

Report.

ESF exploratory workshop

Workshop on genomic approaches to microarray data analysis

Held in the Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain, in 30-31 October 2003

<http://bioinfo.cnio.es/meetings/ESFWorkshopArrays03/>

Organised by:

Joaquín Dopazo, Ramón Díaz-Uriarte

Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Spain

Alvis Brazma

European Bioinformatics Institute, UK

Executive summary

This workshop focused on microarray data analysis with special stress on the state-of-the-art methodologies and the available solutions implemented as methods or software, already available. The main topics for discussion were organized as five different sessions: I) **Databases and tools**, where standards and implementations for microarray data storage were presented. II) **Information mining and automatic annotation methods** where text-free and ontology-based approaches were discussed. III) **Experiment design and normalization**, where the necessity of making a proper design of the experiment and applying the appropriate normalization procedure so as experiments be comparable, was discussed. IV) **Data analysis, clustering, gene selection**, where different analysis procedures and some important statistical remarks were made. And V) **Gene networks**, where different methods for gene networks reconstruction, and the future of the field was discussed.

Regarding the solutions, several tools already implemented were presented and discussed. About data storage, MAGE-ML and the standard MIAME were mentioned. The database ArrayExpress, at the EBI, MIAME compliant, was a public repository for data storage. The EBI has its own resource for microarray data analysis, the EP, which covers mainly clustering problems, as well as some aspects of motif discovery in co-expressed genes. The GEPAS, developed at the CNIO, covered more topics, apart from clustering, which include gene selection, supervised clustering and datamining. As an example of LIMS, and integrated database management system, the BASE initiative was shown.

The workshop was held at the CNIO, whose the facilities (Auditorium, canteen, projection facilities, etc) were offered at no cost. This has been the first ESF meeting

held in the institution and it was really successful in terms of interest, attendance and scientific quality of the presentations and the discussion.

Two companies sponsored the event in addition to ESF: Silicon Graphics (SGI, the computer's manufacturer) and BioAlma S.A, the first Spanish bioinformatics company. SGI took care of the dinners of the speakers and the farewell banquet and BioAlma made the folders and the documentation. Both companies were interested in possible bioinformatic applications in the field of functional genomics.

Scientific content

This workshop brought experts from the areas of functional genomics, bioinformatics and computer science to discuss advances, challenges and opportunities in DNA microarray data analysis.

DNA microarray technology emerges as an essential tool for studying biological processes at genomic level but, at the same time, poses new challenges on the side of the analysis and storage of the huge amounts of data produced. Relevant issues in data analysis are: data processing, including image analysis, and normalization procedures; clustering and multiple correlation techniques are key for understanding the differences between experimental conditions or clinical outcomes; automatic information mining techniques are essential to transfer biological knowledge to genes, when thousands of them are involved in the study; finally, inference of gene networks is crucial in the knowledge on how the biological systems work.

The above mentioned issues were summarised in four sessions, namely:

- I) **Databases and tools**, where standards and implementations for microarray data storage were presented. Also The two most important packages for gene expression data analysis available over the web (EBI's EP cluster and CNIO's GEPAS) were presented.
- II) **Information mining and automatic annotation methods**. In this session, the two main trends for automatic use of information in microarray experiments, text-free and ontology-based were discussed. Both presented advantages and pitfalls, but it was obvious that some automation is needed for converting microarray experiments into knowledge. The package FatiGO, the most popular statistical framework for the use of GO terms in microarray experiments was discussed.
- III) **Experiment design and normalization**. Without a proper experiment design, some results can be, simple, unachievable. In addition, normalization, necessary for a proper comparison of experiments, needs still of some developments. Also, the BASE software for data management and storage was presented within this context. Some plans for the expansion of BASE to more specialized data analysis subjects was discussed.
- IV) **Data analysis, clustering, gene selection**. Two hot topics were discussed here: the inclusion of information on regulation (mainly putative co-regulation bases in sharing putative transcription factor binding sites) and the proper statistical evaluation of the results. To avoid microarray experiments be a "fishing expedition" multiple-testing procedures must be used.
- V) **Gene networks**. Another emerging topic is the derivation of transcriptional gene networks. Several approaches were presented. Despite the complexity of this issue, there is a growing interest, because systems-biology approaches to the knowledge of living organisms require of the comprehension of the interactions among their parts.

