

Figure 1. An example of an analytic tool to facilitate the identification of differentially regulated spots on 2D gels: The top figure shows the location of all detected proteins on the gel (circle centers) separated by the pI (x-axis) and the mass (y-axis) of the protein, and the statistical significance of the protein specific differential regulation, comparing the treatment groups of interest. The diameter of the circle is proportional to the absolute value of the protein specific test statistic, while the sign (positive/negative) of the test statistic is given by the colors of the circles (red/blue), indicating up/down regulation. The lower panel highlights the proteins considered significant in this application, and shows the protein ID number.

Reproducible Research: The study of gene – environment interactions is enhanced by the integration of basic science, clinical and population methods. At each level, the acquisition and analysis of complex data is central to drawing valid and reproducible conclusions. Their integration adds an order of magnitude of complexity. Hence, the success of this and other centers will be enhanced by new informatics tools that enable investigators to share their data, methods and results as part of the publication process.

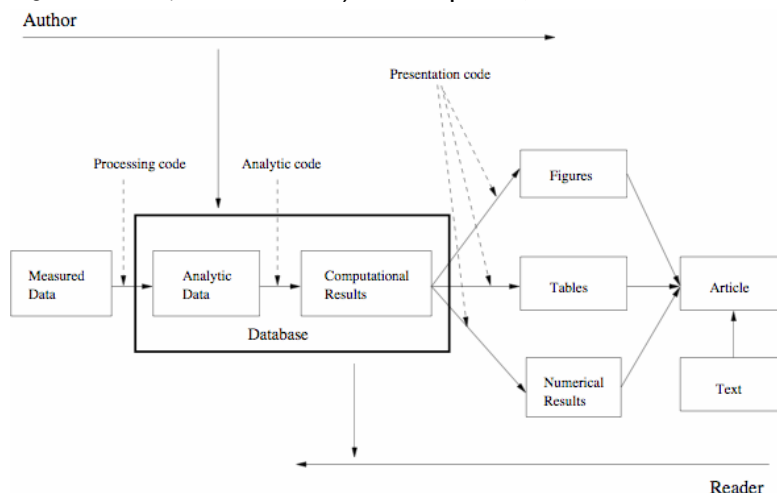


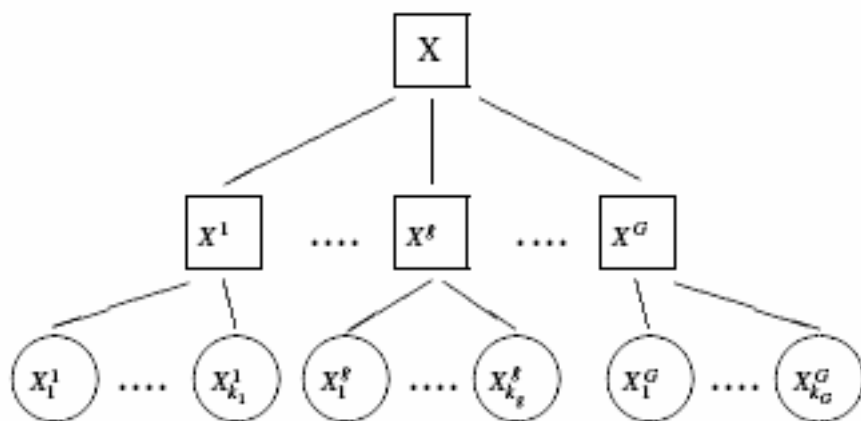
Figure 2. Schematic of the flow for the author (left to right) and reader (right to left) of a reproducible research document with indications for necessary processing, analysis and presentation software.

The ability to make scientific findings reproducible is increasingly important in areas where substantive results are the product of complex statistical computations. Reproducibility allows others to verify the published findings and conduct alternative analyses of the same data. Peng, et al. (2006) established a basic standard for evaluating the reproducibility of published epidemiological research and demonstrated how the standard can be met in a large observational study of the health effects of air pollution.

Figure 2 above is a schematic of a framework for reproducible research. We are currently building modular tools to automate the process of storing the discrete components depicted in the research pipeline. Tools are necessary both for scientists disseminating their research findings and for readers who want to dig deeper into published results. These tools are based on literate programming techniques that mix computer code for statistical analyses with the text of an article or report. The result is a single file containing the necessary code and text for reproducing the computational results and the finished scientific article. We have developed the *stashR* and *cacheSweave* add-on packages for the R software environment that read these mixed code/text files and cache the results of relevant computations in a database (Eckel and Peng, 2006). Once the analytic data and computational results are cached, they can be made available so that others can reproduce the published results or conduct alternative analyses. These packages are now available from <http://www.biostat.jhsph.edu/~rpeng/software/>.

In order to support the dissemination of Center research results, we will develop additional software for constructing “interactive documents” so that readers may explore the data and computations used to produce a scientific result. These software tools will allow readers to automatically download cached computational results from central servers so that they may have immediate access to the data and code corresponding to a published article. This suite of software packages will help to establish a closer connection between the author and reader of an article and has the potential to raise the level of scientific debate by simplifying the process of making all research reproducible.

Particle Epidemiology: We plan to develop new statistical tools for characterizing the toxicity of the PM complex mixture on mortality and cardio-respiratory hospitalizations and deaths. By identifying the toxicity of the PM chemical components, we will test existing and guide development of future hypotheses on biological mechanisms of action. Hypotheses generated from epidemiologic evidence can be tested in experimental models and vice versa. We will apply the proposed methods to national data sets for estimating acute and chronic effects of PM size, chemical components, and emission sources on mortality and morbidity. Specifically we plan to develop and apply regression models with hierarchical predictors. More than fifty correlated chemical components of PM are now monitored by hundreds of stations in US. A critical question is whether a small subset of these chemical constituents, alone or through their interactions, is responsible for the harmful effects of PM. The statistical problem is that a main predictor X (e.g. PM_{2.5} total mass) is composed of multiple chemical constituents. These constituents are then organized in a hierarchy that can be depicted as a tree (see Figure 3).



Groups can be defined according to shared elemental characteristics or according to prior knowledge about their sources (Hopke et al 2006). The groups may be overlapping so that certain constituents appear in more than one group. For example, the constituents in the first group might be related to coal combustion, group 2 might be related to diesel emission, and group 3 related to vehicle emissions. In this case, there would be shared constituents between the three groups (Bell et al 2007). Table 1 summarizes an evidence-based tree

that identifies chemical constituents that share the same emission sources. This tree has been constructed from an extensive EPA database on PM source profiling (Hsu, 2007). Note that the PM2.5 chemical constituents listed in Table 1 provides more 90% of the total PM2.5 mass.

