Generalitat de Catalunya
Departament de Salut

ICO
Institut Català d'Oncologia

Implementation of the statistical pipeline for the SNPStats 2.0 genetic epidemiology web tool

Xavier Solé, PhD

x.sole@iconcologia.net

Jornada I+D MESIO UPC-UB Maig 2018

Cancer Data Science (CaDS) group at ICO-IDIBELL

- Computation-based group aimed at improving cancer diagnosis, prognosis, and treatment through the bioinformatics/statistical analysis and integration of large-scale *omic* data.

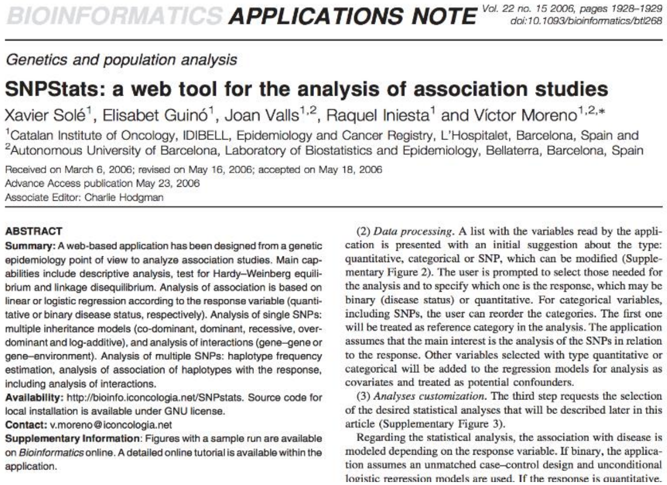
- Created in early 2017. Currently composed of 3 members (1 postdoc + 1 technician + PI), plus a summer intern joining in June 2018.

Cancer Data Science (CaDS) group at ICO-IDIBELL

- Current main projects:
  - Development of a bioinformatics-based platform for precision medicine in lung cancer.
  - Real-time precision medicine: a challenging clinical case.
  - Identification of new biomarkers of prognosis in malignant pleural mesothelioma.
  - Network-based characterization of master regulators in malignant pleural mesothelioma (summer internship).
  - **Development of SNPStats 2.0 genetic epidemiology web tool.**

SNPStats 1.0 (www.snpstats.net)

- Web-based tool for the statistical analysis of small-scale genetic epidemiology analyses.
- Developed in 2006. Since then, it has been used in more than 300,000 analyses from researchers around the world and has been cited more than 1100 times.



BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 15 2006, pages 1928–1929
doi:10.1093/bioinformatics/btl268

Genetics and population analysis

SNPStats: a web tool for the analysis of association studies

Xavier Solé[1], Elisabet Guinó[1], Joan Valls[1,2], Raquel Iniesta[1] and Víctor Moreno[1,2,*]

[1]Catalan Institute of Oncology, IDIBELL, Epidemiology and Cancer Registry, L'Hospitalet, Barcelona, Spain and [2]Autonomous University of Barcelona, Laboratory of Biostatistics and Epidemiology, Bellaterra, Barcelona, Spain

Received on March 6, 2006; revised on May 16, 2006; accepted on May 18, 2006
Advance Access publication May 23, 2006
Associate Editor: Charlie Hodgman

ABSTRACT

Summary: A web-based application has been designed from a genetic epidemiology point of view to analyze association studies. Main capabilities include descriptive analysis, test for Hardy–Weinberg equilibrium and linkage disequilibrium. Analysis of association is based on linear or logistic regression according to the response variable (quantitative or binary disease status, respectively). Analysis of single SNPs: multiple inheritance models (co-dominant, dominant, recessive, overdominant and log-additive), and analysis of interactions (gene–gene or gene–environment). Analysis of multiple SNPs: haplotype frequency estimation, analysis of association of haplotypes with the response, including analysis of interactions.

Availability: http://bioinfo.iconcologia.net/SNPstats. Source code for local installation is available under GNU license.

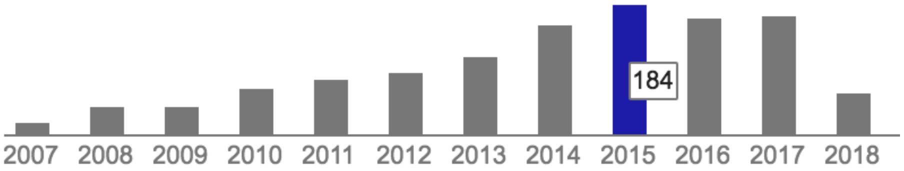
Contact: v.moreno@iconcologia.net

Supplementary Information: Figures with a sample run are available on *Bioinformatics* online. A detailed online tutorial is available within the application.

