Estimating the allele frequency spectrum and detecting selective sweeps from pooled next generation sequencing samples

Simon S. Boitard, Christian Schlötterer, Robert Koffler, Ram Vinay Pandey, Andreas Futschik

Estimating the allele frequency spectrum and detecting selective sweeps from pooled next generation sequencing samples.

Simon Boitard[1], Christian Schlötterer[2], Robert Kofler[2], Ram Vinay Pandey[2], Andreas Futschik[3]. (1) Laboratoire de Génétique Cellulaire, INRA, Toulouse, France. (2) Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria. (3) Institute of Statistics and Operations Research, University of Vienna, Vienna, Austria.

Summary

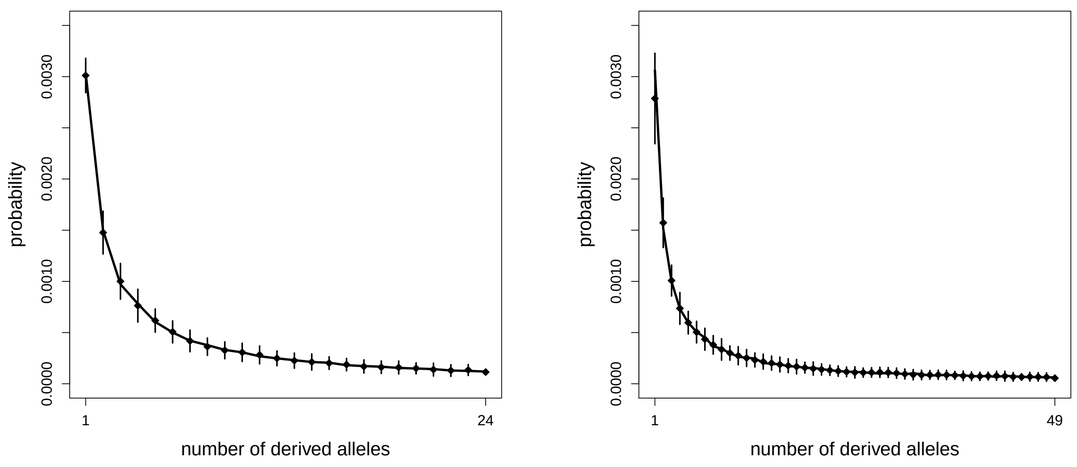
Due to its cost effectiveness, next generation sequencing of pools of individuals (Pool-Seq) is becoming a popular strategy for genome-wide estimation of allele frequencies in population samples. Since the allele frequency spectrum (AFS) provides information about past episodes of selection, Pool-Seq is also a promising design for genomic scans for selection. Here, we introduce a statistical method for estimating the AFS and detecting selective sweeps in a Pool-Seq sample. This method accounts for the uncertainty concerning the true allele frequencies in the pool, which might typically be higher for sites with low coverage or bad quality scores. We apply our method to simulated data at low coverage ($0.5\times$ per chromosome), and to real data from *Drosophila melanogaster*. A python program implementing our method, denoted **Pool-hmm**, is freely available at **https://qgp.jouy.inra.fr/**.

Modelling the Pool-Seq data

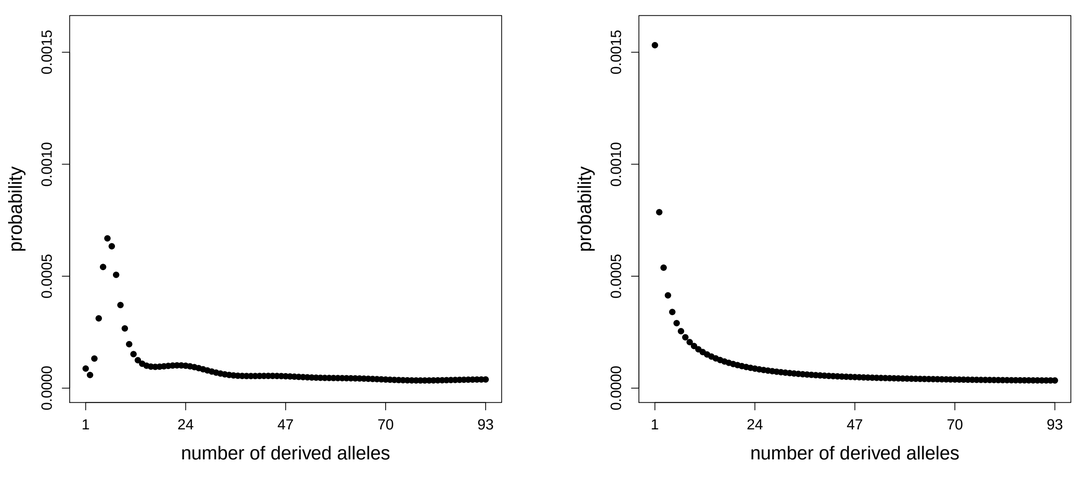
Consider a pool of n chromosomes from the same population. At position i on the genome, let Y_i be the number of derived alleles in the pool ($0 \leq Y_i \leq n$) and r_i be the number of reads covering the position. Observed read data at position i is written $Z_i = (Z_{i,1}, \ldots, Z_{i,r_i})$, where $Z_{i,j}$ is equal to 1 (0) if the allele at read j is derived (ancestral). If $p_{i,j}$ is the probability for a sequencing error at read j (deduced from quality score),

$$\mathbb{P}(Z_i \mid Y_i) = \prod_{j:Z_{i,j}=1} ((1-p_{i,j})\frac{Y_i}{n} + p_{i,j}(1-\frac{Y_i}{n})) * \prod_{j:Z_{i,j}=0} ((1-p_{i,j})(1-\frac{Y_i}{n}) + p_{i,j}\frac{Y_i}{n})$$