		% contribution to PM2.5 total mass		
		average	max	min
Biomass burning	Organic Carbon (OC)	22.852	88.838	1.299
	Elemental Carbon (EC)	8.299	50.416	0.000
	Chlorine atom	8.743	17.670	0.000
	Potassium	5.676	20.660	0.000
Coal Combustion	Sulfate (SO4=)	20.820	71.923	2.120
	Silicon (Si)	14.293	40.031	1.525
	Organic Carbon (OC)	9.698	70.272	0.000
	Aluminum (Al)	8.750	23.103	0.000
	Calcium (Ca)	7.899	34.257	0.146
	Sulfur (S) ***	5.664	40.933	0.199
	Iron (Fe)	5.326	15.667	0.721
	Ammonium (NH4+)	2.104	10.889	0.059
Diesel	Elemental Carbon (EC)	33.399	80.751	0.054
	Organic Carbon (OC)	26.276	91.056	1.223
	Sulfate (SO4=)	1.230	6.668	0.146
Oil Combustion	Sulfate (SO4=)	17.701	34.596	0.000
	Sulfur (S) ***	16.703	103.550	2.942
	Organic Carbon (OC)	10.088	65.445	0.040
	Elemental Carbon (EC)	7.252	42.997	0.000
	Nickel (Ni)	2.777	22.330	0.001
Vehicles	Organic Carbon (OC)	31.736	90.215	0.000
	Elemental Carbon (EC)	24.738	89.880	0.000
	Nitrate (NO3-)	4.211	37.338	0.000
	Sulfate (SO4=)	4.033	24.374	0.000
Crustal	Silicon (Si)	24.206	55.056	1.260
	Calcium (Ca)	8.209	39.930	0.670
	Aluminum (Al)	6.403	15.947	0.000
	Organic Carbon (OC)	4.485	33.186	0.000
	Iron (Fe)	4.263	11.835	0.150

Table 1: Evidence-based tree: emission sources, PM2.5 chemical constituents, their corresponding % contribution to the PM2.5 mass. N indicates number of source profiles, that is the number of sampling experiments that have measured chemical constituents of PM2.5 and have attributed that constituents to emission sources. This evidence-based tree has been constructed by Michelle Bell (as a personal communication) by using the database

<http://oaspub.epa.gov/eims/eimsapi.dispdetail?deid=161519> and a recent EPA documentation report on speciation data (Hsu, 2007)

We are currently developing extensions of regression analysis that take account of the hierarchical structure of the predictor variables to identify individual components or groups of components whose influence on health outcomes is greatest.

E. LOCATION

All BISC faculty except for Dr. Curriero are located on the 3rd floor of the new wings of the School of Public Health proximal to the Department of Environmental Health Sciences, which has laboratories and offices on the 7th floors where Dr. Curriero is located. This has in the past and will continue to facilitate daily informal interactions between BISC and Center faculty.

F. EQUIPMENT

Computing — BISC has a state-of-the-art high-performance computing environment comprising a central cluster, multiple web-servers, desk-top workstations, local area network, high-speed connection to the internet and all major bioinformatics, statistics and mathematical software essential to modern environmental research. This computing environment has been created over the last three years including an expenditure of \$350,000 by the Department of Biostatistics to purchase its current cluster. The department also provides core funding for two computer scientists who maintain the environment.

In addition to desktop computing for its faculty, the BISC shares a high performance Linux computing facility comprising: (1) a 68-cpu AMD Opteron-based computing cluster with 128GB DDR-SDRAM and 2.7TB local disk storage capable of running at 170 gFLOPS; (2) a 4-cpu Opteron-based computing server with 16GB DDR-SDRAM and 36GB local disk storage; and (3) a 2-cpu Pentium-4-based database server with 8GB DDR-SDRAM and 36GB local disk storage. The heart of the system is a Global File System (GFS) that handles 4 terabytes of data over a fiber-channel storage array network (SAN). GFS allows users to seamlessly access their home directories from any of the computing engines that are currently attached to the SAN.

This system is situated behind the School of Public Health firewall ("blue network") for data security. The individual machines are accessible from outside the firewall through an SSH cut-through on the Sun Fire computing server. Outside of the School's firewall ("black network"), the Department of Biostatistics maintains three web servers for serving web pages and hosting individual faculty projects. These servers will be used for BISC projects that depend upon regular public interaction.

The computing environment is supported by two systems specialists: Marvin Newhouse and Jiong Yang who are certified and experienced in both Linux and Solaris. Statistical and mathematics packages include R, S-PLUS, SAS, Matlab, Mathematica and Bioconductor. SAS and SQL are the dominant package used in data management, data editing and updating.

The statistical computing environment described above is strategically directed by associate professor Fernando Pineda, a physicist who now conducts genomic research. He chairs the Bioinformatics/Biostatistics Information Technologies (BIT) Committee of faculty, staff and student representatives. The implementation of their policies is managed by Ms. Cindy Hockett, the Department of Biostatistics administrator. The BIT Committee website (<http://biosun01.biostat.jhsph.edu/~ririzarr/BIT/>) is used to continually educate BISC and other faculty on recent developments in statistical computing and their implementation on our system.

The School Information Systems group (<http://www.jhsph.edu/is/default.html>) is responsible for all networking within the Center and School as well as our internet connections to the outside. IS comprises a staff of more than 30 full-time computer scientists who provide back-up support for Mr. Newhouse and Yang and the BISC

Principal Investigator/Program Director (Last, First, Middle): Groopman, John D. – Bioinformatics/Biostatistics Core faculty.

Web-site – The Department of Biostatistics maintains an active web-site (<http://www.biostat.jhsph.edu/>) that is used as an information center and tool for disseminating our recent research, software and other research products. Ms. Mary Joy Argo is the site manager. She will also support BISC web-pages as needed. The cost of her BISC support is covered by the Department of Biostatistics.

Statistics Library — The Department of Biostatistics provides a library with more than 30 bioinformatics and biostatistics journals and 3,000 reference and textbooks. It provides online access to most major reference tools including for example, PubMed, Lexus/Nexus and the Current Index of Statistics and to the other major indices through the Welch Medical Library with more than 3,000 journals and 300,000 books, located directly across the street.

Institutional Commitment — The School of Public Health currently spends \$6 million per year of general funds to maintain the information systems infrastructure on which BISC relies for its local area network and connection to the internet. The Department of Biostatistics spends roughly \$250,000 per year, 40% of which derive from School funds, to support the computing environment described above. It also expended an additional \$350,000 over the last 3 years to upgrade this environment to its current level. BISC is further supported as described above by Ms. Mary Joy Argo, Mr. Jiong Wang and Mr. Marvin Newhouse through funds provided by the School.

G. BENEFITS TO CENTER INVESTIGATORS

Environmental health research relies on the integration of molecular, clinical and population sciences. It is increasingly dependent on the acquisition, management and use of complex information. Hence, statistical and informatics methods are essential for success. Today, Center scientists routinely rely on bioinformatics, biostatistics, mathematics, and computing tools to design, implement, and analyze data from their experiments and observational studies. BISC introduces Center members to the optimal tools when they exist and create new ones when they do not. Center members are benefited when their research program is made more productive through their consultations and collaborations with BISC faculty. They produce more knowledge from the same investment of resources when they use optimal quantitative methods.

BISC methods research also benefits Center investigators indirectly. When our novel methods such as spatial time series models or gene expression measures are adopted by the larger research community, progress is more rapid and Center members benefit from the discoveries of others that are facilitated by BISC generated tools.

