Online Appendix: Simulation Details

Simulation: We simulated a dataset of 10,000 consecutive probes on a single chromosome. Let S be the indicator for copy number at a probe with S=0 indicating baseline copy number. In the region covered by these probes, we modeled 10 copy-number variable loci (CNV) covering 10 probes each. Each CNV was assigned with equal probability to be a duplication (S=1) or a deletion (S=-1). We assumed that for each CNV, 10% of the sample carried minor allele (S=+/- 1).We then generated a sample of 100 individuals, randomly assigning copy number state for each CNV to each individual. We then sampled probe intensities for every probe from a N(μ_S, σ) distribution with mean $\mu_S = S\delta\sigma$. Here δ represents a effect size parameter; large values of δ indicate an experimental system where copy number variable probes are more easily discernable.

Analysis: We applied two methods to analyze each dataset. First, we performed a standard Hidden Markov Model (Fridlyand J, 2004). We modeled the data with three hidden states and assumed that the intensity distribution for each state was known. We performed a Baum-Welch forward-backward algorithm (Baum LE, 1970) with 30 iterations to estimate the transition matrix. We observed convergence of the transition matrix after 5 iterations. Based on the estimated transition matrix we then applied a Viterbi-algorithm (Viterbi, 1967) to estimate CNV-status for each probe. Computation time of this analysis was approximately 1 min on a 2.33 GHz processor.

Second, we applied CopyMap, setting the minimum length of each CNV to 5 and assigning a prior probability p=0.005 to both non-baseline statuses. Copymap uses a modified Baum-Welch algorithm; we performed 30 iterations to estimate the locus-specific transition rate, observing convergence of the overall likelihood after 10 iterations. To maintain comparability with the first method, we used a Viterbi-algorithm to estimate CNV status of each probe using the site-specific transition rates. Computation time of this analysis was approximately 8.5 min on a 2.33 GHz processor.

To evaluate the performance of each algorithm, we defined each consecutive set of probes inferred to be in the same non-baseline copy number state as one CNV. We compared the location and copy number state of each inferred CNV with the location and copy number state of each simulated CNV. Inferred CNVs that overlapped with at least one CNV of the same copy number state in the same individual were considered correct calls, otherwise inferred CNVs were considered false calls. We calculated the precision of correct calls at estimating the boundary of a CNV the following way: Assume probes are numbered 1,...,10000 and let F_T and L_T be the first and the last probe covering the simulated CNV. Moreover let F_I and L_I be the first and the last probe covering the inferred CNV. The boundary error is then calculated by $B = |F_T - F_I| + |L_T - L_I|$.

For each value $\delta \in \{0.5, 0.75, 1, 1.125, 1.25, 1.375, 1.5, 1.75, 2, 2.25, 2.5\}$ we repeated this analysis 50 times. Figure 1A summarizes the average number of correct calls and false calls; Figure 1B summarized the average boundary error of the correct calls.

Bibliography

Baum, L. E. et al.(1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Statist , 41*, 164-71.

Fridlyand, J. et al.(2004). Hidden Markov models approach to the analysis of array CGH data. *J Multivariate Anal , 90*, 132-53.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory , 13*, 260-9.