The presentations covered several biological problems related to functional genomics and DNA microarray data analysis. They achieved a balance between the introduction of principles and methods and the discussion of applications. Moreover, they provided

the audience with useful insights into the current state of the art, as well as emerging requirements. The meeting offered 21 presentations, 13 of them given by invited speakers.

Assessment of the results

This workshop has provided a broad vision of the field and promoted the exchange of experiences between researchers with different background. Experimental researchers have had the opportunity of learning from experts statisticians and bioinformaticians the state-of-the-art methodologies and, on the opposite, the first ones have transmitted to the second ones real problems and motivations for biomedical research.

One important factor for assessing the results of this event is the level and quality of discussion achieved during the workshop. Despite the multidisciplinary characteristics of the audience, participants actively expressed technical and conceptual concerns, discussed opportunities for improving solutions and shared expertise on different issues of interest for both system designers and users

This workshop has supported the consolidation of existing national and international collaborations by providing feedback and reviewing advances. It also represented an opportunity for the organisation of future meetings. Also it allowed some of the participants to discuss the development of new collaborations involving: Exchange of technical expertise, share of information resources and participation in existing projects. Furthermore, this meeting has significantly promoted potential interdisciplinary cooperation between bioscientists and computer scientists. Some of the participants would be interested in implementing activities such as research visits and exchanges. Options being considered include programmes such as the ESF travel fellowships, and funding provided by local or national organisations.

Final programme

Thursday, 30 October

9.00 Registration

9.15 Welcome

Session I. Databases and tools

9.30 **Alvis Brazma** EBI-EMBL, UK Standards and public repository for microarray data **ArrayExpress**

10.15 **Javier Herrero** Bioinformatics, CNIO, Spain. **GEPAS, a package for DNA microarray data analysis**

11.00 Coffee break

Session II. Information mining and automatic annotation methods.

11.30 **Jan Komorowski** Uppsala University, The Linnaeus Centre for Bioinformatics, Sweden. **Predicting Gene Ontology Biological Process From Temporal Gene Expression Patterns**

12.15 **Alfonso Valencia**, CNB-CSIC, Madrid Spain **Text mining and expression data clustering**

Oral communications

13.00 Diego Araceli

13.15 Javier Tamames

13.30 Joaquín Dopazo

14.00 Lunch

Session III. Experiment design and normalization

15.30 **Anja von Heydebreck**, Max Planck Institute for Molecular Genetics, Berlin, Germany **Error modeling, data transformation and robust normalization for microarray data**

16.15 **Sandrine Dudoit** Division of Biostatistics School of Public Health University of California, Berkeley, USA **Resampling-based multiple testing procedures for identifying differentially expressed genes**

17.00 **Patrick Kemmeren**. Genomics Laboratory, UMC Utrecht The Netherlands. **Microarray data normalization**

Oral communications

17.45 Javier de las Rivas

20.30 **Dinner**

Friday, 31 October

Session IV. Data analysis, clustering, gene selection

9.30 Yves Moreau *ESAT-SISTA/COSIC/DocArch, France* **Discovering cis-regulatory modules**

10.15 Mark van der Laan. Division of Biostatistics School of Public Health University of California, Berkeley, USA **Cross-Validated Deletion/Substitution/Addition Algorithms in Regression**

11.00 Yoav Benjamini. Department of Statistics, Tel Aviv University, Israel: **False discovery rate control in complex microarray studies**

Oral communications

11.45 Ramón Díaz-Uriarte

12.15 Anat Reiner

12.30 Giorgio Valentini

13.00 **Lunch**

Session V. Gene networks

15.00 Ron Shamir School of Computer Science, Tel Aviv University, Israel. **Computational dissection of regulatory networks using gene expression and other high throughput data**

15.45 Yoseph Barash School of Computer Science & Engineering, Hebrew University, Israel. **Probabilistic Models for Identifying Regulation Networks**

16.30 Thomas Schlitt, EMBL-EBI. UK **Identification of functional relationships between genes based on the comparison of gene networks.**

Oral communications

17.15 Francesco Falciani

17.45 **Closing**

Final list of participants

See accompanying excel file

Statistical information on participants.

