BIOGRAPHICAL SKETCH

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

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. The following papers represent some of the methodological work that these applications have motivated.

1. Bogdan, M., E. van den Berg, **C. Sabatti**, W. Su, and E. Candès (2015) "SLOPE – adaptive variable selection via convex optimization," *Annals of Applied Statistics*, **9**: 1103-1140. PMID: 26709357.
2. 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.
3. Katsevich, E. and **C. Sabatti** (2019) "Multilayer Knockoff Filter: Controlled variable selection at multiple resolutions," *The Annals of Applied Statistics*, available at https://arxiv.org/abs/1706.09375.
4. Sesia, M., **C. Sabatti**, and E. Candès (2019), "Gene hunting with knockoffs for hidden markov models," *Biometrika* **106**: 1-18. PMID: 30799875.

B. Positions and Honors

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-Present Professor, Biomedical Data Science and Statistics, Stanford University
2018-Present Associate Chair, 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
2002 UCLA Career Award

C. Contributions to Science

Linkage disequilibrium and haplotype models

The process of recombination is such that adjacent portions of the DNA tend to be transmitted together, a phenomena that induces dependence between the alleles at neighboring polymorphic loci. In turn, the observation of statistical association between the alleles at different markers (linkage disequilibrium) can be used to gather information on the historical recombination events along the DNA segment separating these loci. This reconstruction is complicated by the fact that polymorphic sites differ in allele frequencies, reflecting diverse histories; limited sample sizes offer only a partial view of the process; and researchers typically have access only to genotype data and lack information on which alleles are on the same ancestral chromosome (phase). Dr. Sabatti introduced novel measures of linkage disequilibrium that address some of these challenges. She showed that by relying on homozygosity it is possible to gather information on linkage disequilibrium while bypassing the need for phasing genotype data [5]. Additionally, the volume measure she introduced accounts for the effects of both variable allele frequency and small sample size [6]. Finally, the phenomenological dictionary model for haplotypes [8] permits the incorporation of the effects of a complex genealogy without attempting to reconstruct it explicitly. High-density genotyping data have revealed the common presence of homozygous segments across the genome [7] and these have been considered as signatures of selection, identity by state, presence of copy number variants, etc. The observation that linkage disequilibrium alone can produce homozygous segments has been crucial in devising methods to detect these different phenomena that have adequate specificity.

5. **Sabatti, C**. and N. Risch (2002) "Homozygosity and linkage disequilibrium," *Genetics,* **160**: 1707-1719.
6. Chen, Y., C. Lin, **C. Sabatti** (2006) "Volume Measures for Linkage Disequilibrium," *BMC Genetics*, **7**: 54. PMCID: 1665459.
7. Wang, H., C. Lin, S. Service, The international collaborative group on isolated populations, Y. Chen, N. Freimer, **C. Sabatti** (2006) "Linkage disequilibrium and haplotype homozygosity in population samples genotyped at a high marker density," *Human Heredity*, **62**: 175-189. PMID: 17077642.
8. Ayers, K., **C. Sabatti** and K. Lange (2007) "A Dictionary Model for Haplotyping, Genotype Calling and Association Testing," *Genetic Epidemiology*, **31**: 672-683. PMID: 17487885.

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 [10], but what is the appropriate threshold for significance in this context? [9] argues in favor of False Discovery Rate (FDR) as a measure of global error for studies of complex traits, where we expect more than one gene implicated, and [10] shows how this criteria adapts to the study of multiple disease. The advantages and challenges of adopting FDR as a target error rate are further explored in [12]. [10] is also one of the first papers to underscore, on the one hand, how the `hits' from GWAS appear to explain a small percentage of the variance of phenotypes, and, on the other hand, that SNP variation overall seems to account for a larger portion of heritability than captured by the `hits'. In [10], Dr. Sabatti and co-authors used a statistical test known as higher criticism to illustrate this point, that has since been made abundantly clear using polygenic scores or heritability estimates. Another feature that [10] 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 [11], 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.

9. **Sabatti, C.**, S. Service, and N. Freimer (2003) "False discovery rate in linkage and association genome screens for complex disorders," *Genetics,* **164**: 829-833. PMCID: 1462572.
10. **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.
11. 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.
12. Peterson, C., M. Bogomolov, Y. Benjamini, **C. Sabatti** (2016) "Many phenotypes without many false discoveries: error controlling strategies for multi-trait association studies," *Genetic Epidemiology*, **40**: 45-56. PMCID: 2662603

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 [12], to study variation across tissues in model organisms [15] and humans [16]. We have also developed new statistical methodology that guarantees control of FDR while increasing the power of detecting regulation shared across tissues [14].

13. 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 Genet- ics*, **12**:e1006046
14. C. B. Peterson, M. Bogomolov, Y. Benjamini, and **C. Sabatti**, "TreeQTL: hierarchical error control for eQTL findings," *Bioinformatics*, 2016. PMID: 27153635.
15. 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.
16. 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 [18], and pioneering a de-convolution approach to the analysis of gene expression data [17], 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 [19] or a penalized likelihood method [20], 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.