H. POLICIES FOR OPERATING

• *Personnel*

Scott L. Zeger, PhD, Director
Frank Curriero, PhD
Francesca Dominici, PhD
Rafael Irizzary, PhD
Roger Peng, PhD
Ingo Ruczinski, PhD

• *Policies*

Personnel — Dr. Scott Zeger will serve as Core Director at 7% effort and will be responsible for matching Biostatistics Facility Core members to individual scientists who request support through the Core. He will also be responsible for providing statistical input to environmental epidemiology research activities. Drs. Dominici and Peng will work most closely with Center investigators conducting epidemiologic and clinical research. Dr. Dominici will

Principal Investigator/Program Director (Last, First, Middle): Groopman, John D. – Bioinformatics/Biostatistics Core also lead the working group on Environmental Biostatistics as discussed below. Drs. Irizzary and Ruczinski will collaborate with Center members conducting molecular (e.g. genomic) and clinical research and will lead the Environmental Bioinformatics working group. Dr. Curierro will provide general support for any center faculty member with a shorter-term statistical issue and also be involved in longer-term collaborations involving GIS and other spatial or time series applications.

Collaboration — To promote effective consultations and collaborations among environmental and bioinformatics/biostatistics scientists, we will rely on two strategies. First, many of the environmental scientists and statisticians encounter one another daily. The Biostatistics Core members will make themselves available for informal consultations through these daily contacts around the school. Such impromptu discussions often meet an investigator's needs and foster ongoing dialogue. For those investigators with less daily contact, they will secure Core services by contacting Dr. Zeger's office either by phone, fax, or email. Dr. Zeger's assistant will then schedule an initial meeting between the environmental scientist and the appropriate member of the BISC.

Longer-term collaborations typically start through shorter consultations. Collaborations will be supported through the bi-weekly meeting of the BISC faculty and Center members in the Environmental Biostatistics and Epidemiology Working Group (EBEG) as described further below.

Methods Research — Two existing working groups will be used to promote collaborations among environmental and BISC faculty as well as the development and dissemination of new methods and software for environmental research. The first is the Environmental Biostatistics and Epidemiology working group (EBEG) that currently meets twice per month (<http://www.biostat.jhsph.edu/bstproj/ebeg/>). It is organized by Dr. Dominici. The second is the Bioinformatics working group organized by Drs. Rafael Irizzary and Giovanni Parmigiani (<http://astor.som.jhmi.edu/hex//index.html>). This group meets weekly. Meetings of both groups include presentations of work in progress, discussions of recent important papers, and presentations of new problems by environmental and other scientists using bioinformatics/biostatistics tools.

- *Cost*

In the absence of a Biostatistics Core, Center investigators would be required to make individual arrangements for statistical collaboration on an as-needed basis. It is unlikely that the BISC members would be available to many of the Center scientists, since they would be committed to the support of other investigations. By securing a small fraction of the effort of these bioinformatics and biostatistics experts for the use of all environmental scientists at the school and by promoting research on improved designs and methods, a larger group of environmental scientists will have access to state-of-the-art methods for their research.

It is difficult to quantify the cost savings of collaboration with statistical experts; however, a few points are relevant. First, the efficiency of a study designed jointly by an environmental and statistical scientist will likely be greater in some cases than what could be achieved by the environmental scientist alone. It is not unreasonable to assume that an average efficiency gain of at least 10% would result from their interaction. Hence, the same quality of information will be collected for 10% lower costs. Second, substantial gains will result from the avoidance of serious design and analysis flaws, which can lead to incorrect scientific conclusions. The opportunity costs to the original and new research teams of pursuing false leads that result from poor designs, analyses or interpretations of environmental data can be enormous.

I. USAGE OF RESOURCES

Environmental Epidemiology

NMMAPS: Over the last five years, the BISC faculty continued collaborations and consultations with several Center members to address health issues in urban populations. Drs. Zeger, Dominici and Peng collaborated with Drs. Samet, Breyse, Geyse and Buckley to complete four studies of the health effects of particulate air pollution in U. S., Canadian and European cities and to start three new ones. The four completed studies all relied upon the National Morbidity and Mortality Air Pollution Study (NMMAPS-funded by Health Effects Institute, NIEHS, EPA)

Principal Investigator/Program Director (Last, First, Middle): Groopman, John D. – Bioinformatics/Biostatistics Core database that comprises daily mortality, air pollution and weather data on the 100 largest cities in the U.S. and on nearly 50 cities from Canada and Europe. Our group built an on-line repository for the U.S. data through the internet Health and Air Pollution Surveillance Study (iHAPSS) funded by the Health Effects Institute, but initiated by the BISC core. The data is available as an R module from the iHAPSS web-site (<http://www.ihapss.jhsph.edu/>). Software to conduct basic spatial time series analyses of these data is also available there. These methods have been adopted after modification to conduct an international version of the NMMAPS study, the APHENA project (funded by HEI) involving cities from the U.S., Canada and Europe.

The NMMAPS study has produced significant public health findings and has created novel methods for risk assessment. Our collaboration has produced estimates a national and regional relative rates of mortality associated with a change in PM 10, PM2.5 and ozone for the largest U.S. These results have been relied on by the U.S. EPA to develop new guidelines for the regulation of these pollutants. A complete reference list is given in Section I.

MCAPS: The Center has made it possible for us to compete for six new grants from the NIH, EPA and Health Effects Institute over the last 5 years; three have already been funded; 3 are pending review. We have received EPA funding to create the Johns Hopkins Particle Research Center that has supported the Medicare Cohort Air Pollution Study or MCAPS. As described above, we have matched roughly 12 of the 50 million Medicare enrollees to air pollution monitors nearby and to weather and zip-level SES data. The MCAPS database is now an important BISC and Center resource that is already producing key findings. For example, Dominici, Peng, Bell, et al (2006) demonstrated the differential affect of PM2.5 exposure on rates of hospitalization for cardiovascular and respiratory causes in the eastern and western U.S. generating interesting hypotheses about the possible toxic constituents of particles. In addition to the EPA Particle Center grant, Dominici, Peng, Zeger and Samet have successfully funded an NIEHS grant to build novel methods to exploit the MCAPS and other similar datasets. They are also key personnel in a new NIEHS training grant (PI, Tom Louis) to support PhD candidates in biostatistics.

Relying on the MCAPS data, we have submitted two other grants to the EPA to investigate the health effects of coarse particles (PM10-PM2.5, PI: Francesca Dominici) and to quantify the total health effects caused by air pollution (EPA, PI: Francesca Dominici). Finally, Roger Peng has submitted an HEI Rosenblith proposal to develop methods for making the MCAPS studies reproducible as discussed in Section D.

In developing the MCAPS database, Dr. Dominici and other BISC members developed expertise in working with the Center for Medicare and Medicaid Services (MCS), the Johns Hopkins IRBs and other administrative tasks necessary to use Medicare data to address environmental questions. They have established the Collaboration on Health Information, Computing and Statistics (CHICAS) to facilitate the use of these data by other Johns Hopkins scientists who can avoid the substantial start-up costs of acquiring and managing the data, obtaining permissions and developing statistical and database tools to do the appropriate analyses. Several new environmental and other projects are underway. For example, Pulmonary Medicine fellow, Dr. Emily Smith and colleagues have created a subset of the Medicare population with chronic diabetes and are estimating the health effects of exposure to air pollution in this at-risk population.

