A statistical base-calling methodology for the Illumina high-throughput DNA sequencing platform

Wally Gilks

Department of Statistics, University of Leeds

We propose statistical base-calling methodology for use with the Illumina Genome Analyzer, a high-throughput next-generation DNA sequencing platform. This methodology takes account of cross-talk between dye labels and three problems which accrue in clusters of DNA sequence over successive cycles of the sequencer: accumulation of dye reactants (the "sticky T" problem); base-incorporation errors (the "phasing" problem); and terminal failure of the sequencing reactions (the problem of "drop-off"). Our methodology allows for differential rates of phase inaccuracy and drop-off between clusters, and calculates the probability of miscall for each base called. The resulting base-calling algorithm is linear in the number of cycles. We evaluate the performance of this algorithm, and compare its base-calling accuracy to that of the Illumina pipeline.