Estimation of the AFS



Simulation results. 100 pools of $n = 25$ (left) and $n = 50$ (right) chromosomes of length $L = 100$kb were simulated under the coalescent with $\theta = 0.003$ and $\rho = 0.003$. Solid lines (diamonds) show the average AFS estimated from the complete sequences (Pool-Seq data at $100\times$ coverage). **If quality scores are reliable, our estimation of the AFS is very accurate.**



Folded AFS on the X chromosome of *D. melanogaster*. Estimated from a pool of 97 flies sequenced at $100\times$ coverage, using all base calls (left) or only those with PHRED score greater than 35 (right). **The AFS is biased because quality scores overestimate the sequencing error probablities.**

Detection of selective sweeps - Methods

Detection of selective sweeps is based on a Hidden Markov Model. Each position i is assumed to have an hidden state X_i, with 3 possible values : "Selection", "Intermediate" or "Neutral". These 3 values are associated to different AFS. The "Neutral" AFS is estimated from whole genome data. The "Intermediate" and "Selection" AFS are derived from the "Neutral" AFS as in (Nielsen et al., 2005). To account for the uncertainty concerning the true allele frequencies in the pool, emission probabilities are

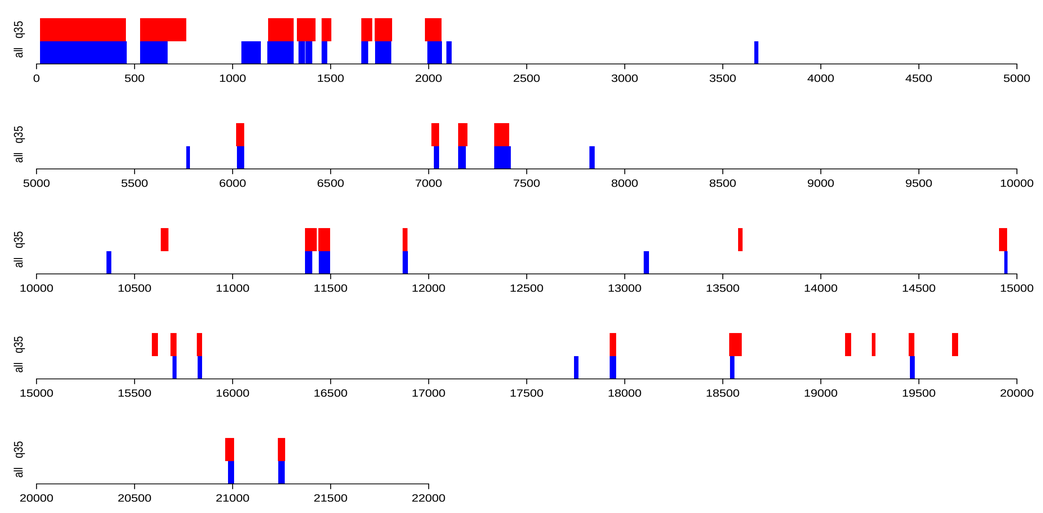
$$\mathbb{P}(Z_i \mid X_i) = \sum_{Y_i=0}^{n} \mathbb{P}(Z_i \mid Y_i)\, \mathbb{P}(Y_i \mid X_i)$$

Hidden states (and consequently selective sweeps) are predicted from the observed Pool-Seq data using the Viterbi algorithm.

Detection of selective sweeps - Results

Sample size	$n = 25$	$n = 50$	$n = 100$	$n = 200$
Pool-Seq data	0.91	0.90	0.91	0.90
Sequence data	0.89	0.88	0.87	0.87

Simulation results. Samples were simulated with selection intensity $s = 500$, other parameters are as in the left box. **Pool-Seq data and complete sequence data provide the same detection power.**



Selective sweeps on the X chromosome of *D. melanogaster*, detected using all base calls (blue) or base calls with PHRED score greater than 35 (red). **Sweep detection is not affected by the bias of the "Neutral" AFS.**

References

Boitard, S., Schlötterer, C., Nolte, V., Pandey, R., Futschik, A., 2012. Detecting selective sweeps from pooled next generation sequencing samples. Mol. Biol. Evol., doi: 10.1093/molbev/mss090.

Nielsen, R., Williamson, L., Kim, Y., Hubisz, M., Clark, A., Bustamante, C., 2005. Genomic scans for selective sweeps using SNP data. Genome Research 15, 1566–1575.