The following figures show statistical information on participants: gender, country of origin, main scientific activity or background and age structure.

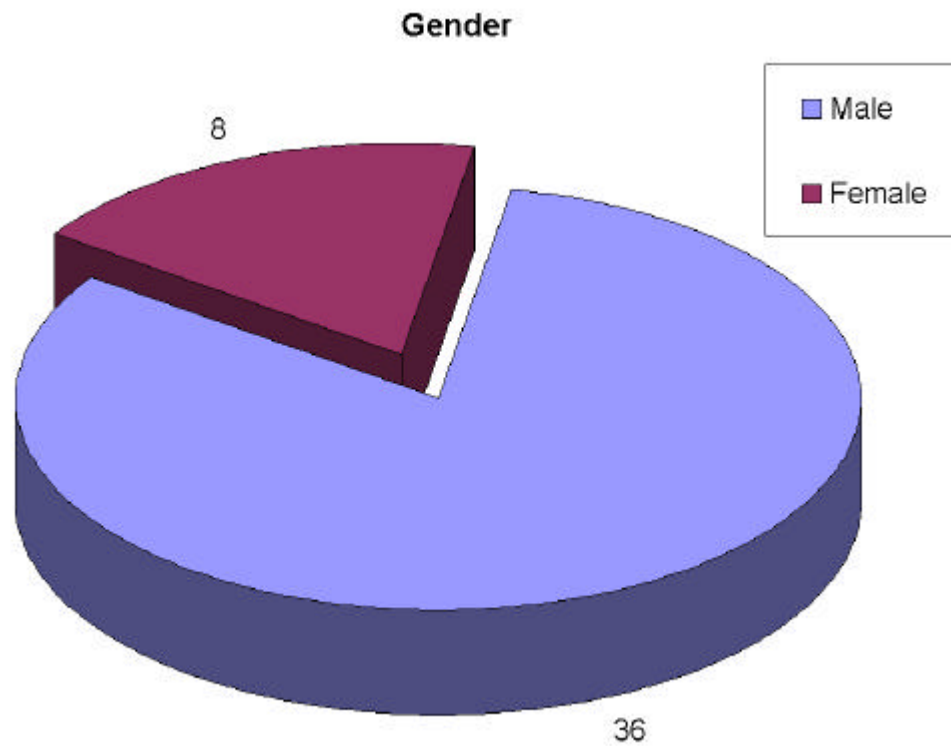


Figure 1. Statistical information on participants: Gender.