Methodology: In the course of their collaborative research, BISC members have developed new statistical methods with important application to Center projects and beyond. For example, in the area of spatial time series analysis, Dominici et al (2004) developed methods for choosing the optimal degree of adjustment for potential confounding by time-varying factors. These are now regularly used in our MCAPS analyses. Welty, et al (2007) developed a novel method for specifying and estimating distributed lag models to determine the total affect of acute exposure rather than the effect from a single days exposure. Peng, et al (2007) have extended this method to the hierarchical case where we have many time series from cities around the country. Lu and Zeger (2007) have demonstrated the equivalence of time series and case-cross-over methods for estimating pollution effects. Dominici et al (2005) developed the method of smooth quantile ratio estimation (SQUARE) to estimate the costs of medical services that are attributable to an environmental exposure such as smoking or air pollution. Nearly all of these methods are implemented in standard software (e.g. R) and are available from the BISC investigators. In the next period, we will consolidate them into a single location making access more immediate.

Genomics

What was previously was a Biostatistics Core became a Biostatistics and Bioinformatics Core during the last cycle. Rafael Irizarry and later Ingo Ruczinski joined the BISC and have dramatically expanded our ability to provide support for the measurement, management and analysis of genomic data. In the last five years they have made important advances that have benefited the Center investigators and others beyond.

Measuring gene expression: Affymetrix introduced its original measure of expression that was a contrast between the degree of binding of perfect oligonucleotides relative to the binding of “mismatched” ones in which a middle base pair was changed. The idea was to use the mismatch to control for non-specific binding. While correcting for some bias due to non-specific binding, the mismatch data added substantial noise making it difficult to detect fold changes in genes with moderate to low abundance. Irizarry and colleagues discovered this problem and, in two 2003 papers (Irizarry, et al, 2003a,b) that are among the most cited in the recent biomedical literature, proposed RMA an alternate measure of gene expression that is now an industry standard. Center scientists were early adopters of the improved approach to measurement and benefited directly.

Bioconductor: Irizarry is one of the leaders of a group of statisticians and bioinformaticians who, during the last 5 years, have develop Bioconductor, a computing environment for the manipulation and analysis of genomic data (Gentleman, et al, 2004). Bioconductor (<http://www.bioconductor.org/>) is now a leading data analysis tool found in many genomics labs around the world. In addition to developing these tools, Irizarry, Ruczinski and their colleagues have presented 5 Hopkins workshops on the use of Bioconductor and on principles for the analysis of genomic data. They have done two one-day workshops and 3 lectures series. Approximately 500 Johns Hopkins scientists have participated in the series. Irizarry has just received a 3rd percentile score for his NIH grant to develop additional tools for genomic data analysis to expand the utility of Bioconductor.

Single nucleotide polymorphism (SNP) data analysis: Both Irizarry and Ruczinski have made significant contributions to the measurement and analysis of high throughput SNP data. Irizarry and colleagues (Carvahlo, et al, 2007) have used ideas like those that led to RMA to develop an alternate way to “call” base pairs from SNP chips. It performs much better than the commercial methods. Ruczinski and colleagues have developed a methodology called “logic regression” (Ruczinski, et al, 2003; Scharpf, et al, 2007) that finds a small number of Boolean combinations (and/or) of the SNPs that optimally predict a phenotype. With 100,000 SNPs, there are $2^{100,000}$ possible models; their technique efficiently search among subsets of these models for ones that optimally predict the phenotype.

Epigenetics: With support from the Center, BISC member Irizarry has developed a group of statisticians including Ingo Ruczinski to research statistical methods for epigenetic data. Professor Andrew Feinberg of the Johns Hopkins Department of Medicine is leading a multi-disciplinary team that includes Irizarry’s statistical core in a new study for the Epigenetics of Common Human Disease. It recently received a \$5 million, five-year grant from the National Human Genome Research Institute and the National Institute of Mental Health to develop methods and to begin systematically examining the epigenetics of autism and bipolar disorder. The biostatistics methods core will develop and broadly disseminate methods for epigenetic analysis. Given the potential centrality of epigenetics to the study of gene-environment interaction, these methods are likely to be important to Center members.

NIGMS Training Grant on Bioinformatics: During the past cycle, the Center facilitated our ability to respond to a new NIH initiative to train more bioinformaticians. We successfully competed for a new bioinformatics training grant that is directed by Dr. Giovanni Parmigiani of Biostatistics and Oncology. Drs. Zeger, Irizarry and Ruczinski are core members of the training faculty. The program supports five pre-doctoral trainees at a time. The first PhD graduate, Rob Scharpf has just finished and is a lead author with Dr. Ruczinski on a SNP data analysis paper. There is also a NIEHS-funded training grant in Environmental Biostatistics.

J. FUTURE PLANS

The proposed BISC represents a near doubling of our expertise and effort from the previous period. Five years ago, we added Rafael Irizarry to strengthen our bioinformatics component. His influence on the Center in the measurement and analysis of genomic data prompted us to add Ingo Ruczinski who brings additional experience in proteomics and SNP data analysis. We have now added Dr. Roger Peng whose database and literate programming skills complement the genomics group and better integrate it with the environmental epidemiology members of BISC.

Our specific aims clearly state our intention to continue consultation and collaboration with Center members to ensure that their research programs take full advantage of modern bioinformatics and biostatistics methods and expertise. We will continue the very successful strategy of promoting informal “coffee-pot” discussions so that problems get handled naturally in the course of each day. This is facilitated by the decision of BISC member Dr. Frank Curriero to move his office and primary appointment to Environmental Health Science the 7th floor. We will supplement the regular informal contacts with more formal consulting meetings whenever needed. Collaboration meetings are naturally scheduled by the team members around specific projects. BISC members will keep abreast of each other’s progress through weekly or bi-weekly meetings of our two working groups and with a quarterly meeting of the BISC members to discuss policies and procedures.

Our second aim is to continue to develop original environmental science resources that facilitate Center research. In the next 5 years, we specifically propose to focus on the following topics.

1. *MCAPS*: We plan to use the Medicare Cohort Air Pollution Study (MCAPS) database to identify constituents of particulate pollution whose control will most reduce the burden of attributable disease and death. We also plan to promote the use of the MCAPS resource by other Center members who are pursuing other environmental questions.

2. *Regression for spatial time series data with hierarchical predictors*: The MCAPS database and the constituent question have motivated the development of new regression methods that are designed for the situation where the predictor variables form a hierarchy as do particulate components. We plan to develop such methods and disseminate them widely so that others can use and improve upon them.

3. *Reproducible Research*: We plan to implement new software and databases so that Center research results can be reproduced by others and can stimulate additional analysis and discussion to speed the rate at which environmental health science questions are resolved.

4. *SNP data analysis*: We plan to continue to develop better SNP measurements, pre-processing methods and methods for identifying combinations of SNPs that predict phenotypes.

