



Inferring the Genome Mosaic in the Collaborative Cross

Department of Computer Science

University of North Carolina at Chapel Hill

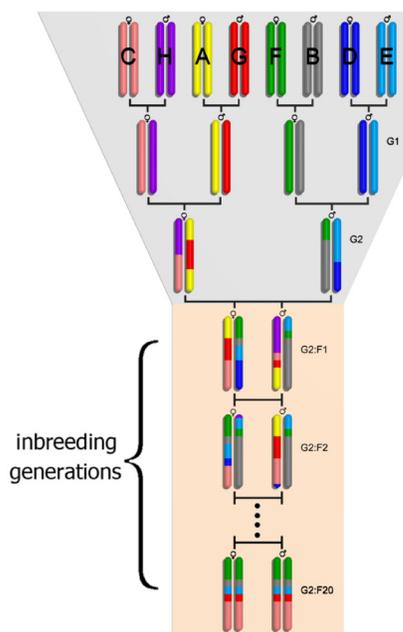
March 2012

Introduction

The Collaborative Cross (CC) is a large panel of new inbred mouse strains derived from a common set of eight inbred founder strains. We have developed novel methods to determine the haplotype origins in developing CC lines at any stage of inbreeding and used them to monitor the developing CC lines. Our methods consider how the structured pedigree (an 8-way breeding funnel) influences the genetic structure of the resulting line. Ignoring such breeding constraints leads to spurious inferences.

CC: Breeding Design and Constraints

The CC strains are derived using a funnel breeding design which randomly recombines the genomes of the eight founders, as shown below. A resulting inbred CC strain is a mosaic of these genomes resulting from 2 crosses (G1 and G2), followed by at least 20 generations of inbreeding to fix the genome leading to a new, reproducible, inbred line. Over 300 lines have undergone more than five generations of inbreeding. Given the founder haplotypes and an extant genotype, we investigate the problem of inferring the haplotype ancestry of the descendant genotype.

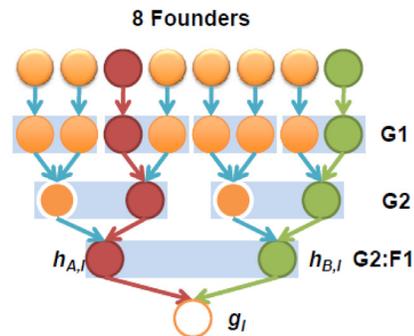


Breeding constraints that are induced by the CC breeding structure at generation G2:F1 are obvious. For a single site, it is impossible to have founder ancestry either solely from the left 4 founders or solely from the right 4 founders. At later generations, the founder ancestry pairs seen in generation G1 never reappear.

Methods

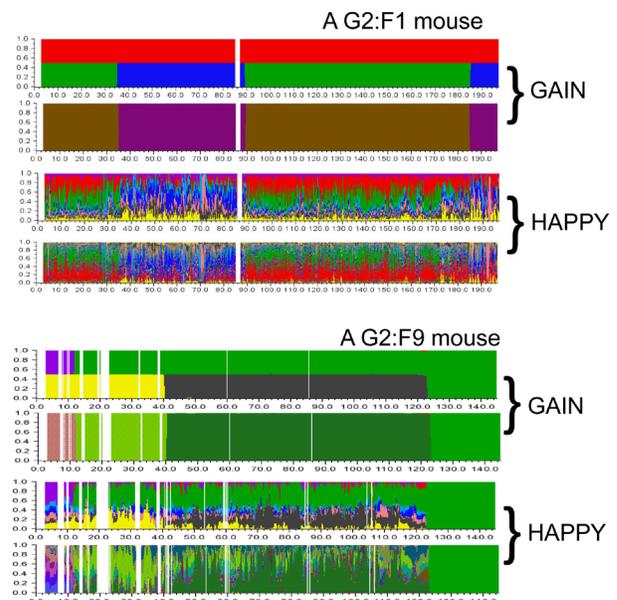
1. A Dynamic Programming approach that constructs an optimal mosaic solution by minimizing the number of recombinations. Breeding constraints are explicitly coded as forbidden choices in this approach.

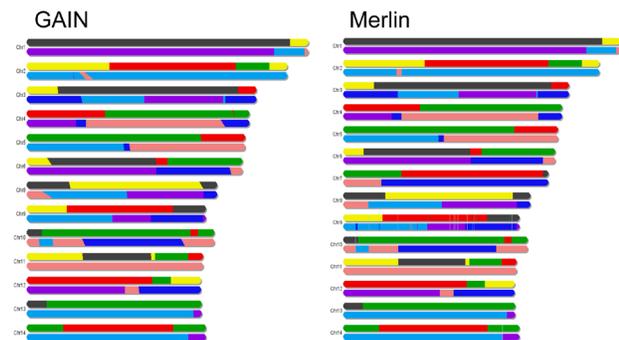
2. A Hidden Markov Model (HMM) that extends the classic Lander-Green approach to handle any number of inbreeding generations. The breeding constraints are naturally supported through representing inheritance vector as hidden states.



Experiment Results

We have applied these methods to more than 200 Pre-CC lines. By resolving ambiguities using breeding constraints, our method (GAIN) provides much cleaner results than alternative methods that do not consider pedigree (e.g., HAPPY). Compared with Merlin (the state of the art implementation of Lander-Green approach), GAIN infers highly consistent mosaic structures in much shorter time.





We provide a complete framework to investigate the genome structure of CC lines. We have been able to detect funnel and breeding errors and extend association mapping techniques.

Current Project Members

Wei Wang, faculty member

Leonard McMillan, faculty member

Fernando Pardo-Manuel de Villena, collaborator, UNC Department of Genetics

Qi Zhang, collaborator, Biostatistics, University of Washington

Eric Yi Liu, graduate research assistant

Research Sponsors

NSF IIS 0448392, NSF IIS 0812464

NIH U01 CA134240, NIH P50 MH090338-01

Related Publications

Zhang et al. "Genotype sequence segmentation: handling constraints and noise," *WABI*, pp. 271-283, 2008.

Liu et al. "Efficient genome ancestry inference in complex pedigrees with inbreeding," *Bioinformatics*, vol. 26, no. 12, pp. 199-207, 2010.

For More Information

Wei Wang

Professor

919-962-1744

weiwang@cs.unc.edu