Lecture Notes in Bioinformatics 2983

Edited by S. Istrail, P. Pevzner, and M. Waterman



Subseries of Lecture Notes in Computer Science

Springer
*Berlin*
*Heidelberg*
*New York*
*Hong Kong*
*London*
*Milan*
*Paris*
*Tokyo*

Sorin Istrail  Michael Waterman
Andrew Clark (Eds.)

# Computational Methods for SNPs and Haplotype Inference

DIMACS/RECOMB Satellite Workshop
Piscataway, NJ, USA, November 21-22, 2002
Revised Papers

Springer

Volume Editors

Sorin Istrail
Celera Genomics, Applied Biosystems
45 West Gude Drive, Rockville, MD 20850, USA
E-mail: sorin.istrail@celera.com

Michael Waterman
Celera Genomics and
University of Southern California
1042W 36th Place, DRB 155, Los Angeles, CA 90089-1113, USA
E-mail: msw@hto.usc.edu

Andrew Clark
Celera Genomics and
Cornell University, Molecular Biology and Genetics
Ithaca, NY 14853, USA
E-mail: Andy.Clark@celera.com

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

CR Subject Classification (1998): H.2.8, F.2.2, G.2.1, G.3, H.3.3, I.5, J.3

ISSN 0302-9743
ISBN 3-540-21249-3 Springer-Verlag Berlin Heidelberg New York



Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10993057 06/3142 5 4 3 2 1 0

# Preface

The 1st Computational Methods for SNPs and Haplotype Inference Workshop was held on November 21–22, 2002 at the DIMACS Center for Discrete Mathematics and Theoretical Computer Science.

The workshop focused on methods for SNP and haplotype analysis and their applications to disease associations. The ability to score large numbers of DNA variants (SNPs) in large samples of humans is rapidly accelerating, as is the demand to apply these data to tests of association with diseased states. The problem suffers from excessive dimensionality, so any means of reducing the number of dimensions of the space of genotype classes in a biologically meaningful way would likely be of benefit. Linked SNPs are often statistically associated with one another (in "linkage disequilibrium"), and the number of distinct configurations of multiple tightly linked SNPs in a sample is often far lower than one would expect from independent sampling. These joint configurations, or haplotypes, might be a more biologically meaningful unit, since they represent sets of SNPs that co-occur in a population. Recently there has been much excitement over the idea that such haplotypes occur as blocks across the genome, as these blocks suggest that fewer distinct SNPs need to be scored to capture the information about genotype identity. There is need for formal analysis of this dimension reduction problem, for formal treatment of the hierarchical structure of haplotypes, and for consideration of the utility of these approaches toward meeting the end goal of finding genetic variants associated with complex diseases.

The workshop featured the following invited speakers:
Peter Donnelly (Oxford University), Kathryn Roeder (Carnegie Mellon University), Jonathan Pritchard (University of Chicago), Molly Przeworski (Max Planck Institute), Maoxia Zheng (University of Chicago), Elizabeth Thompson (University of Washington), Monty Slatkin (University of California, Berkeley), Dahlia Nielsen (North Carolina State University), Matthew Stephens (University of Washington), Andrew Clark (Cornell University and Celera/Applied Biosystems), Sorin Istrail (Celera/Applied Biosystems), David Cutler (Johns Hopkins University), Magnus Nordborg (University of Southern California), Bruce Rannala (University of Alberta), Russell Schwartz (Carnegie Mellon University), Fengzhu Sun (University of Southern California), Jun Liu (Stanford University), Jinghui Zhang (National Cancer Institute, NIH), Dan Gusfield (University of California, Davis), Eran Halperin (University of California, Berkeley), Vineet Bafna (The Center for the Advancement of Genomics), David Altschuler (Harvard Medical School), Nancy Cox (University of Chicago), Francisco de la Vega (Applied Biosystems), Li Jin (University of Cincinnati) and Steve Sherry (National Center for Biotechnology Information, NIH).

This volume includes papers on which the presentations of Andrew Clark, Dan Gusfield, Sorin Istrail, Tianhua Niu, Jonathan Pritchard, Russell Schwartz, Elizabeth Thompson, Bruce Rannala, Fengzhu Sun, and Maoxia Zheng were based. It also includes the collection of abstracts of all the presentations.

We would like to thank Merissa Henry for outstanding workshop organization and editorial support. We would also like to thank the DIMACS Center for Discrete Mathematics and Theoretical Computer Science for providing financial support and excellent organization of the workshop.

January 2004

Andrew G. Clark
Sorin Istrail
Michael Waterman

Workshop Organizers

# Table of Contents



## Abstracts