5. *Statistical methods for epigenetics*: While these are early stages of this initiative, we recognize the potential centrality of epigenetic mechanisms in environmental sciences. We therefore plan to develop methods for the analysis of epigenetic data and to disseminate them using Bioconductor and short-courses within and beyond the Center.

K. EFFECT OF FACILITY ON STIMULATING SCIENTIFIC INTERACTIONS

When collaborating effectively, statisticians are cross-fertilizers. By interacting with numerous environmental scientists, they discover opportunities for the ideas of one scientist to cross-fertilize another. When they create a novel design or analytic approach to solve one kind of problem, it often applies to many other related problems. An effective statistician discovers common issues in different scientific studies and brings together investigators to find common solutions. This has been the pattern over the last five years of our core and will continue for the next period. For example, Dominici, Zeger, Peng and Curriero have developed and applied spatial time series models to estimate the relative risk of mortality and morbidity associated with air pollution in two major studies, NMMAPS and MCAPS. Drs. Irizarry and Ruczinski have developed new methods of measuring and analyzing genomic data that similarly advances the Center mission. Drs. Dominici and Curriero have interacted with COEC in their participation in the MPT-JHBSPH EnviroHealth Connections, Winter Colloquium and Summer Institutes.

L. REPRESENTATIVE PUBLICATIONS

The following publications have used the resources and facilities of the Bio-Statistics/ -Informatics Core (BISC).

References

Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, Foy C, Fuscoe J, Gao X, Gerhold DL, Gilles P, Goodsaid F, Guo X, Hackett J, Hockett RD, Ikonomi P, **Irizarry RA**, Kawasaki ES, Kaysser-Kranich T, Kerr K, Kiser G, Koch WH, Lee KY, Liu C, Liu ZL, Lucas A, Manohar CF, Miyada G, Modrusan Z, Parkes H, Puri RK, Reid L, Ryder TB, Salit M, Samaha RR, Scherf U, Sendera TJ, Setterquist RA, Shi L, Shippy R, Soriano JV, Wagar EA, Warrington JA, Williams M, Wilmer F, Wilson M, Wolber PK, Wu X, Zadro R. Nat Methods (2005) The External RNA Controls Consortium: a progress report. *Nature Methods* 2(10): 731-4.

Basu R, **Dominici F**, **Samet JM** (2005) Characterizing the Relationship Between Temperature and Cardio-Respiratory Mortality Among the Elderly U.S. Population, *Epidemiology*, 16:58-66.

Bell M. **Samet J.M.** **Dominici F** (2004) Time-Series Studies of Particulate Matter, *Annual Review of Public Health*, 25:247-280.

Bell M, **Samet JM**, McDermott A, **Zeger SL**, **Dominici F** (2004) Ozone and Mortality in 95 U.S. Urban Communities from 1987 to 2000, *Journal of the American Medical Association*, 292:2372-2378

Bell M, **Dominici F**, **Samet JM** (2005) A Meta-Analysis of Time-Series Studies of Ozone and Mortality with Comparison to the National Morbidity Mortality Air Pollution Study, *Epidemiology with discussion*, 16:436-445.

Bell M, **Peng R**, **Dominici F** (2006) The Exposure-Response Curve for Ozone and Risk of Mortality and the Adequacy of Current Ozone Regulations, *Environmental Health Perspectives*, 114: 532-536.

Bolstad BM, **Irizarry RA**, Astrand M, and Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* 19:185-193

Bolstad, BM, Collin, F, Simpson, KM, **Irizarry RA**, and Speed, TP (2004) Experimental design and low-level analysis of microarray data. *International Review of Neurobiology* 60:25-58.

Brewster AM, Jorgensen TJ, **Ruczinski I**, Huang HY, Hoffman S, Thuita L, Newschaffer C, Lunn RM, Bell D, Helzlsouer KJ (2006). *Polymorphisms of the DNA repair genes XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp): relationship to breast cancer risk and familial predisposition to breast cancer*. *Breast Cancer Research and Treatment* 95(1): 73-80.

Brown EE, Fallin D, **Ruczinski I**, Hutchinson A, Staats B, Vitale F, Lauria C, Serraino D, Rezza G, Mbisa G, Whitby D, Messina A, Goedert JJ, Chanock SJ, and the Kaposi Sarcoma Working Group (2006). Associations of classic kaposi sarcoma with common variants in genes that modulate host immunity. *Cancer Epidemiology, Biomarkers and Prevention* 15(5): 926-34.

Cappola TP, Cope L, Cernetich A, Barouch LA, Minhas, K, **Irizarry RA**, Parmigiani G, Durrani S, Lavoie T, Hoffman EP, Ye SQ, Garcia JGN, Hare JM (2003) Deficiency of different nitric oxide synthase isoforms activates divergent transcriptional programs in cardiac hypertrophy. *Physiological Genomics* 14:25-34

Carvalho B, Bengtsson H, Speed TP, **Irizarry RA** (2007) Exploration, normalization, and genotype calls of high density oligonucleotide SNP array data. *Biostatistics* (To appear)

Colantuoni C, Henry G, **Zeger S**, Pevsner J: SNOMAD (Standardization and Normalization of MicroArray Data):

Principal Investigator/Program Director (Last, First, Middle): Groopman, John D. – Bioinformatics/Biostatistics Core web-accessible gene expression data analysis. *Bioinformatics* 18(11):1540-1541, 2002.

Colantuoni C, Henry G, **Zeger S**, Pevsner J: Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques* 32(6):1316-20, 2002.

Cope LM, **Irizarry RA**, Jaffee H, Wu Z, and Speed TP (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 20: 323-331.

Cope L, Hartman SM, Golmann HWH, Tiesman JP, and **Irizarry RA** (2006) Analysis of Affymetrix GeneChip data using amplified RNA. *Biotechniques* 40:165-70.

Curriero FC, Patz JA, Rose JB, Lele S: The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948-1994. *American Journal of Public Health* 91:1194-1199, 2001.

Curriero FC, Heiner KS, **Samet JM**, **Zeger SL**, Strug L, Patz JA: Temperature and mortality in eleven cities of the eastern United States. *American Journal of Epidemiology* 155(1):80-87, 2002.

Curriero FC, Shone SM, Glass GE: Cross correlation maps: a tool for visualizing and modeling time lagged associations. *Vector Borne and Zoonotic Diseases* 5(3):267-75, 2005.

Dai J, **Ruczinski I**, LeBlanc M, Kooperberg C (2006). Imputation methods to improve inference in SNP association studies. *Genetic Epidemiology* 30(8): 690-702.

De Los Rios MA, Muralidhara BK, Wildes D, Sosnick TR, Marqusee S, Wittung-Stafshede P, Plaxco KW, **Ruczinski, I** (2006). On the precision of experimentally determined protein folding rates and ϕ -values. *Protein Science* 15(3): 553-63.

DiPietro JA, Caulfield LE, **Irizarry RA**, Chen P, Merialdi M and Zavaleta N (2006). Prenatal development of intra- and inter-individual synchrony. *Behavioral Neuroscience* 20(3):687-701.

