BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Sabatti, Chiara

eRA COMMONS USER NAME (credential, e.g., agency login): CSABATTI

POSITION TITLE: Professor of Biomedical Data Science and of Statistics

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Bocconi University, Milan, Italy	B.S., M.S.	07/93	Economics & Statistics
Stanford University, Stanford, CA	Ph.D.	08/98	Statistics
Stanford Medical School, Stanford, CA	Post Doc	06/00	Genetics

A. Personal Statement

My research group works to develop statistical methodology to analyze genomic data (most commonly genetic variation and gene expression measurements), paying particular attention to the challenges presented by its high dimensionality. Section C. below details some of the specific scientific questions I have tackled, many of which represent areas of interest for BMI students. In recent years I have also been focusing on questions of data ethics and fairness, [1] and causality [2].

As a member of the BMI exec committee, and the Associate Chair for Education and Training in the Biomedical Data Science department, I contribute to the decisions regarding curriculum, admission, and administration of the BMI program. I am currently the academic advisor of three master students, and the member of the thesis committee of one student. I teach Stats 200, which is a required Statistical Methodology class for both the BMI master and PhD students.

During the past year, I have worked closely with the BMI students to tackle questions of social justice, diversity, equity and fairness in the academic and data science world. I participated in the student book club and discussions, and I am a member of the department committee on justice, equity, diversity and inclusion. I have introduced and taught the class BIODS/BIOMEDIN 240 "Race, data, algorithms and health", which is recognized by the BMI program as satisfying an ethics requirement. I have championed a new initiative aiming to connect the Stanford data science community with undergraduates in institutions that do not focus on research in this area, with the goals of increasing recruitment and retention of URM. This resulted in the pilot class BIODS 360 "Inclusive mentorship in data science", which pairs up 20 Stanford PhD students (the vast majority from the BMI program) with 20 undergraduates in non R1 institutions.

1. Romano, Y., R. Barber, C. Sabatti, E, Candès (2020) "With Malice Towards None: Assessing Uncertainty via Equalized Coverage," Harvard Data Science Review 2.2
2. Bates, S., M. Sesia, C. Sabatti, and E. Candès (2020) "Causal inference in genetic trio studies" PNAS, 117 24117–24126 PMCID: PMC7533659

B. Positions and Honors

Positions and Employment

1993-1994 Research Fellow, Department of Statistics, Bocconi University
2000-2006 Assistant Professor, Human Genetics and Statistics, UCLA
2006-2009 Associate Professor, Human Genetics and Statistics, UCLA
2004-2009 Associate of the UCLA Center for Society and Genetics
2009-2011 Professor, Human Genetics and Statistics, UCLA (on leave)

2009-2015 Associate Professor, HRP (Biostatics), Stanford; member of BioX.
2015-2016 Associate Professor, Biomedical Data Science and Statistics, member of BioX.
2016- Professor, Biomedical Data Science and Statistics, Stanford University
2019- Associate director for the undergraduate major in Mathematical and Computational Science
2020- Associate director for Education, Stanford Data Science Initiative
2020- Associate Chair for Education and Training, Biomedical Data Science

Honors

1990, 1992 Two times winner of the Credito Bergamasco Award
1993 Amici della Bocconi dissertation award
1998 Best Teaching Assistant Award, Statistics Department, Stanford
2003 NSF Career Award

C. Contributions to Science

Bayesian modeling for high throughput biological data

Progress in biotechnology has opened (and continues to open) new horizons for biology and medicine. The "-omics" datasets tend to be very high dimensional and, while they offer the opportunity of a global view, they come with a specific set of challenges. Often, each single measurement is fairly noisy and is best interpreted in the context of other aspects of the experiment. The Bayesian inferential approach offers a natural framework to incorporate multiple sources of information, but comes with computational challenges. Dr. Sabatti has both developed novel computational strategies that adapt well to the investigation of high-dimensional parameter spaces [3] and introduced new models designed to capture the specificity of genetic investigations [4-6]. The star-genealogy model introduced in [4] has become popular in a wide class of haplotype reconstruction algorithms. The careful modeling of signal intensities for genotyping arrays introduced in [5] turned out to be crucial for the detection of genomic variation other than single nucleotide polymorphisms. In [6] we explore a Bayesian model selection for the goal of gene mapping.

