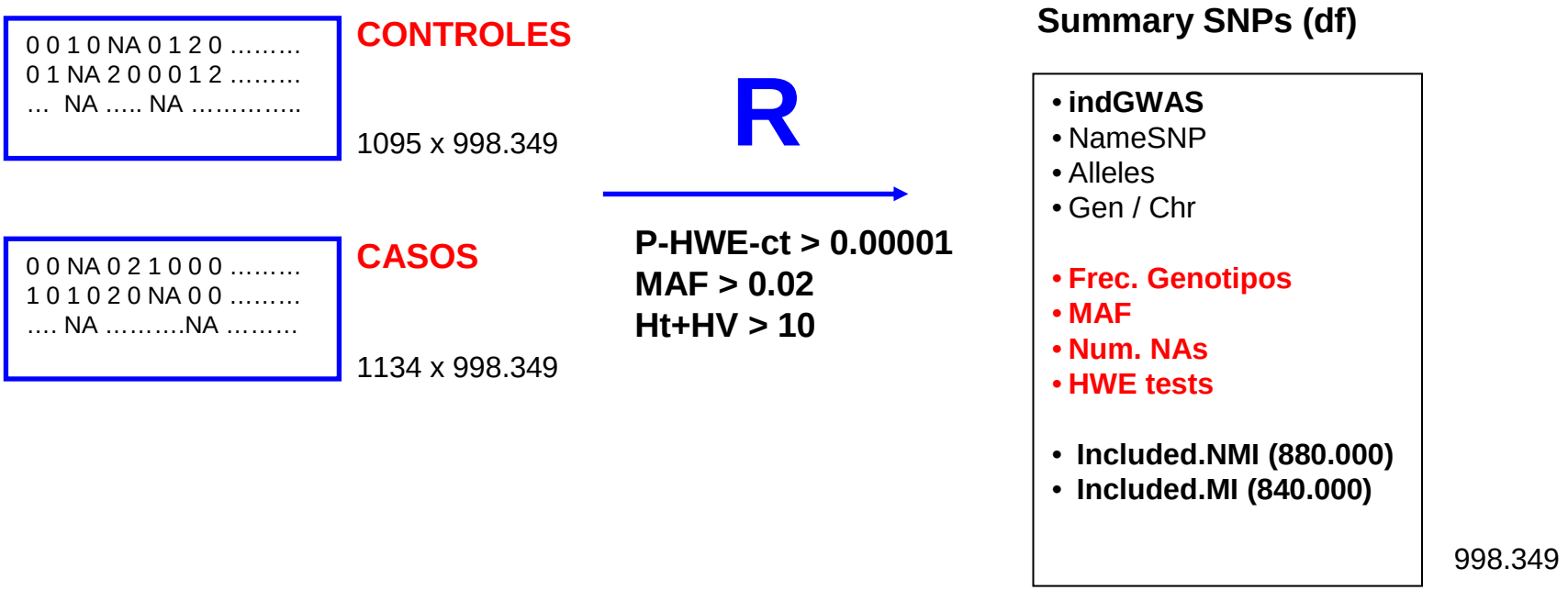
- La **librería ff** permite crear objetos
  - Capacidad de almacenamiento, rápido acceso, no usa memoria de R
  - Columnas: **nombres de los SNPs**
  - Filas: **identificadores de los pacientes**