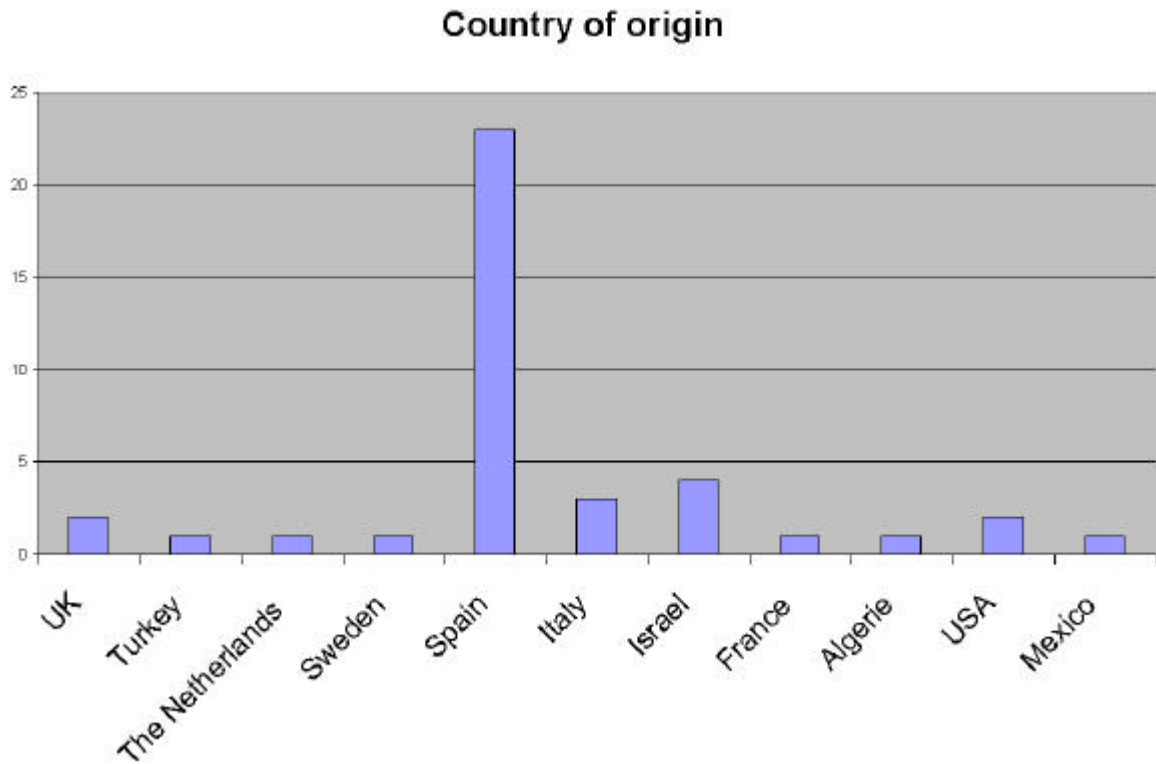


Figure 2. Statistical information on participants: Country of origin.

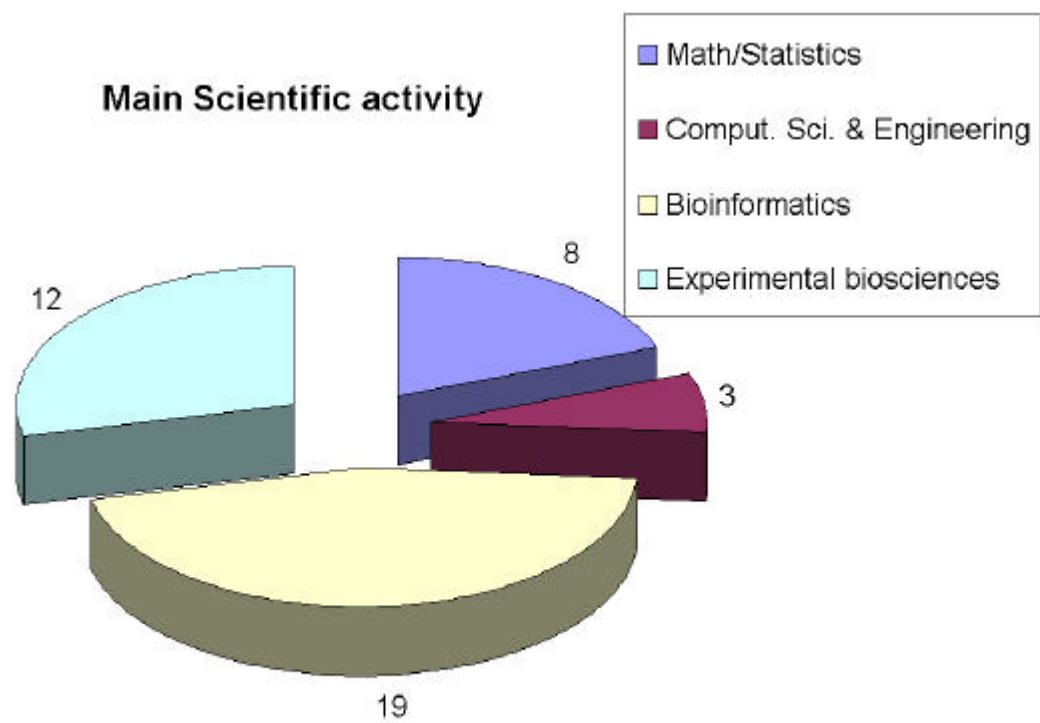


Figure 3. Statistical information on participants: Country of origin.