(2) *Data processing*. A list with the variables read by the application is presented with an initial suggestion about the type: quantitative, categorical or SNP, which can be modified (Supplementary Figure 2). The user is prompted to select those needed for the analysis and to specify which one is the response, which may be binary (disease status) or quantitative. For categorical variables, including SNPs, the user can reorder the categories. The first one will be treated as reference category in the analysis. The application assumes that the main interest is the analysis of the SNPs in relation to the response. Other variables selected with type quantitative or categorical will be added to the regression models for analysis as covariates and treated as potential confounders.

(3) *Analyses customization*. The third step requests the selection of the desired statistical analyses that will be described later in this article (Supplementary Figure 3).

Regarding the statistical analysis, the association with disease is modeled depending on the response variable. If binary, the application assumes an unmatched case–control design and unconditional logistic regression models are used. If the response is quantitative.

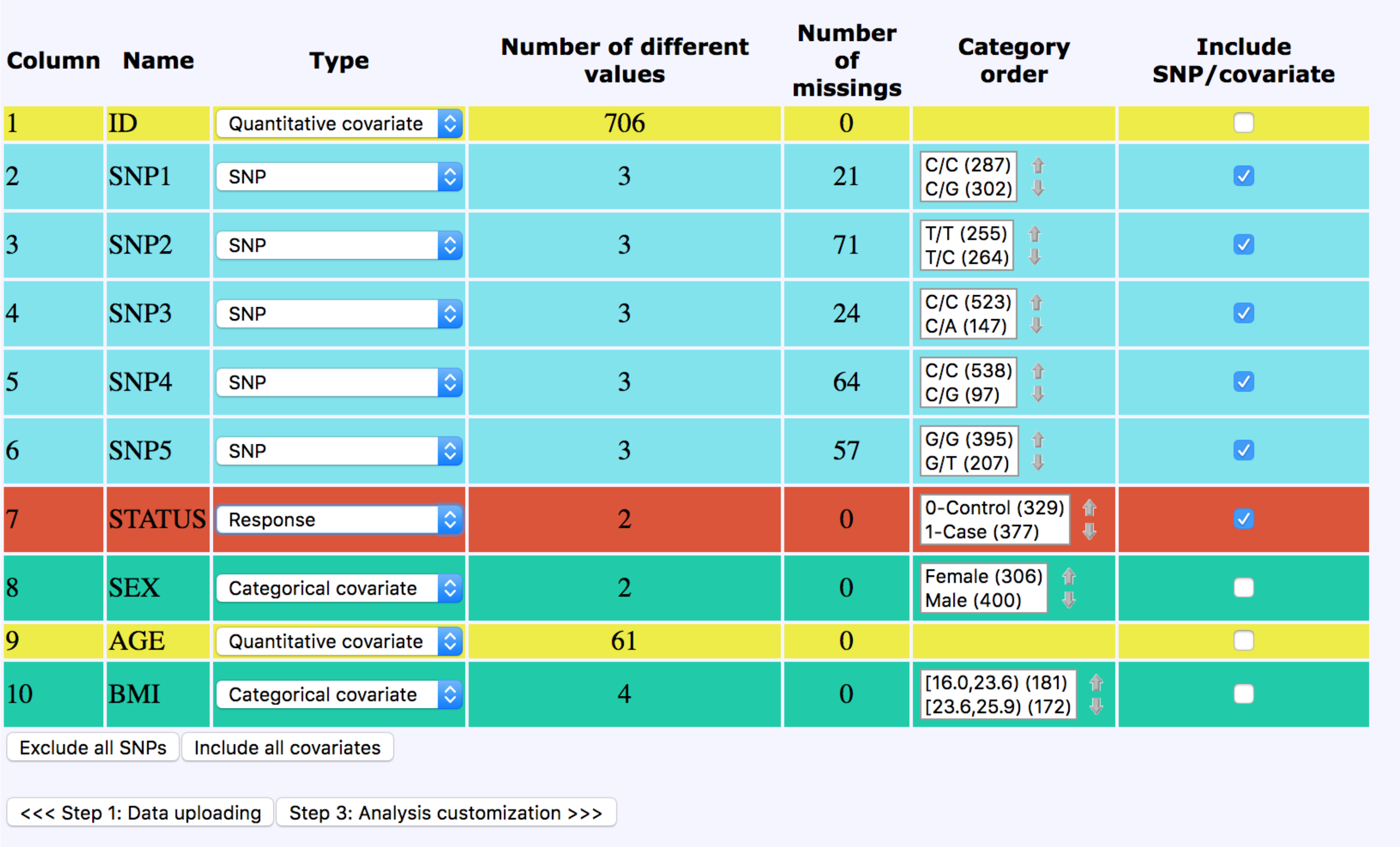


Solé X, *et al*. Bioinformatics 22(15):1928-9, 2006.

What does SNPStats 1.0 do?

- In essence, SNPStats evaluates the association of a genetic marker (i.e., A/A, A/T, T/T) with a response variable (i.e., case/control).
- Additionally, also computes descriptive statistics, haplotype analyses, and performs interaction analyses between the genetic traits and other covariates,

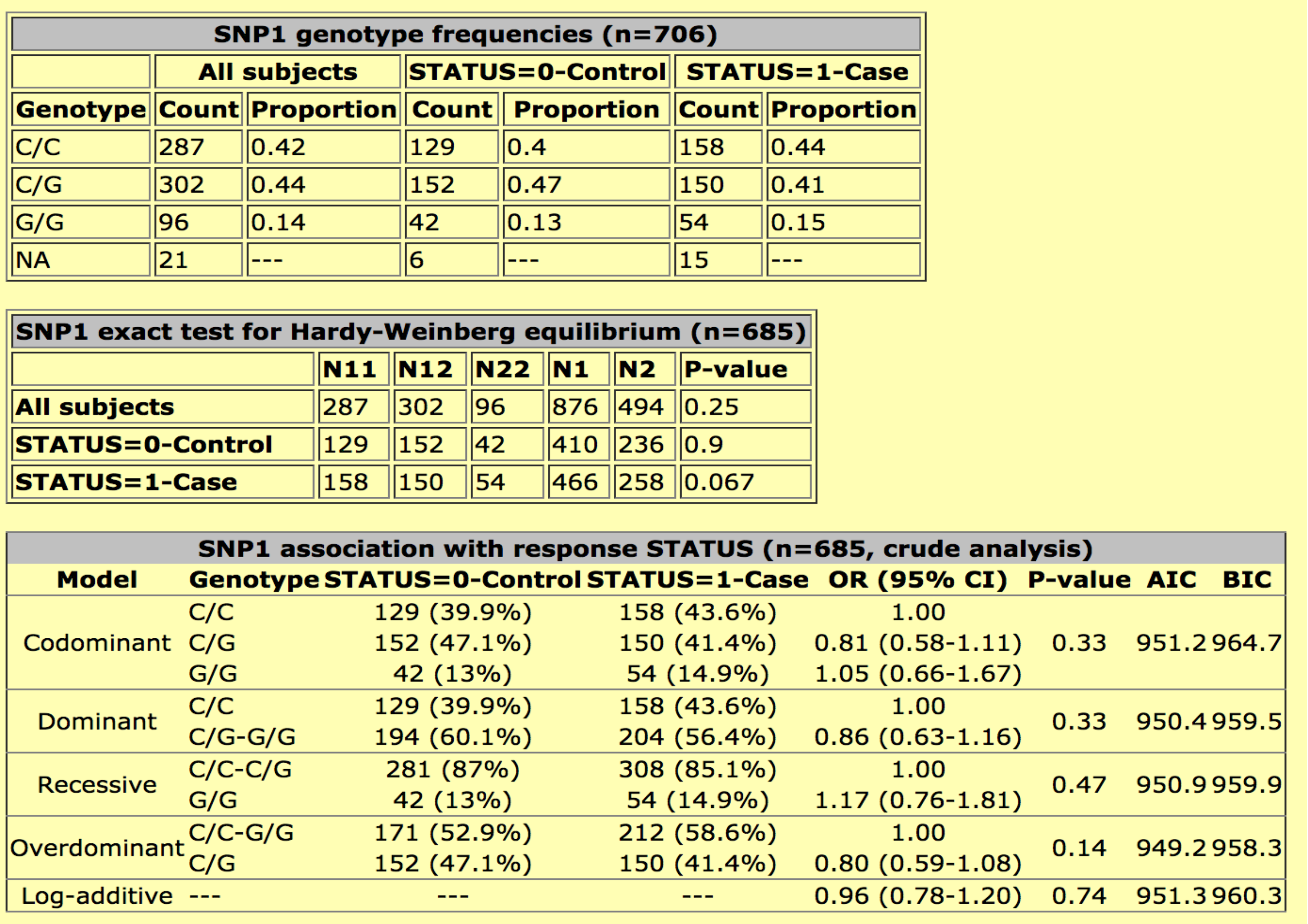
What does SNPStats 1.0 do?