3. Liu, J. and C. Sabatti (2000) "Generalized Gibbs sampler and multigrid Monte Carlo for Bayesian computation," Biometrika, 87: 353-369.
4. Liu, J., C. Sabatti, J. Teng, B. Keats, and N. Risch (2001) "Bayesian analysis of haplotypes for linkage disequilibrium mapping," Genome Research, 11: 1716-24. PMCID 311130.
5. Sabatti, C. and K. Lange (2008) "Bayesian Gaussian mixture models for high density genotyping arrays," Journal of the American Statistical Association 103: 89–100. PMCID: PMC3092390.
6. Stell, L. and C. Sabatti (2016) "Genetic variant selection: learning across traits and sites," Genetics 202: 439–55. PMCID: PMC4788227.

Gene expression and its genetic regulation.

Our group has also studied how genetic variations regulate gene expression. We have been involved in the analysis of datasets collected to understand the biological pathways underlying Bipolar Disorder [7], to study variation across tissues in model organisms [9] and humans [10]. We have also developed new statistical methodology that guarantees control of FDR while increasing the power of detecting regulation shared across tissues, of which [8] is one example.

7. Peterson, C., S. Service, A. Jasinska, F. Gao, I. Zelaya, T. Teshiba, C. Bearden, V. Resus, G. Macaya, C. Lopez, M. Bogomolov, Y. Benjamini, E. Eskin, G. Coppola, N. Freimer, and C. Sabatti (2016) "Genetic regulation of LCL gene expression in families segregating bipolar disorder," PLOS Genetics, 12:e1006046, PMCID: PMC4866754
8. Peterson, C., M. Bogomolov, Y. Benjamini, and C. Sabatti, "TreeQTL: hierarchical error control for eQTL findings," Bioinformatics, 2016. PMID: 27153635.
9. Jasinska, A., I. Zelaya, S. Service, C. Peterson, R. Cantor, O. Choi, J. DeYoung, E. Eskin, L. Fair- banks, S. Fears, A. Furterer, Y. Huang, V. Ramensky, C. Schmitt, H. Svardal, M. Jorgensen, J. Kaplan, D. Villar, B. Aken, P. Flicek, R. Nag, E. Wong, J. Blangero, T. Dyer, M. Bogomolov, Y. Benjamini, G. Weinstock, K. Dewar, C. Sabatti, R. Wilson, J. Jentsch, W. Warren, G. Coppola, R. Woods, N. Freimer (2017) "Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate," Nature Genetics, 49: 1714-1721. PMID: 29083405.

10. The GTEx Consortium (2017) "Genetic effects on gene expression regulation across human tissues," Nature 550: 204-213. PMID: 29022597.

Reconstruction of regulatory networks

Gene expression arrays and mRNA sequencing produce measurements of the expression levels of thousands of genes simultaneously, offering the opportunity to study how genes interact with each other and how they respond to the environment. Transcription factors play a crucial role in defining the cellular response to external conditions: by switching between active and inactive states they modulate the changes in expression of the genes they control, even when the transcription factors themselves do not undergo changes in expression. By studying DNA sequences to identify transcription factor binding sites [12], and pioneering a de-convolution approach to the analysis of gene expression data [11], Dr. Sabatti has outlined a program to reconstruct regulatory networks that capitalizes on all available information and is flexible enough to infer missing links. In particular, by using a Bayesian approach [13] or a penalized likelihood method [14], it is possible to identify groups of genes regulated by the same transcription factor and infer the changes in concentration of activity levels of the TF. The use of factor-like models for the analysis of gene expression data has become mainstream and the papers below contributed to the success of this approach.

11. Liao, J., R. Boscolo, Y. Yang, L. Tran, C. Sabatti, and V. Roychowdhury (2003) "Network component analysis: reconstruction of regulatory signals in biological systems," Proceedings of the National Academy of Science, 100: 15522-15527. PMCID: 307600.
12. Sabatti, C., L. Rohlin, K. Lange, and J. Liao (2005) "Vocabulon: a dictionary model approach for reconstruction and localization of transcription factor binding sites," Bioinformatics, 21: 922-931. PMID: 15509602.
13. Sabatti, C. and G. James (2006) "Bayesian sparse hidden components analysis for transcription regulation networks," Bioinformatics, 22: 739-746. PMID: 16368767.
14. James, G., C. Sabatti, N. Zhou, and J. Zhu (2010) "Sparse Regulatory Networks," The Annals of Applied Statistics, 4(2): 663-686. PMCID: 3102251.