Etapa 2: Imputación de Missings



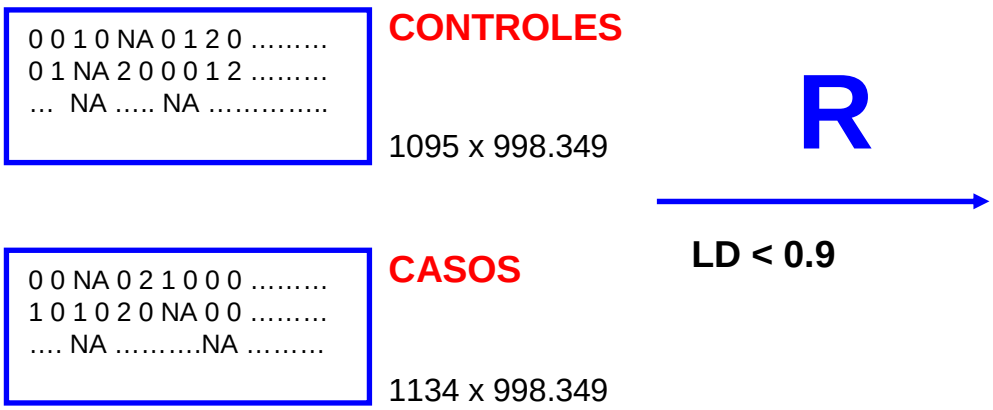
- Imputación por **Random Forests**
  - Ventana con los SNPs más próximos (correlacionados, LD)
- Otros métodos de imputación
  - **Mantener los datos originales sin imputar**
  - Crear otros ficheros de datos imputados

Etapa 3: Control de Calidad



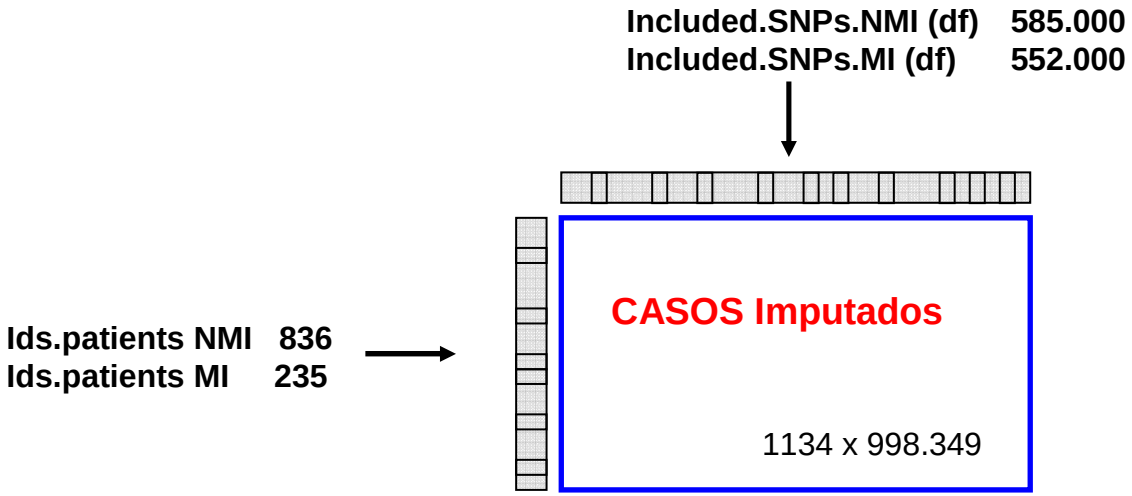
- **SNPs eliminados** del análisis
  - **MAF (minor allele frecuency):** poca variabilidad en la población
  - **HWE (Hardy-Weinberg equilibrium):** errores de genotipado
- Otros criterios de inclusión
  - No se recalcula MAF / HWE

Etapa 4: Linkage disequilibrium (LD)



- **LD – SNPs correlacionados**
  - Se detectan **bloques de LD**
  - Se detectan **singletons** (SNPs no correlacionados)
  - **Representantes del bloque**: entran en las fases de análisis de screening
  - No representantes del bloque: entran en las fases finales (comp. múlt.)

Análisis simultáneos (4 estudios)



NMI – EPICURO (df - 836)

- Tiempo / Status RECUR
- Tiempo / Status PROGR
- Covariables RECUR
- Covariables PROGR

MI – EPICURO (df - 235)

- Tiempo / Status PROGR
- Tiempo / Status DEATH
- Covariables PROGR
- Covariables DEATH

- **R Scripts únicos y flexibles**
  - Inclusión de otras **submuestras**
  - Inclusión de otro **conjunto de SNPs** (criterios de inclusión, LD)
  - Inclusión de otros **eventos clínicos** de interés (tiempo y status)
  - Inclusión de otras **variables de ajuste**
  - **Parámetro de entrada:** estudio

Análisis estadístico (NMI – Progresión)