Dominici F, McDermott A, **Zeger SL**, **Samet JM** (2003) National Maps of the Effects of Particulate Matter on Mortality: Exploring Geographical Variation, *Environmental Health Perspectives*, 111(1):39-43.

Dominici F, McDermott A, **Samet JM**, **Zeger SL** (2003) Airborne Particulate Matter and Mortality: Time-Scale Effects in Four US Cities, *American Journal of Epidemiology* (with invited commentary), 157:1055-1065.

Dominici F and Burnett R. (2003) Risk Models for Particulate Air Pollution, *Journal of Toxicology and Environmental Health*, 66:1879-1885.

Dominici F Sheppard L. Clyde M. (2003) Health effects of air pollution: A statistical review, *International Statistical Review*, 71:243-276.

Dominici F, Zanobetti A, **Zeger SL**, Schwartz J, **Samet JM** (2004) Hierarchical bivariate time-series models: a combined analysis of the effects of particulate matter on morbidity and mortality, *Biostatistics*, 5, 341-360.

Dominici F, McDermott A, Hastie T (2004) Improved semi-parametric time series models of air pollution and mortality, *Journal of the American Statistical Association*, 468:938-948.

Dominici F, McDermott A, Daniels M, **Zeger SL**, **Samet JM** (2005) Revised analysis of the National Morbidity, Mortality and Air Pollution Study: mortality among residents of 90 cities, *Journal of Toxicology and Environmental Health*, 68:1071-1092.

Dominici F, Cope L, Naiman D, **Zeger SL** (2005) Smooth quantile ratio estimation (SQUARE), *Biometrika*, 92:543-557.

Dominici F, **Zeger SL** (2005) Smooth quantile ratio estimation with regression: estimating medical expenditures for smoking attributable diseases, *Biostatistics*, 6:505-519.

Dominici F Peng D. Bell M. Pham M. McDermott A. **Zeger S.L. Samet J.M.** (2006) Fine particulates air pollution and hospital admission for cardiovascular and respiratory diseases, *Journal of the American Medical Association*, March 8, 295:1127-1135.

Eckel, S. P. and **Peng, R. D.** (2006), "Interacting with local and remote data repositories using the stashR package for R," Tech. Rep. 127, Johns Hopkins University Department of Biostatistics, <http://www.bepress.com/jhubiostat/paper127>.

Gaffney SH, **Curriero FC**, **Strickland PT**, Glass GE, Helzlsouer KJ, **Breyse PN** (2005) Influence of geographic location in modeling blood pesticide levels in a community surrounding a U.S. Environmental protection agency superfund site. *Environmental Health Perspectives* 113(12):1712-6.

Gautier L, Cope LM, Bolstad BM, and **Irizarry RA** (2004) affy - An R package for the analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 20:307-315.

Gentleman RC, Carey VJ, Bates DJ, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, **Irizarry RA**, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth GK, Tierney L, Yang YH, and Zhang J (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* 5:R80

Graczyk TK, Lewis EJ, Glass G, Dasilva AJ, Tamang L, Girouard AS, **Curriero FC**: Quantitative assessment of viable *Cryptosporidium parvum* load in commercial oysters (*Crassostrea virginica*) in the Chesapeake Bay. *Parasitology Research* 100(2):247-53, 2007.

Grigoryev DN, Ma SF, **Irizarry RA**, Ye SQ, Quackenbush J, Garcia JG (2004) Orthologous gene-expression profiling in multi-species models: search for candidate genes. *Genome Biology* 5:R34.

Grigoryev DN, Ma SF, Simon BA, **Irizarry RA**, Ye SQ, Garcia JG (2005) In vitro identification and in silico utilization of interspecies sequence similarities using GeneChip technology. *BMC Genomics* 6:62.

Hansel NN, Hilmer SC, Georas SN, Cope LM, Guo J, **Irizarry RA**, Diette GB (2005) Oligonucleotide-microarray analysis of peripheral-blood lymphocytes in severe asthma. *Journal of Lab Clinical Medicine*. May;145(5): 263-74.

Henshaw SL, **Curriero FC**, Shields TM, Glass GE, **Strickland PT**, **Breyse PN**. (2004) Geostatistics and GIS: tools for characterizing environmental contamination. *Journal of Medical Systems* 28(4):335-348.

Huang IC, Frangakis C, **Dominici F**, Diette GB, Wu A (2005) Application of a Propensity Score Approach for Risk Adjustment in Multiple Physician Group Profiling on Asthma Care, *Health Services and Research*, 40, 253-278.

Huang Yi **Dominici F** Bell M. (2005) Bayesian Hierarchical Distributed Lag Models for Summer Ozone Exposure and Cardio-Respiratory Mortality, *Environmetrics*, 16:547-562.

Fung KY, Krewski D, Burnett R, **Dominici F** (2005) Testing the Harvesting Hypothesis, *Journal of Toxicology and Environmental Health* 68:1137-1154.

Principal Investigator/Program Director (Last, First, Middle): Groopman, John D. – Bioinformatics/Biostatistics Core

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249-264

Irizarry RA, Ooi SL, Wu Z, and Boeke JD (2003) Use of mixture model in a genome-wide DNA microarray-based genetic screen for components of the NHEJ pathway in yeast. *Statistical Applications in Genetics and Molecular Biology* 2:Article 1

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* 31:e15

Irizarry RA (2004) Parameters with Musical Interpretations. *Chance* 17: 30-38. Wu Z and **Irizarry RA** (2004) Processing of oligonucleotide array data. *Nature Biotechnology* 22: 4-5.

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W (2005) Multiplelaboratory comparison of microarray platforms. *Nature Methods* 2:329-30.

Irizarry RA, Zhijin Wu, and Jaffee, H (2006) Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 22(7):789-94.

Irizarry RA, Cope LM, Wu Z (2006) Feature-level exploration of a published Affymetrix GeneChip control dataset. *Genome Biology* 7(8):404

Jackson L, Correa A, **Lees PSJ**, **Dominici F**, Stewart P, **Breyse PA**, Matanoski G. (2004) Parental Lead Exposure and Total Anomalous Pulmonary Venous Return, *Birth Defects Research*, 70:185-193.

Janes H, **Dominici F**, **Zeger SL**. Trends in Particulate Matter and Mortality in 113 U.S.Counties, 2000-2002: Evidence on the Long Term Effects of Air Pollution. *American Journal of Epidemiology* (under revision).

Johnson E, **Dominici F**, Griswold M, **Zeger SL** (2003) Disease cases and their medical costs attributable to smoking: an nalysis of the National Medical Expenditure Survey, *Journal of Econometrics*, 112(1):135-151.

Jorgensen J, Visvanathan K, **Ruczinski I**, Thuita L, Helzlsouer KJ. Breast cancer risk is not associated with polymorphic forms of Xeroderma Pigmentosum genes in a cohort of women from Washington County, Maryland. *Breast Cancer Research and Treatment* (to appear).

Katz S, **Irizarry RA**, Lin X, Tripputi M, Porter MW (2006) A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database. *BMC Bioinformatics* 7:464.

