Bioinformatic Analysis of Imprinted CpG Islands in *Mus Musculus*

Daniel Patrick Riordan
Churchill College
University of Cambridge
August 2003

This dissertation is submitted for the degree of Master of Philosophy.

PREFACE

This dissertation is my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

ACKNOWLEDGEMENTS

I wish to acknowledge my advisor, Richard Durbin, for his valuable guidance and direction in helping me to complete this work. The other members of the Durbin Lab – Ian Korf, Lachland Coin, and Ashwin Hajarnavis – also helped me in various ways throughout the year, for which I am thankful. I would also like to recognize Rachel Smith and Gavin Kelsey for providing the initial dataset upon which this analysis was based and for their helpful suggestions. I am grateful to have been supported by a Churchill Scholarship from the Winston Churchill Founation of the United States of America and also by the Wellcome Trust in carrying out this research. I would especially like to recognize all of my friends and family who have always been there for me. Most of all, I want to thank my parents, Kathleen and John Riordan, without whose constant love and support I wouldn't be the person I am today.

TABLE OF CONTENTS

ABSTRACT

Bioinformatic Analysis of Imprinted CpG Islands in *Mus musculus*

For some genes the maternally and paternally inherited alleles are differentially labeled with epigenetic markings leading to mono-allelic expression from only one copy, a process known as genomic imprinting. Previous studies have shown that imprinting is often associated with differential methylation of CpG islands (CGIs) and that SINE repeats occur less frequently in imprinted regions. This study identifies additional features of differentially-methylated CpG islands (DMR-CGIs) and uses them to help predict novel imprinted loci. A database containing sequences for 60 imprinted genes at 41 loci and 13,070 control genes in mouse was created and analyzed according to repeat content and CGI sequence properties. The SINE repeat content was significantly reduced at imprinted versus control loci, confirming previous reports. The sequence composition of CGIs associated with imprinted and control genes was also examined, and a considerable number of oligonucleotides with significantly different frequencies between DMR and control CGIs were found. A scoring function based on these oligonucleotides was developed that assigns greater scores to DMR-CGIs than controls at a highly significant level ($p < 10^{-5}$), and this scoring function was used in conjunction with regional SINE repeat content to predict novel imprinted loci. Genes associated with the predicted novel imprinted loci were compared with another set of candidate imprinted genes that were recently experimentally identified by large-scale expression profiling of FANTOM2 mouse cDNAs, and there was considerable overlap between both sets.