- NameSNP
- indGWAS
- indBOOST
- Modelo Herencia

Ids. NMI 836

BOOST NMI Progr

0	0 0 1 1 0 1 0 1 2 0
0	1 1 2 0 0 0 0 0 1 2
1	0 0 0 0 0 1 0 2 1 0

836 x 585.000

Boost - C

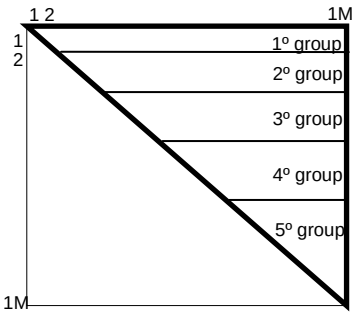
Regresión Logística
171.000 M interacciones
$P < 10^{-4}$

5 procesos (4 días)

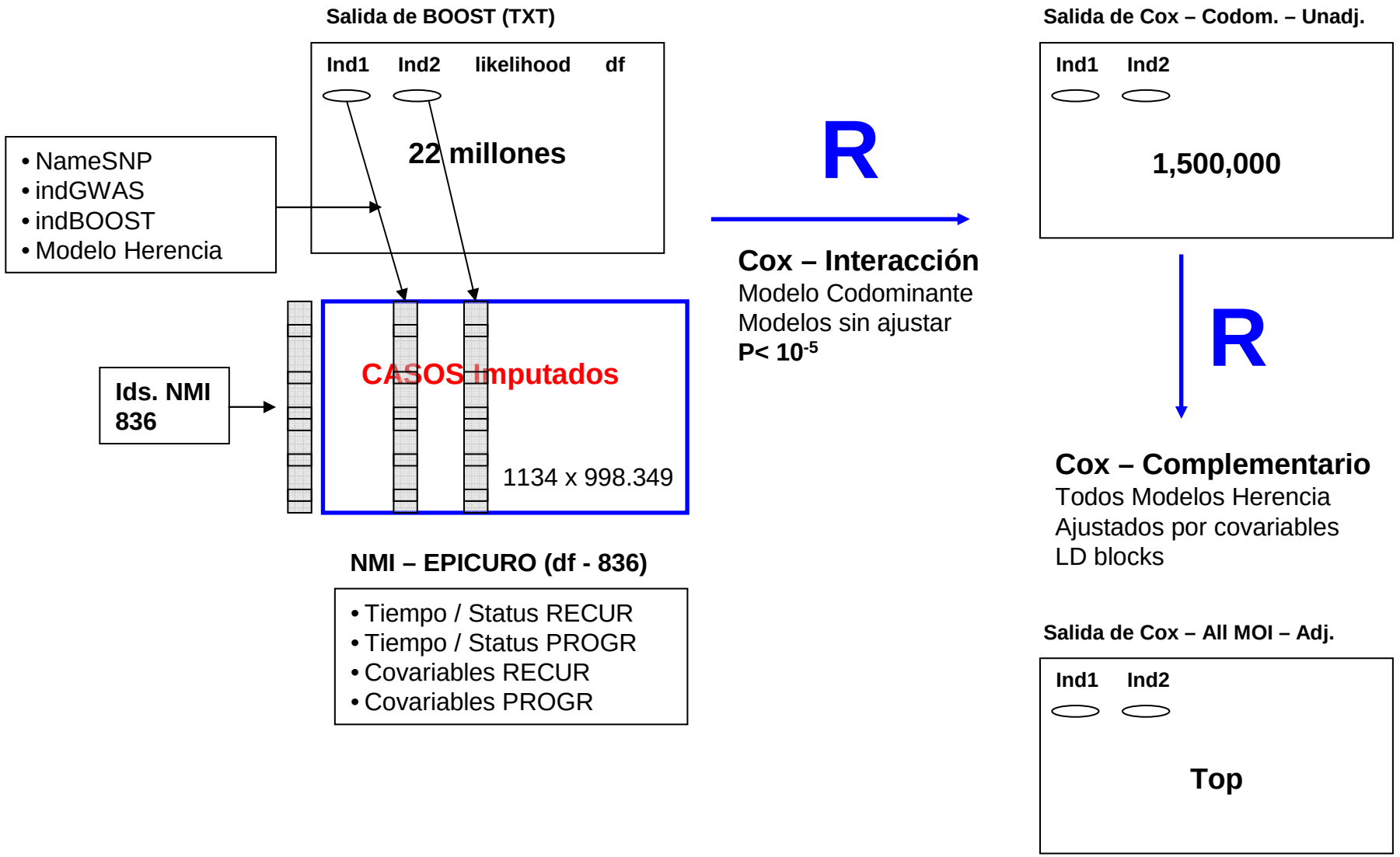
Salida de BOOST (TXT)