Kendziorski C, **Irizarry RA**, Chen K-S, Haag JD, Gould MN (2005) On the utility of pooling biological samples in microarray experiments. *Proceedings of the National Academy of Science* 102: 4252-4257.

Kim JH, Sherman ME, **Curriero F**, Guengerich FP, **Strickland PT**, Sutter TR (2004) Expression of cytochromes P450 1A1 and 1B1 in human lung from smokers, non-smokers, and ex-smokers. *Toxicol. Appl. Pharmacol.* 199(3): 210-9.

Klassen AC, **Curriero F**, Kulldorff M, **Alberg AJ**, Platz EA, Neloms ST (2006) Missing stage and grade in Maryland prostate cancer surveillance data. *Am J. Preventive Medicine* 30(2 Suppl.): S77-87.

Klassen AC, Kulldorff M, **Curriero F** (2005) Geographic clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *International Journal of Health Geographics* 4(1):1.

Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR, Hasan MZ, Pande VS, **Ruczinski I**, Doniach S, Plaxco KW (2004). Random-coil behavior and the dimensions of chemically unfolded proteins. *Proceedings of the National Academy of Sciences* 101(34): 12491-6.

Kooperberg C, **Ruczinski I** (2005). Identifying interacting SNPs using Monte carlo logic regression. *Genetic Epidemiology* 28(2): 157-70.

Maxwell KL, Wildes D, Zarrine-Afsar A, De Los Rios MA, Brown AG, Friel CT, Hedberg L, Horng JC, Bona D, Miller EJ, Vallee-Belisle A, Main ER, Bemporad F, Qiu L, Teilum K, Vu ND, Edwards AM, **Ruczinski I**, Poulsen FM, Kragelund BB, Michnick SW, Chiti F, Bai Y, Hagen SJ, Serrano L, Oliveberg M, Raleigh DP, Wittung-Stafshede P, Radford SE, Jackson SE, Sosnick TR, Marqusee S, Davidson AR, Plaxco KW (2005). Protein folding: defining a standard set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Science* 14(3): 602-16.

McCarney ER, Werner JH, Bernstein SL, **Ruczinski I**, Makarov DE, Goodwin PM, Plaxco KW (2005). Site-specific dimensions across a highly denatured protein; a single molecule study. *Journal of Molecular Biology* 352(3): 672-82.

Peng R, Dominici F, Pastor-Barriuso R, **Zeger SL, Samet JM** (2005) Seasonal analyses of air pollution and mortality in 100 U.S. cities, *American Journal of Epidemiology*, 161:585-594.

Peng R, Dominici F Louis T. (2006) Model choice in multi-site time series studies of air pollution and mortality, *Journal of the Royal Statistical Society, Series A* 169, Part 2:179- 203.

Peng R, Dominici F, Zeger SL (2006) Reproducible epidemiological research, *American Journal of Epidemiology*, 163: 783-789.

Peyser BD, **Irizarry RA**, Tiffany CW, Chen O, Yuan DS, Boeke JD, Spencer FA (2005) Improved statistical analysis of budding yeast TAG microarrays revealed by defined spike-in pools. *Nucleic Acids Research* 33(16):e140.

Quackenbush J, **Irizarry RA** (2006) Response to Shields: 'MIAME, we have a problem. *Trends in Genetics* 22(9): 471-472.

Quackenbush J, Stoeckert C, Ball C, Brazma A, Gentleman R, Huber W, **Irizarry R**, Salit M, Sherlock G, Spellman P, Winegarden N (2006) Top-down standards will not serve systems biology. *Nature* 440(7080):24.

Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, **Irizarry RA**, Liu J, Maier DS, Miller M, Petersen K, Quackenbush J, Sherlock G, Stoeckert CJ Jr, White J, Whetzel PL, Wymore F, Parkinson H, Sarkans U, Ball CA, Brazma A. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB (2006) *BMC Bioinformatics*, 7:489

Ruczinski I, Kooperberg C, LeBlanc M (2003). Logic regression. *Journal of Computational and Graphical Statistics* 12(3): 475-511.

Ruczinski I, Kooperberg C, LeBlanc M (2003). Logic regression - methods and software. *Non-linear Estimation and Classification, Lecture Notes in Statistics* 171: 333-44.

Ruczinski I, Kooperberg C, LeBlanc M (2004). Exploring interactions in high dimensional genomic data: an overview of logic regression, with applications. *Journal of Multivariate Analysis* 90: 178-95.

Ruczinski I, Sosnick TR, Plaxco KW (2006). Methods for the accurate estimation of confidence intervals on protein

Principal Investigator/Program Director (Last, First, Middle): Groopman, John D. – Bioinformatics/Biostatistics Core
folding ϕ -values. *Protein Science* 15(10): 2257-64.

Samet JM, Dominici F, McDermott A, Zeger SL (2003) New problems for an old design: time-series analyses of air pollution and health, *Epidemiology*, 14(1):11-12.

Sapkota A, Halden R, **Dominici F, Groopman J. Buckley T** (2006) Evaluation of 1,3-Butadiene biomarkers in environmental setting using liquid chromatography isotope dilution tandem mass spectrometry, *Chemico-Biological Interactions* 160:70-79.

Scharpf RB, Ting JC, Pevsner J, **Ruczinski I** (2006). SNPchip: R classes and methods for SNP array data. *Bioinformatics* (to appear).

Shone SM, **Curriero FC**, Lesser CR, Glass GE: Characterizing population dynamics of *Aedes sollicitans* (Diptera: Culicidae) using meteorological data. *Journal of Medical Entomology* 43(2):393-402, 2006.

Sidransky D, **Irizarry RA**, Califano JA, Li X, Ren H, Benoit N, and Mao L (2003) Serum protein MALDI profiling distinguishes upper aerodigestive tract cancer patients from controls. *Journal of the National Cancer Institute* 95:1711-1777.

Symons JM, Wang L, Guallar E, Howell E, **Dominici F**, Schwab M, Ange BA, **Samet JM, Ondov J**, Harrison D, **Geyh A** (2006) A Case-Crossover Study of Fine Particulate Matter Air Pollution and Congestive Heart Failure Hospitalization, *American Journal of Epidemiology*, 164: 421-433.

Tankersley C, Irizarry RA, Flanders S, Rabold R, and Frank R (2003) Unstable heart rate and temperature regulation predict mortality in ARK/J mice. *American Journal of Physiology-Regulatory, Integrative, and Comparative Physiology* 284:R742-750.

Ting JC, Ye Y, Thomas GH, **Ruczinski I**, Pevsner J (2006). Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. *BMC Bioinformatics* 18:7(1): 25.

Welty LJ, **Zeger SL** (2005) Are the acute effects of particulate matter on mortality in the National Morbidity, Mortality, and Air Pollution Study the result of inadequate control for weather and season? A sensitivity analysis using flexible distributed lag models. *American Journal of Epidemiology* 162(1):80-8.

Wheelan S, Scheifele L, Martinez-Murillo F, **Irizarry RA**, Boeke J (2006) The Transposon Insertion site Profiling Chip (TIP-chip). *Proceeding of the National Academy of Science* 103(47):17632-7.