Column	Name	Type	Number of different values	Number of missings	Category order	Include SNP/covariate
1	ID	Quantitative covariate	706	0		☐
2	SNP1	SNP	3	21	C/C (287) C/G (302)	☑
3	SNP2	SNP	3	71	T/T (255) T/C (264)	☑
4	SNP3	SNP	3	24	C/C (523) C/A (147)	☑
5	SNP4	SNP	3	64	C/C (538) C/G (97)	☑
6	SNP5	SNP	3	57	G/G (395) G/T (207)	☑
7	STATUS	Response	2	0	0-Control (329) 1-Case (377)	☑
8	SEX	Categorical covariate	2	0	Female (306) Male (400)	☐
9	AGE	Quantitative covariate	61	0		☐
10	BMI	Categorical covariate	4	0	[16.0,23.6) (181) [23.6,25.9) (172)	☐

Exclude all SNPs
Include all covariates
<<< Step 1: Data uploading
Step 3: Analysis customization >>>

What does SNPStats 1.0 do?



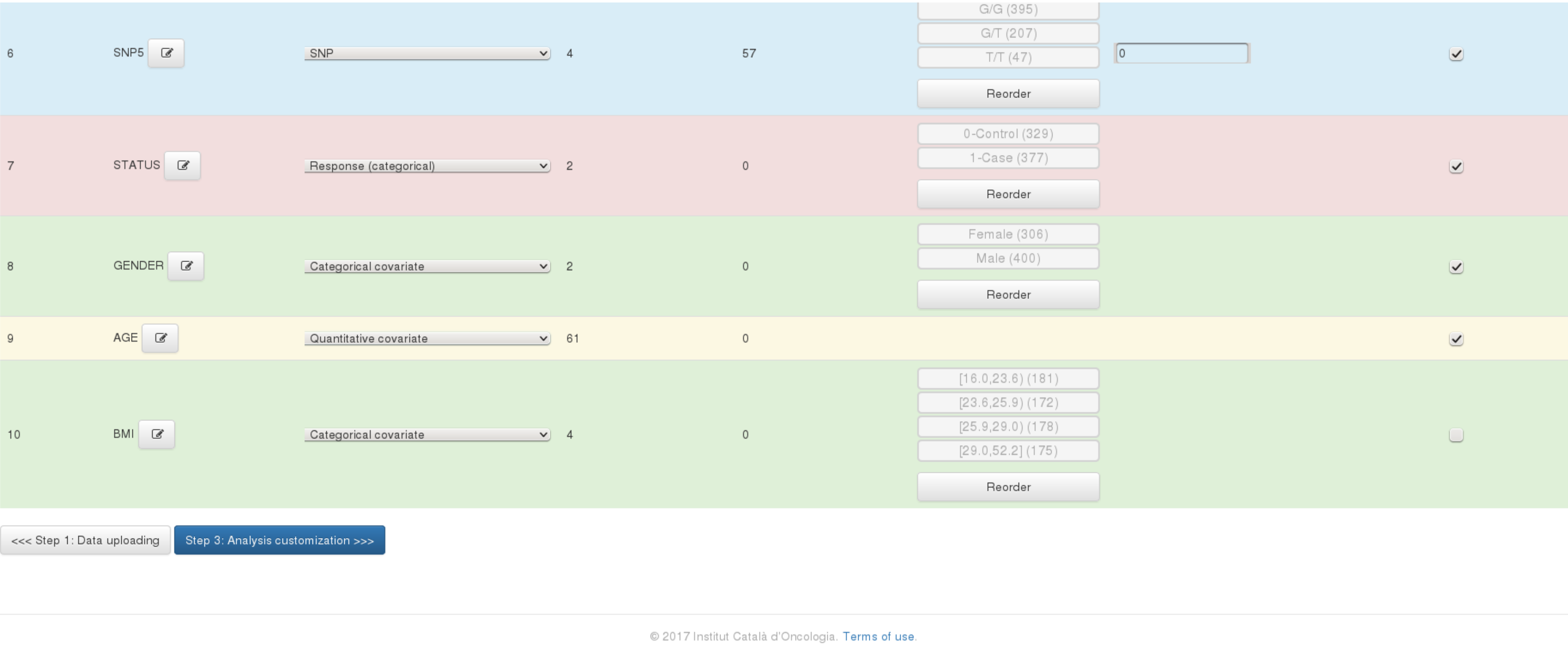
SNP1 genotype frequencies (n=706)						
	All subjects		STATUS=0-Control		STATUS=1-Case	
Genotype	Count	Proportion	Count	Proportion	Count	Proportion
C/C	287	0.42	129	0.4	158	0.44
C/G	302	0.44	152	0.47	150	0.41
G/G	96	0.14	42	0.13	54	0.15
NA	21	---	6	---	15	---