Ind1	Ind2	likelihood	df

22 millones



Análisis estadístico (NMI – Progresión)



Resultados

Subphenotype	Outcome	Chr - Gen 1		Chr - Gen 2		MOI 1	MOI 2	P-Value	Threshold Informative SNF 8.60E-14	Threshold 0.4 - Factor 1.06E-13
NMI	**Recurrence**	16	*Gen1*	19	*Gen2*	A	A	**5.53E-14**	**Sign.**	**Sign.**
	Recurrence	14	*Gen3*	20	*Gen4*	R	A	**7.31E-14**	**Sign.**	**Sign.**
	Recurrence	11	*Gen5*	14	*Gen6*	D	C	**1.91E-13**	NS	NS
	Recurrence	10	*Gen7*	21	*Gen8*	D	D	**1.94E-13**	NS	NS
NMI	**Progression**	18	*Gen9*	X	*Gen10*	C	C	**5.33E-15**	**Sign.**	**Sign.**
	Progression	5	*Gen11*	14	*Gen12*	R	D	**7.78E-14**	**Sign.**	**Sign.**
	Progression	7	*Gen13*	10	*Gen14*	A	C	**2.46E-13**	NS	NS

Subphenotype	Outcome	Chr - Gen 1		Chr - Gen 2		MOI 1	MOI 2	P_Value	Threshold Informative SNF 9.46E-14	Threshold 0.4 - Factor 1.17E-13
MI	**Progression**	2	*Gen15*	8	*Gen16*	A	D	**7.54E-14**	**Sign.**	**Sign.**
	Progression	2	*Gen17*	16	*Gen18*	D	R	**1.66E-13**	NS	NS
MI	**Death**	8	*Gen19*	12	*Gen20*	D	D	**4.11E-14**	**Sign.**	**Sign.**
	Death	7	*Gen21*	10	*Gen22*	C	C	**7.24E-14**	**Sign.**	**Sign.**
	Death	7	*Gen23*	14	*Gen24*	R	D	**9.87E-14**	**Sign.**	**Sign.**
	Death	17	*Gen25*	22	*Gen26*	C	D	**1.17E-13**	NS	**Sign.**
	Death	1	*Gen27*	3	*Gen28*	D	D	**1.56E-13**	NS	NS

- Modelos de Cox **ajustados por covariables clínicas**
- Distintos **modelos de herencia** genética (dominante, recesivo, aditivo, codominante)
- Significación basada en **comparaciones múltiples**

Conclusiones

- **Estrategia analítica novedosa y viable**
  - Exhaustivamente todas los millones de Interacciones gen-gen
  - Análisis de supervivencia – Pronóstico – GWAS
  - Tiempo de computación aceptable (15-20 días)
- Hemos encontrado **varias interacciones gen-gen estadísticamente significativas**

Limitaciones

- **Interacciones perdidas** (con tamaños muestrales bajos)
- Los resultados deberían ser **replicados**

Agradecimientos

- Nuria Malats
- Toni Picornell
- Roger Milne
- Evangelina López
- Gaëlle Marene
- Mirari Márquez
- André Amaral
- Salman Tajuddin
- Matt Czachorowski

cnio Centro Nacional de Investigaciones Oncológicas

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN

NATIONAL CANCER INSTITUTE

- Francisco X Real (CNIO)
- Stephen Chanock (NCI)
- Nathaniel Rothman (NCI)
- M. García-Closas (NCI)
- Debra Silverman (NCI)
- María L. Calle (Unv. Vic)
- Manolis Kogevinas (CREAL)