Statistical methods for genome wide association studies

The availability of high-density genotyping has made it possible to search for the signature of functional alleles in case-control or population samples, making genome-wide association studies (GWAS) the "bread and butter" of gene mapping in the last decade. The statistical analysis of these data appeared deceptively simple: the complicated likelihood maximization required for linkage studies are substituted by a series of t-tests or linear regressions. In reality, a new set of challenges are associated with GWAS and the papers below address some of them. One of the attractive aspects of a population design is that the same set of subjects can be used to study multiple traits, as their measurements are often available in cohorts [15], but what is the appropriate threshold for significance in this context? [17] argues in favor of opting for a False Discovery Rate (FDR) controlling criteria, carefully exploring advantages and challenges. Another feature that [15] points out is that--even in genetically uniform populations as Finland--genomic variation mirrors the geographical origin of individuals. The presence of such population structure has important consequences for the analysis of GWAS: there are effectively various degrees of distant relations between individuals in the samples and it is necessary to account for these. In [16], Dr. Sabatti and co-authors show how the variance-component approach traditionally used in linkage analysis could be effectively adapted to this challenge: this paper had an instrumental role in introducing mixed models to the GWAS literature. Finally, [18] documents some exciting developments of most recent years, when we have been able to leverage sophisticated machine learning models to analyze the totality of genetic markers to detect those reproducibly and distinctly associated with changes in disease risks.

15. Sabatti. C., S. Service, A. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. Jones, N. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, C. Lachlan, C. Hoggart, P. Elliott, A. Collins, H. Turunen, S. Gabriel, M. McCarthy, M. Daly, M-R. Jarvelin, N. Freimer, L. Peltonen (2009) "Genomewide association analysis of metabolic phenotypes in a birth cohort from a founder population," Nature Genetics, 41: 35-46. PMCID: 2687077.
16. Kang, H., J-H. Sul, S. Service, N. Zaitlen, S.Kong, N. Freimer, C. Sabatti*, E. Eskin* (2010) "Accounting for sample structure in large scale genome-wide association studies using a variance component model," Nature Genetics, 42: 348-54. PMCID: 3092069.
17. Brzyski, D., C. Peterson, P. Sobczyk, E. Candès, M. Bogdan and C. Sabatti (2017) "Controlling the rate of GWAS false discoveries," Genetics, 205: 61-75. PMID: 27784720.

18. Sesia, M., E. Katsevich, S. Bates, E. Candès and C. Sabatti (2020) "Multi-resolution localization of causal variants across the genome" Nature Communications 11: 1–10. PMID:32107378.

Understanding biological mechanisms to improve disease treatment

Gene expression and genetic variation are but some of the biological features we can work with to understand disease mechanisms and identify biomarkers. In collaborations, I have analyzed data of many other forms, as partially illustrated in the following publications.

19. Riley, R., C. Lee, C. Sabatti, and D. Eisenberg (2005) "Measuring evidence for protein domain interactions from multi-species protein interaction data," Genome Biology, 6: R89. PMCID 1257472.
20. Hattori, D., Y. Chen, B. Matthews, L. Salwinski, D. Eisenberg, C. Sabatti, W. Grueber, L. Zipuski (2009) "Robust discrimination between self and non-self neurites requires thousands of Dscam1 isoforms," Nature, 461: 644-648. PMCID: PMC2836808.
21. Visnyei, K., H. Onodera, R. Damoiseaux, K. Saigusa, S. Petrosyan, D. De Vries, D. Ferrari, J. Saxe, E. Panosyan, M. Masterman-Smith, J. Mottahedeh, K. Bradley, J. Huang, C. Sabatti, I. Nakano, H. Kornblum (2011) "A molecular screening approach to identify and characterize inhibitors of glioblastoma stem cells." Mol Cancer Ther. 10:1818-28. PMID: 21859839
22. Salahudeen, A., S. Choi, A. Rustagi, J. Zhu, V. van Unen, S. de la O, R. Flynn, M. Margalef-Català, A. Santos, J. Ju, A. Batish, T. Usui, G. Zheng, C. Edwards, L. Wagar, V. Luca, B. Anchang, M. Nagendran, K. Nguyen, D. Hart, J. Terry, P. Belgrader, S. Ziraldo, T. Mikkelsen, P. Harbury, J. Glenn, K. Garcia, M. Davis, R. Baric, C. Sabatti, M. Amieva, C. Blish, T. Desai, C. Kuo (2020) "Progenitor identification and SARS-CoV-2 infection in human distal lung organoids." Nature, 588: 670–675. PMID:33238290.