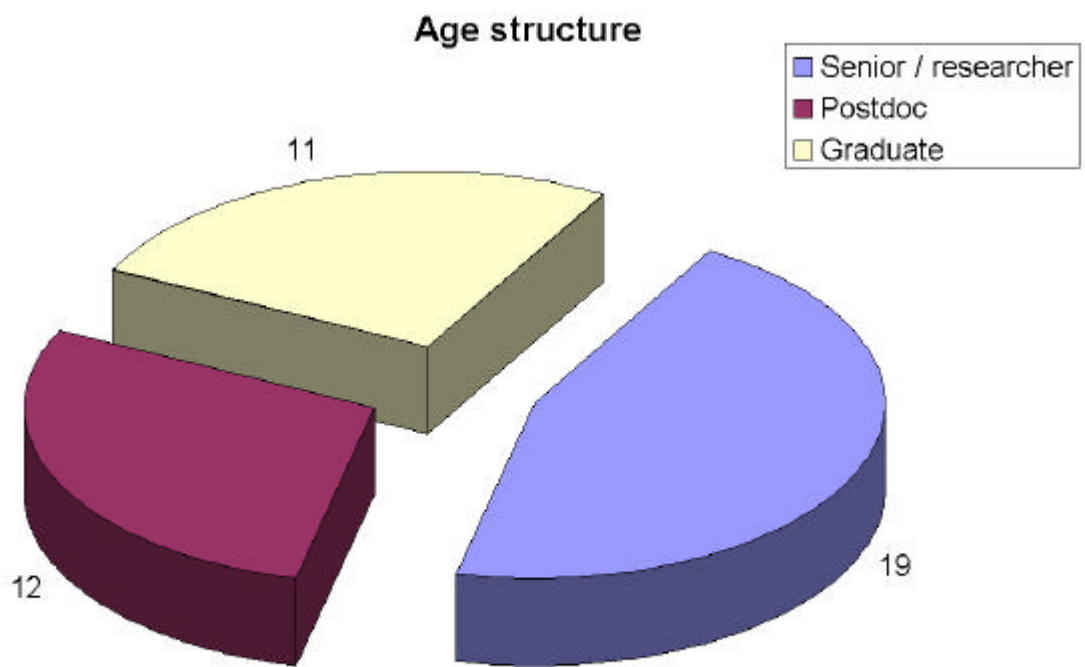


Figure 4. Statistical information on participants: Age structure.