Wheelan S, **Irizarry RA**, Martinez-Murillo F, Boeke J (2006) Stacking the deck: double-tiled DNA microarrays. *Nature Methods* 3(11):903-7.

Wu Z, **Irizarry RA**, Gentleman R, Martinez Murillo F, and Spencer F (2004) A model based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* 99: 909-917.

Wu Z, **Irizarry RA**: Preprocessing of oligonucleotide array data. (2004) *Nature Biotechnology* 22(6):656-8.

Wu Z and **Irizarry RA** (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *Journal of Computational Biology*. 12(6): 882-93.

Yuan DS, Pan X, Ooi SL, Peyser BD, Spencer FA, **Irizarry RA**, Boeke JD (2005) Improved microarray methods for profiling the yeast knockout strain collection. *Nucleic Acids Research* 33(12):e103.

Yuan DS, **Irizarry RA** (2006) High-Resolution Spatial Normalization for Microarrays. *Bioinformatics* 22(24):3054-60.

Zeger SL, Irizarry RA, Peng RD (2006) On time series analysis of public health and biomedical data. *Annual Review of Public Health* 27:57-79.

Submitted:

Crainiceanu C.M. **Dominici F**. Parmigiani G. Adjustment uncertainty in effect estimation, *Submitted to Biometrika*.

Dominici F. Evaluation framework for environmental public health tracking research: Criteria development and case studies. *Submitted to American Journal of Public Health*.

Peng R. **Dominici F**. Welty L. A Bayesian hierarchical distributed lag model for constrained distributed lag functions: estimating the time course of hospitalization associated with air pollution exposure. Submitted to *Journal of American Statistical Association*.

Zeger S.L. Dominici F. McDermott A. **Samet J.M.** Mortality in the Medicare Population and Chronic Exposure to Fine Particulate Matter. *Submitted to American Journal of Epidemiology*.

Books, Chapters, Monographs, Peer-reviewed Reports

Bell M, **Dominici F**, McDermott A, **Zeger SL**, Samet JM (2005) Multi-Site Time-Series Analysis of Ozone and Mortality in 95 U.S. Cities from 1987 to 2000: Results from the National Morbidity Mortality Air Pollution Study. Final Report to the Environmental Protection Agency.

Daniels M, **Dominici F**, Samet JM, **Zeger SL** (2004) The National Morbidity, Mortality, and Air Pollution Study Part III: Estimating PM10-Mortality Dose-Response Curves and Threshold Levels: An Analysis of Daily Time-Series for the 20 Largest US Cities, The Health Effects Institute, Cambridge, MA.

Dominici F (2004) Time Series Analysis of Air Pollution and Mortality: A Statistical Review. Final Report on the Walter A. Rosenblith New Investigator Award, The Health Effects Institute, Cambridge, MA.

Dominici F, Zanobetti A, **Zeger SL**, Schwartz J, **Samet JM** (2005) The National Morbidity, Mortality, and Air Pollution Study Part IV: Hierarchical Bivariate Time Series Models: A Combined Analysis of the Effects of Particulate Matter on Morbidity and Mortality, The Health Effects Institute, Cambridge, MA.

Dominici F, McDermott A, Daniels M, **Samet JM, Zeger SL** (2003) A Special Report to the Health Effects Institute on the Revised Analyses of the NMMAPS II Data, The Health Effects Institute, Cambridge, MA.

Dominici F Peng D. Bell M. Pham M. McDermott A. **Zeger S.L. Samet J.M.** (2006) Reply to Letters by Drs Ricci and Thurston on the article: Particles, Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases, *Journal of the American Medical Association* (to appear)

Health Effects from Exposure to PAVE PAWS Low-Level Phased-Array Radiofrequency Energy, (2005) The National Academies of Science, Washington DC.

Parmigiani G, Garrett ES, **Irizarry RA, Zeger SL**. *The Analysis of Gene Expression Data*: New York, NY: New York: Springer-Verlag, 2003.

Samet JM, Dominici F, Zeger SL, Dockery D, Schwartz J (2000) The National Morbidity, Mortality, and Air Pollution Study Part I: Methods and Methodological Issues, Number 94, The Health Effects Institute, Cambridge, MA.

Principal Investigator/Program Director (Last, First, Middle): Groopman, John D. – Bioinformatics/Biostatistics Core

Samet JM, Zeger SL, Dominici F, Curriero F, Coursac I, Dockery D, Schwartz J, Zanobetti A (2000) The National Morbidity, Mortality, and Air Pollution Study Part II: Morbidity and Mortality from Air Pollution in the United States, Number 94, The Health Effects Institute, Cambridge, MA.

Zeger SL, McDermott A, Dominici F, Peng R, Samet JM,(2005) Internet-based Health and Air Pollution Surveillance System (iHAPSS), The Health Effects Institute, Cambridge, MA

Zeger SL, Dominici F, McDermott A, Samet JM (2004) Modeling the Health Effects of Environmental Exposures: Particulate Air Pollution and Mortality in Urban America, in *Monitoring the Health of Populations: Statistical Methods for Public Health Surveillance* (Edited by Brookmeyer R. and Stroup D.), Oxford University Press

Additional Papers Cited

Hsu, Y., Strait, S., and Holoman, D. (2007). Speciate 4.0: Speciation Data Documentation - Final Report. U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-06/161.

National Research Council (2004). Research Priorities for Airborne Particulate Matter: IV. Continuing Research Progress. National Research Council of the National Academies.

Nikolov, M., Coull, B., Catalano, P., and Goldeski, J. (2007). "An informative Bayesian structural equation model to assess source-specific health effects of air pollution." *Biostatistics*, to appear.

Health Effects Institute (2002). Understanding the Health Effects of Components of the Particulate Matter Mix: Progress and Next Steps. Perspectives 2, Health Effects Institute, Boston MA.

Hopke, P. K., Ito, K., Mar, T., Christensen, W. F., Eatough, D. J., Henry, R. C., Kim, E., Laden, F., Lall, R., Larson, T. V., Liu, H., Neas, L., Pinto, J., Stolzel, M., Suh, H., Paatero, P., and Thurston, G. D. (2006). PM source apportionment and health effects: 1. Intercomparison of source apportionment results. *Journal of Exposure Science and Environmental Epidemiology*, 16, 3, 275-286.

Kelsall JE, **Samet J, Zeger SL**, Xu J: Air pollution and mortality in Philadelphia, 1974-1988. *American Journal of Epidemiology* 146(9):750-762, 1997.

Liang K-Y, **Zeger SL**: Longitudinal data analysis using generalized linear models. *Biometrika* 73(1):13-22, 1986.

Thurston, G. D., Ito, K., Mar, T., Christensen, W. F., Eatough, D. J., Henry, R. C., Kim, E., Laden, F., Lall, R., Larson, T. V., Liu, H., Neas, L., Pinto, J., Stolzel, M., Suh, H., and Hopke, P. K. (2005a). Workgroup report: workshop on source apportionment of particulate matter health effects-intercomparison of results and implications. *Environmental Health Perspectives*, 113, 12, 1768-1774.

Zeger SL: Regression model for time series of counts. *Biometrika* 75:621-630, 1988.