Complete List of Published Works in PubMed:
http://www.ncbi.nlm.nih.gov/pubmed?term=Sabatti%20C%5BAuthor%5D

D. Additional Information: Research Support and/or Scholastic Performance

Ongoing research support

R56HG010812 (Sabatti) 1/1/21-12/31/2022
National Institutes of Health/NHGRI
Title: The pursuit of genetic causal mechanisms
Goal: To zoom in on genetic variants with causal effects, we capitalize on powerful machine learning algorithms and, crucially, equip their results with precise replicability guarantees. It will also improve the precision of personalized risk evaluations based on genotypes: if we can construct risk scores using variants that are truly causal, their performance will remain solid across ethnicities and environmental exposures.
Role: PI

NSF 1934578 (Candes) 10/1/19-9/30/2021
National Science Foundation
The Stanford Data Science Collaboratory
Goal: to develop the collaborative infrastructure necessary to enable data science research at Stanford.
Role: co-PI

Math+X (Candes) 4/1/18-6/30/2021
Simons Foundation
Title: Math+X: Encouraging Interactions Program
Goal: Application of model selection approaches to the discovery of relevant genetic variants
Role: co-investigator.

UL1TR003142-01 (O'Hara) 7/15/19-6/30/2024
National Institutes of Health
Stanford Center for Clinical and Translational Education and Research
Goal: accelerate the translation of basic scientific discoveries into practical solutions that improve human health, through educational programs, research support, infrastructure streamlining and innovation funding.
Role: Statistician

R01MH113078 (Freimer, Stanford PI: Sabatti) 4/1/17-1/31/2022
UCLA/NIH primary
Title: Genetics of Severe Mental Illness
Goal: This project aims to use genetics to help develop an approach for classifying severe mental illness (SMI) that has a stronger scientific foundation than the systems currently used in both research and clinical practice. We are using genetic data as well as extensive phenotyping of individuals that receive treatment in one hospital in Colombia.
Role: subcontract PI

2R01GM064798-09 (Herschlag) 9/11/19-8/31/2023
National Institutes of Health
Quantitative, High-throughput Mechanistic Enzymology
Goal: To develop HT-MEK, HT-MES, and QDE experimental high through-put platforms to improve our understanding of how enzyme sequence and structure dictate functional properties.
Role: co-investigator

R01 HL148704-01A1 (Mignot) 4/1/20-3/31/2024
National Institutes of Health
Proteomic and Transcriptomic Biomarkers of Circadian Timing
Goal: to develop robust diagnostic biomarkers for circadian timing that can identify, from a single biospecimen, the biological time within an individual. To do this, we will use two state of the art methods, a plasma proteomics-based method to identify a panel of rhythmic proteins and a whole blood-derived monocyte-based method using a panel of 15 transcripts.
Role: co-investigator

R01MH123157 (Freimer, Stanford PI: Sabatti) 5/1/20-2/28/2025
UCLA/ NIH primary
A Latin American biobank for large-scale genetics research on severe mental illness
Goal: To create a resource to study the genetic basis of mental illness using electronic medical records, genome-wide genotyping, and a large, ethnically matched control sample
Role: Sub-contract PI

R01DK11572801 (Kuo) 8/15/18-6/30/2023
NIH
Title: Structure-based Bioengineering of Wnt Surrogates for Intestinal Stem Cell Biology and Therapy
Goal: To image the entire Wnt/Frizzled/Lrp6 ternary transmembrane complex by X-ray crystallography and cryo-Electron Microscopy, building on preliminary successes in expressing and purifying this multimolecular assembly, and to overcome two major obstacles for translation of Wnts into therapeutics: 1) difficulty of expressing natural Wnts as recombinant proteins due to their lipidation, and 2) Fz cross-reactivity.
Role: co-investigator

Completed research support

DMS 1712800 (Sabatti) 9/01/17-8/31/2020
NSF
Title: Discovering what matters: informative and reproducible variable selection with applications to genomics
Goal: Develop new statistical methods to identify variables that influence outcome of interest. Partly in response to a change in budget, this project focuses on theoretical and methodological aspects and does not include a component of direct analysis of genetic datasets.
Role: PI

Discovery Innovation Fund (Sabatti) 10/01/17-9/31/2020
Stanford
Title: Reproducible identification of cancer cell types via scRNAsec
Goal: Develop statistical methods to identify clusters of cells that are significantly distinct indicating that they correspond to sub-populations representing different biological roles.
Role: PI