Study of crossover detection using RTIGER on low coverage snRNA-seq data from *Arabidopsis thaliana* F1 hybrids

Masters in Bioinformatics - Spring 2026

Johan Olesen
202104408
Msc. Student Bioinformatics,
Aarhus University

Thomas Bataillon
Supervisor
Professor, Department of
Molecular Biology and Genetics - BiRC,
Aarhus University

12.06.2026

Abstract

Meiotic crossover (CO) events are a major source of genetic variation and essential for chromosome segregation. Accurate mapping of CO positions at scale is critical for understanding the mechanism they abide by for their number, distribution and interference. Conventional approaches based on F2 or recombinant inbred line (RIL) populations require growing and individually genotyping many plants across multiple generations, thereby limiting throughput. Single-nucleus RNA sequencing (snRNA-seq) of gametes offers an alternative, enabling thousands of recombinant haplotypes to be harvested and sequenced in a single experiment, by exploiting the natural product of meiosis. RTIGER is a rigid hidden Markov model method, designed to predict COs from allele count data, but its performance on real snRNA-seq data have not been evaluated.This thesis evaluates RTIGER for CO detection using snRNA-seq pollen data from a Col-0 x Db-1 *Arabidopsis thaliana* cross hybrid. Two separate datasets were analysed to characterise SNP marker coverage and density across chromosome 4. To assess the main determinants for detection accuracy, simulated datasets were generated from the real data with known CO positions. Allowing precision, recall and positional accuracy to be quantified across various conditions. SNP density and SNP coverage emerged as the primary factors in determining performance, while sequencing depth and cell number were not limiting factors.Beyond testing RTIGER, this thesis produced a fully reproducible computational framework including automated data acquisition, software installation and management, Snakemake-based workflow execution and simulation based validation. Together, the results demonstrate that RTIGER is a robust method for CO detection from snRNA-seq pollen data, but that its performance is fundamentally constrained by the distribution and coverage of informative marker SNPs.

Abbreviations

RNA: ribonucleic acid

SNP: single nucleotide polymorphism

HMM: hidden Markov model

CO: crossover

DNA: deoxyribonucleic acid

DSB: double-strand break

DSBR: double-strand break repair

dHJ: double Holliday junction

SDSA: synthesis-dependant strand annealing

MMR: mismatch repair

gBCG: GC-biased gene conversion

rHMM: rigid hidden Markov model

EM: expectation–maximisation

SyRI: Synteny and Rearrangement Identifier

HPC: high performance computer

ORF: open reading frame

snRNA-seq: single nuclear RNA sequencing

RIL: recombinant inbred line

Units

Mb: mega bases

cM: centimorgan

Contents

1 Introduction 1
1.1 Recombination 2
1.1.1 Crossover 2
1.1.2 Genetic conversion / Non-Crossover 4
1.2 Plant model *Arabidopsis thaliana* 4
1.3 RTIGER 6
2 Methods 9
2.1 Data availability 9
2.2 Technical decisions 10
2.2.1 Workflow management with Snakemake 10
2.2.2 `rtigertools` R package structure 10
2.2.3 Environment and dependency challenges 11
2.2.4 Reproducibility design principles 12
2.3 Focus on Chromosome 4 of Col-0 x Db-1 F1 hybrid cells 12
2.4 Low coverage snRNA-seq data analysis 13
2.5 Simulating crossover events 16
2.5.1 Conceptual framework and decisions 16
2.5.2 Models 18
2.5.2.1 No-Interference Model (Geometric Waiting Times) 18
2.5.2.2 Counting / Chi-Square (Stahl) Model 19
2.5.2.3 SC Diffusion / Coarsening Model 20
2.5.2.4 Model comparison 21
2.5.3 Simulation results 23
2.6 Global counts 25
3 Discussion 29
4 Conclusion 34
References 35
Appendix A Source code for models i
A.1 No-Interference Model (Geometric Waiting Times) i
A.2 Counting / Chi-Square (Stahl) Model i
A.3 SC Diffusion / Coarsening Model ii
Appendix B Genotype examples iv
B.1 col0xdb1 iv

B.2 col0xmany v
Appendix C RTIGER performance metrics vii
GAI Declaration viii