SNP1 exact test for Hardy-Weinberg equilibrium (n=685)						
	N11	N12	N22	N1	N2	P-value
All subjects	287	302	96	876	494	0.25
STATUS=0-Control	129	152	42	410	236	0.9
STATUS=1-Case	158	150	54	466	258	0.067

SNP1 association with response STATUS (n=685, crude analysis)							
Model	Genotype	STATUS=0-Control	STATUS=1-Case	OR (95% CI)	P-value	AIC	BIC
Codominant	C/C	129 (39.9%)	158 (43.6%)	1.00	0.33	951.2	964.7
	C/G	152 (47.1%)	150 (41.4%)	0.81 (0.58-1.11)			
	G/G	42 (13%)	54 (14.9%)	1.05 (0.66-1.67)			
Dominant	C/C	129 (39.9%)	158 (43.6%)	1.00	0.33	950.4	959.5
	C/G-G/G	194 (60.1%)	204 (56.4%)	0.86 (0.63-1.16)			
Recessive	C/C-C/G	281 (87%)	308 (85.1%)	1.00	0.47	950.9	959.9
	G/G	42 (13%)	54 (14.9%)	1.17 (0.76-1.81)			
Overdominant	C/C-G/G	171 (52.9%)	212 (58.6%)	1.00	0.14	949.2	958.3
	C/G	152 (47.1%)	150 (41.4%)	0.80 (0.59-1.08)			
Log-additive	---	---	---	0.96 (0.78-1.20)	0.74	951.3	960.3

SNPStats 2.0

- We want to upgrade SNPStats to a new version, with improved functionality and a better user interface.
- The new user interface was already developed by Jon Aldazabal, a CFGS intern at the CaDS group between Dec 2016-April 2017.



6	SNP5	SNP	4	57	G/G (395) G/T (207) T/T (47) Reorder	0	☑
7	STATUS	Response (categorical)	2	0	0-Control (329) 1-Case (377) Reorder		☑
8	GENDER	Categorical covariate	2	0	Female (306) Male (400) Reorder		☑
9	AGE	Quantitative covariate	61	0			☑
10	BMI	Categorical covariate	4	0	[16.0,23.6) (181) [23.6,25.9) (172) [25.9,29.0) (178) [29.0,52.2] (175) Reorder		☐

<<< Step 1: Data uploading
Step 3: Analysis customization >>>

© 2017 Institut Català d'Oncologia. Terms of use.

SNPStats 2.0

- We need now to develop the new statistical engine for SNPStats with R:
  - Adding continuous and survival response variables.
  - Adding SNP-SNP interactions.
  - Modularize the code.
  - Generate output in text format.
- Knowledge of biology/genetics is NOT required, but if you are curious about the field it would be a plus. ☺

Generalitat de Catalunya
Departament de Salut

ICO
Institut Català d'Oncologia

If you are interested please come outside and ask!

x.sole@iconcologia.net

Institut Català d'Oncologia

ICO l'Hospitalet
Hospital Duran i Reynals
Av. Granvia de L'Hospitalet, 199-203
08908 L'Hospitalet de Llobregat

ICO Badalona
Hospital Germans Trias i Pujol
Ctra. del Canyet s/n
08916 Badalona

ICO Girona
Hospital Doctor Trueta
Av. França s/n
17007 Girona

ICO Camp de Tarragona i Terres de l'Ebre
Hospital Joan XXIII
C. Dr. Mallafrè Guasch, 4 43005 Tarragona
Hospital Verge de la Cinta
C. de les Esplanetes, 14 43500 Tortosa