LIST OF PARTICIPANTS
ESF WORKSHOP ON GENOMIC APPROACHES TO MICROARRAY DATA ANALYSIS

	NAME	SURNAME	POSITION	INSTITUTION	DEPARTMENT	ADDRESS	CITY	COUNTRY	TELEPHONE	FAX	E-MAIL
	Jesús	Aguilar-Ruiz	Research Scientist	University of Seville	Computer Science	ETS Ingeniería Informática	Seville	Spain	34 954 55 38 71	34 954 55 71 39	aguilar@si.us.es
	Andrés	Alonso-Fernández	Postdoc	Universidad Autónoma de Madrid	Dpto. Matemáticas	Campus de Cantoblanco	Madrid	Spain	34 91 396 76 44	34 91 397 48 89	andres.alonso@uam.es
	Fátima	Al-Shahrou		CNIO	Bioinformática	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	91 224 69 00	91 224 69 72	falshahrou@cnio.es
GRANT	Diego	Arcelli	Postdoc	Università di Ferrara	Functional Genomics Lab	Via Fossato di Mortara 64/b	Ferrara	Italy	39 0532291714	39 0532291533	diego.arcelli@unife.it
GRANT	Mukesh	Bansal	Graduate Student	ITC-irst	ECTRL	Via Ghiaie 20/23 - 38100 Trento	Trento	Italy	39-33.35.02.50.57		mukesh_iitk@yahoo.com
	Yoseph	Barash		Hebrew University	Computer Science	91904 - Jerusalem	Jerusalem	Israel	+972-2-658.4881/6757	+972-2-5617723	hoan@cs.huji.ac.il
	Alvis	Brazma		EMBL Outstation -- Hinxton		European Bioinformatics Institute - Wellcome Trust Genome Campus - Hinxton, Cambridge CB10 1SD	Cambridge	UK	+44-(0)1223 494658	+44-(0)1223 494468	brazma@ebi.ac.uk
	Yoav	Benjamini		Tel Aviv University	Department of Statistics		Tel Aviv	Israel			ybenia@post.tau.ac.il
	Mario	Cardaba Arranz		CNE	CNE	Pza. Rosa Chacel, 4	Valladolid	Spain	620 547 622		mcardaba@isciii.es
	Lucía	Conde		CNIO	Bioinformática	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	91 224 69 00	91 224 69 72	lconde@cnio.es
	Ana	Conesa	Postdoc	IVIA	Centro de Genómica	Carretera Moncada - Náquera	Valencia	Spain	34 96 342 40 00	34 96 342 40 01	estecrepusculo@hotmail.com
	Javier	de las Rivas		Centro de Investigación del Cáncer CIC		Universidad de Salamanca - CSIC, Campus de Miguel de Unamuno - 37007 - Salamanca	Salamanca	Spain	34 923 29 47 20 / 620 29 17 37	34 923 29 47 43	jrivas@usal.es
	Ramón	Díaz		CNIO	Bioinformática	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	91 224 69 00	91 224 69 72	rdiaz@cnio.es
	Joaquín	Dopazo		CNIO	Bioinformática	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	91 224 69 00	91 224 69 72	jdopazo@cnio.es
	Sandrine	Dudoit		University of California at Berkeley	Division of Biostatistics		California	USA			sandrine@stat.berkeley.edu
GRANT	Francesco	Falciani	Research Director	University of Birmingham / School of Biosciences	ImmunoGenomics	Edgbaston	Birmingham	UK	44 (0) 121 414 30387		f.falciani@bham.ac.uk
	Javier	Galán		CNIO		Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	34 91 224 69 00		jgalan@cnio.es
	Juan Fernando	García	Research Scientist	CNIO	Lymphomas Unit	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	34 91 224 69 08	34 91 224 69 23	jfgarcia@cnio.es
	Javier	Herrero		CNIO	Bioinformatics Unit	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	+34 91 224 69 00	+34 91 224 69 72	jherrero@cnio.es
GRANT	Iñaki	Inza	Postdoc	University of the Basque Country	Department of Computer Science and Intelligent Systems Group	P.O. Box 649 - 20080 Donostia	Donostia-San Sebastián	Spain	34 943 01 50 26	34 943 21 93 06	ccpincal@si.ehu.es
	Patrick	Kemmeren		UMC Utrecht	Genomics Lab	Postbus 85060 - 3508 AB	Utrecht	The Netherlands	+31-30-2538959	+31-30-2539035	P.P.C.W.kemmeren@med.uu.nl
	Jan	Komorowski	Director	Uppsala University and The Swedish University of Agricultural Sciences	The Linnaeus Centre for Bioinformatics	BMC, Box 598 - Husargatan 3 - SE-751 24 Uppsala	Uppsala	SWEDEN	+46 18 471 66 92	+46 18 471 66 98	Jan.Komorowski@lcb.uu.se
	Luis	Lombardía	Research Scientist	CNIO	Genomic Analysis Unit	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	34 91 224 69 00	34 91 224 69 72	llombardia@cnio.es
	Luis	López Fernández		Centro Nacional de Biotecnología	Department of Immunology and Oncology	Campus Universidad Autónoma de Madrid	Madrid	Spain	34 91 585 49 18	34 91 585 45 06	llfdez@cnb.uam.es
	José Ramón	Macías González				Pilar de Zaragoza, 47 - 6° A	Madrid	Spain	667 95 99 11		jr_macias@terra.es
	Alvaro	Mateos Gil		CNIO	Bioinformática	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	34 91 224 69 00	91 224 69 72	amateos@cnio.es
	Pedro Pablo	Medina Vico	Graduate Student	CNIO	Cáncer de Pulmón	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	34 91 224 69 00		pmolina@cnio.es
	Hanafi	Menouar		University of Sciences and Technology			Oran	Algeria		213.41.39.84.35	hanafi951@yahoo.com
	Yves	Moreau		ESAT-SISTA/COSIC/DocArch				France			moreau@esat.kuleuven.ac.be
	Juan Carlos	Oliveros		Alma Bioinformática							oliveros@almabioinfo.com
	Marina	Pollán	Research Scientist	Instituto de Salud Carlos III	Centro Nacional de Epidemiología	Sinesio Delgado, 6 - 28029 Madrid	Madrid	Spain	34 91 822 26 35	34 91 387 78 15	mpollan@isciii.es

