Genotyping-by-sequencing of thousands of individuals

Vipul Patel, Geo Velikkakam James, Ales Pecinka, Korbinian Schneeberger

Max Planck Institute for Plant Breeding Research, Cologne, Germany

Like in every other genetically tractable species, genetic mapping starts with the design of mapping populations. During the establishment, recombination introduces a unique combination of parental alleles within the genome of each of the progenitors. Following the principles of conventional mapping all individuals of these populations are genotyped for known sequence differences, like single nucleotide polymorphism (SNPs). Those then can be used to reconstruct the recombination events and finally allow for association analysis between genotypes and phenotypes. The resolution of these predictions though depends on the density of genetic marker. Whole genome sequencing would not only by-pass time-intense genotyping efforts but would also allow for the reconstruction of recombination breaks at great resolution as essentially all markers could be genotyped at the same time. However, complex trait analyses requires populations of more than hundreds or even thousands of individuals. For this amount of individuals applying standard resequencing methods would still require huge sequencing investment.
A solution to avoid this, but still apply whole genome sequencing, is low-fold sequencing enabled through pooling of multiple individuals in one sequencing reaction. For this, the DNA of each individual is tagged by a DNA barcode allowing for a unique assignment of sequence reads to samples. This reduces the sequencing cost per individual, but comes at the price of having only partial genomes sequenced. It has already been shown that the minimum of one twenty-fifth (0.04x) of a genome worth of sequencing data is enough for genotyping at a recombination breakpoint resolution of 50kb (1). As the current data yield of one lane of an Illumina HiSeq 2000 machine can be as large as the equivalent of 400 *Arabidopsis thaliana* genomes. 10,000 individuals could be genotyped of this species in one reaction, in theory. Sophisticated genotyping-by-sequencing methods dealing with variable, sparse and sequencing error-prone data have been introduced but it remains unclear how well they perform (2, 3). The analysis of recombination breakpoints that have been missed and the effect of coverage on the resolution of recombination breakpoint detection are yet to be determined as these can only be assessed for some of the breakpoints that have been detected.
Here we introduce a novel Hidden Markov Model (HMM) based approach, which is designed for recombination breakpoint prediction of low-fold sequenced individuals. For this we have established *in-silico* simulated mapping population for training and testing of our HMM. As our simulations follow the rules of Mendelian inheritance and generate individual genomes based on empirically determined data for recombination location and frequency, we expect them to be very close to real populations. Like-wise, sequencing-specific coverage biases and sequencing errors are considered for simulating the result of whole genome sequencing.
We assessed the quality of our model with a 10-fold cross validation using 5,000 plants and 331,000 SNP markers. Depending on the coverage, we are hoping to reach a significantly better breakpoint resolution than the genetic resolution provided by our mapping populations. We have observed missing recombination breakpoints in our prediction though those were rare and only occurred at very low coverage levels. Currently we are extending our model for the analysis of mapping populations based on multiple parental genotypes.

References:

1. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-Throughput genotyping by whole-genome resequencing. Genome Res 2009, Jun;19(6):1068-76.

2. Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res 2011, Jan 13.

3. Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, Yu S, Han B, Zhang Q (2010) Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc Natl Acad Sci USA 107: 10578-10583