17. 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.
18. 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.
19. **Sabatti, C.** and G. James (2006) "Bayesian sparse hidden components analysis for transcription regulation networks," *Bioinformatics*, **22**: 739-746. PMID: 16368767.
20. James, G., **C. Sabatti**, N. Zhou, and J. Zhu (2010) "Sparse Regulatory Networks," *The Annals of Applied Statistics*, 4(2): 663-686. PMCID: 3102251.

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 [21] and introduced new models designed to capture the specificity of genetic investigations [22,23,24]. The star-genealogy model introduced in [22] has become popular in a wide class of haplotype reconstruction algorithms. The careful modeling of signal intensities for genotyping arrays introduced in [23] turned out to be crucial for the detection of genomic variation other than single nucleotide polymorphisms. In [24] we explore a Bayesian model selection for the goal of gene mapping.

21. Liu, J. and **C. Sabatti** (2000) "Generalized Gibbs sampler and multigrid Monte Carlo for Bayesian computation," *Biometrika*, **87**: 353-369.
22. 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.
23. **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: 3092390.
24. Stell, L. and **C. Sabatti** (2016) "Genetic variant selection: learning across traits and sites," *Genetics* **202**: 439–55.

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

DMS 1712800 (Sabatti) 09/01/17-08/31/20
NSF
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) 10/01/17-09/31/19
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

Math+X (Candes) 4/1/2018-02/28/2021
Simons Foundation
Title: Math+X: Encouraging Interactions Program
Goal: Application of model selection approaches to the discovery of relevant genetic variants
Role: co-investigator.

R01MH113078 (Freimer, Stanford PI: Sabatti) 4/1/2017—1/31/2022
UCLA/NIH primary
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

R01DK11572801 (Kuo) 8/15/2018--6/30/2019
NIH
Goal: Structure-based Bioengineering of Wnt Surrogates for Intestinal Stem Cell Biology and Therapy
Role: co-investigator

Completed research support

R01 HG006695 (Sabatti) 4/1/13 – 3/31/16
NIH
New Statistical Methods for High Resolution Mapping of Multiple Phenotypes
Goal: The project proposes statistical method developments to analyze rare variants, map multivariate phenotypes, while controlling for False Discovery Rates in powerful manners. It includes a subcontract with the University of Tel Aviv and Prof. Y. Benjamini as a principal investigator.
Role: Principal Investigator

R01MH101782 (Sabatti) 8/1/13- 6/30/16
NIH/NIMH
Genetic Regulation of Gene Expression and its Impact on Phenotypes
Goal: This project will develop statistical methods for the study of the genetic regulation of gene expression, capitalizing on the datasets collected under the umbrella of the GTEX project. It includes a subcontract to UCLA, with Prof. E. Eskin principal investigator.
Role: Co-investigator

U01MH105578 (Freimer, Palotie; Sabatti Stanford PI) 9/23/2014 - 7/31/2018
UCLA/NIH primary
Genomic Strategies to Identify High-impact Psychiatric Risk Variants
Goal: Schizophrenia (SCZ) and bipolar disorder (BP) are the major adult psychotic disorders; uncertainty about their relationship is a central issue in psychiatry. By discovering variants with a high impact on either or both disorders (or on their endophenotypes) we can transform our understanding of their biology. This multisite project, from investigators with a track record of successful collaboration, will leverage exceptional, extensively phenotyped pedigree and population samples and integrate a combination of bioinformatics and experimental genomics approaches to identify such variants and demonstrate their relationship to these diseases.
Role: Co-investigator

U01HG007419 (Matise; Bustamante Stanford PI) 6/1/13 – 5/31/17
Rutgers University/NIH Primary
NHGRI PAGE Coordinating Center
Goal: The CC will serve as a centralized resource to facilitate and support the activities of the investigators in the Population Architecture Using Genomics and Epidemiology (PAGE) research program. The ultimate goal of our CC is to facilitate the identification and characterization of genotype-phenotype associations, especially as relevant to non-European populations, thereby accelerating our understanding of ancestral differences in the genetic and environmental causes of common diseases
Role: Co-investigator

R01HL113315 (Freimer, Palotie; Sabatti Stanford PI) 6/15/12 - 3/31/17
NIH/NHLBI
Genomic and Metabolomic Profiling of Finnish Familial Dyslipidemia Families
Subcontract from UCLA (Sabatti, Stanford PI)
Goal: This proposal is to re-investigate, using metabolomic profiling and advanced genomics technologies, 92 Finnish pedigrees that were ascertained for two forms of complex heritable dyslipidemia: familial combined hyperlipidemia (FCHL) and low serum levels of high density lipoprotein cholesterol (HDL-C).
Role: Dr. Sabatti is the chief statistician in the project.