	NAME	SURNAME	POSITION	INSTITUTION	DEPARTMENT	ADDRESS	CITY	COUNTRY	TELEPHONE	FAX	E-MAIL
GRANT	Anat	Reiner	Graduate Student	Tel-Aviv University	Yoav Benjamini's Statistics in Genetics Lab	69978	Tel-Aviv	Israel	972-3-640 53 86	972-3-640 93 57	anatr@post.tau.ac.il
	Juliana	Salazar		Facultat de Medicina i Ciències de Salut	Unitat de Recerca de Lípids	Universitat Rovira i Virgili - c/ San Llorenç, 21 - 43201 Reus	Reus	Spain	977.75.93.18	977.75.93.22	iusb@fmcs.urv.es
	Javier	Santoyo		CNIO	Bioinformática	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	91 224 69 00	91 224 69 72	jsantoyo@cnio.es
	Tomas	Schlitt		EMBL-EBI Hinxton		Wellcome Trust Genome Campus - CB10 1SD Cambridge		UK	+44-1223-494651		schlitt@ebi.ac.uk
GRANT	Osman	Sezerman		Sabanci University	Computational Biology	Sabanci University MDBF 34956	Istanbul	Turkey	90-216-483 95 13	90-216-483 95 50	uqur@sabanciuniv.edu
	Ron	Shamir		School of Computer Science		Tel Aviv University - 69978 Tel Aviv	Tel Aviv	Israel	+972-3-640-5383	+972-3-640-5384	rshamir@tau.ac.il
	Javier	Tamames		Alma Bioinformática							tamames@almabioinfo.com
	Amada	Torres	Graduate Student	IMSS-Mexico	Central de Instrumentos	Av. Cuauhtemoc, 3 30	Mexico D.F.	Mexico			torresa19@terra.com.mx
	Alfonso	Valencia		CNB-CSIC		Campus Universidad Autónoma de Madrid	Madrid	Spain	+34 91 5854570	+34 91 585 4506	valencia@cnb.uam.es
GRANT	Giorgio	Valentini	Postdoc	University of Milan	DSI - Computer Science Dept.	Via Comelico 39	Milano	Italy	39 02 50 31 62 55	39 02 50 31 63 73	valentini@dsi.unimi.it
	Mark	Van Der Laan		University of California at Berkeley - School of Public Health	Division of Biostatistics		California	USA			laan@stat.berkeley.edu
	Juan Manuel	Vaquerizas		CNIO	Bioinformática	Melchor Fernández Almagro, 3 - 28029 Madrid	Madrid	Spain	91 224 69 00	91 224 69 72	ivaquerizas@cnio.es
	Ms. Anja	Von Heydebreck		Max Planck Institute for Molecular Genetics	Computational Molecular Biology (Abt. Vingron)	lnestr. 73 - 14195 Berlin	Berlin	Germany	+49-30-8413-1168	+49-30-8413-1152	Anja.Heydebreck@molgen.mpg.de