Supplementary information to CGHcall

Mark A. van de Wiel, Kyung In Kim, Sjoerd J. Vosse, Wessel N. van Wieringen, Saskia M. Wilting and Bauke Ylstra

1 Methods

1.1 Assumptions

Firstly, the mean levels of the six states are shared by the profiles. This assumption implies the use of tumor samples with reasonably similar proportions of tumor cells or knowledge of these proportions so that these can be corrected for (see Discussion). Secondly, variability within a segment is assumed to be constant within a profile, but may differ between independent array CGH profiles to allow for experimental variation such as DNA quality. Thirdly, to save a parameter in the fitting procedure, it is assumed the (average) log-ratio of a double gain (four copies) is proportional to that of a single gain (three copies): $\gamma_5 = \log_2(4/2)/\log_2(3/2) * \gamma_4 = 1.7 * \gamma_4$. Fourthly, $\gamma_4 - \gamma_3 > 0.1, \gamma_3 - \gamma_2 > 0.1, \gamma_2 - \gamma_1 > 0.2$ and $\gamma_6 - \gamma_5 > 0.2$, thereby separating the modes of the six states, because these represent discrete levels. The exact value of these separation lower bounds is not so crucial and, in general, estimated values were away from these bounds. Finally, $\tau_1 \geq \tau_2$ and $\tau_6 \geq \tau_5 \geq \tau_4$ is enforced to induce higher variability of segment-wise log-ratio levels for higher level aberrations such as amplifications.

1.2 Derivation of log-likelihood

Here, we derive the log-likelihood function $\mathcal{L}(\Theta|\{X_{ijk}\}, \{Y_{jk}\})$. Note that the σ_k's are fixed. The model (see also form. (1) in the article):

$$\begin{aligned} X_{ijk} &\sim N(\mu_{jk}, \sigma_k^2) \\ \mu_{jk} &\sim \sum_{\ell=1}^{6} p_\ell N(\gamma_\ell, \tau_\ell^2) \end{aligned} \tag{1}$$

may be re-written as

$$\begin{aligned} \Theta &= (p_1, p_2, \ldots, p_6, \gamma_1, \ldots, \gamma_6, \tau_1, \ldots, \tau_6) \\ (Y_{jk}|\Theta) &\sim \mathcal{M}(p_1, \ldots, p_6) \\ (\mu_{jk}|Y_{jk}, \Theta) &\sim N(\gamma_{Y_{jk}}, \tau_{Y_{jk}}^2) \\ (X_{ijk}|\mu_{jk}) &\sim N(\mu_{jk}, \sigma_k^2), \end{aligned}$$

with $\mathcal{M}(p_1, \ldots, p_6)$: the multinomial distribution. The joint density of $(\{X_{ijk}\}, \{\mu_{jk}\}, \{Y_{jk}\})$ is

$$\begin{aligned}
& P(\{X_{ijk}\}, \{\mu_{jk}\}, \{Y_{jk}\}|\Theta) \\
= \; & P(\{X_{ijk}\}|\{\mu_{jk}\}, \{Y_{jk}\}, \Theta) P(\{\mu_{jk}\}|\{Y_{jk}\}, \Theta) P(\{Y_{jk}\}|\Theta) \\
= \; & P(\{X_{ijk}\}|\{\mu_{jk}\}) P(\{\mu_{jk}\}|\{Y_{jk}\}, \Theta) P(\{Y_{jk}\}|\Theta) \\
= \; & \left(\prod_{j,k} \prod_{i} P(X_{ijk}|\mu_{jk}) \right) \left(\prod_{j,k} P(\mu_{jk}|Y_{jk}, \Theta) \right) \left(\prod_{j,k} P(Y_{jk}|\Theta) \right) \\
= \; & \prod_{j,k} \left(P(\mu_{jk}|Y_{jk}, \Theta) P(Y_{jk}|\Theta) \prod_{i} P(X_{ijk}|\mu_{jk}) \right),
\end{aligned}$$

where index i ranges from 1 to I_{jk}, the number of clones for segment j in profile k. Hence, the likelihood of $(\{X_{ijk}\}, \{Y_{jk}\})$ equals

$$\begin{aligned}
& \prod_{j,k} \left(\int P(\mu_{jk}|Y_{jk}, \Theta) P(Y_{jk}|\Theta) \prod_{i} P(X_{ijk}|\mu_{jk}) d\mu_{jk} \right) \\
= \; & \prod_{j,k} \left(P(Y_{jk}|\Theta) \int P(\mu_{jk}|Y_{jk}, \Theta) \prod_{i} P(X_{ijk}|\mu_{jk}) d\mu_{jk} \right).
\end{aligned}$$

The log-likelihood of $(\{X_{ijk}\}, \{Y_{jk}\})$ is

$$\begin{aligned}
& \mathcal{L}(\Theta|\{X_{ijk}\}, \{Y_{jk}\}) \\
= \; & \sum_{j,k} \left(\log P(Y_{jk}|\Theta) + \log \int P(\mu_{jk}|Y_{jk}, \Theta) \prod_{i} P(X_{ijk}|\mu_{jk}) d\mu_{jk} \right) \\
= \; & \sum_{j,k} \left(\log p_{Y_{jk}} + \log \int \phi(\mu_{jk}|\gamma_{Y_{jk}}, \tau_{Y_{jk}}) \prod_{i} \phi(X_{ijk}|\mu_{jk}, \sigma_k) d\mu_{jk} \right) \\
= \; & \sum_{j,k} \left(\log p_{Y_{jk}} + \log h(\{X_{ijk}\}|\gamma_{Y_{jk}}, \tau_{Y_{jk}}, \sigma_k) \right),
\end{aligned}$$

where $\phi(x|\mu, \sigma)$ is the Normal(μ, σ^2) density evaluated at x and h is an explicit function, because the integral can be computed analytically as follows. Set $Y_{jk} = \ell$ and $I_{jk} = n$. Then,

$$h(\{X_{ijk}\}|\gamma_\ell, \tau_\ell, \sigma_k) = \int \phi(\mu_{jk}|\gamma_\ell, \tau_\ell) \prod_{i} \phi(X_{ijk}|\mu_{jk}, \sigma_k) d\mu_{jk} \tag{2}$$

where

$$\begin{aligned}
& \phi(\mu_{jk}|\gamma_\ell, \tau_\ell) \prod_{i} \phi(X_{ijk}|\mu_{jk}, \sigma_k) \\
= \; & \frac{1}{\sqrt{2\pi}\tau_\ell} \exp(-\frac{(\mu_{jk} - \gamma_\ell)^2}{2\tau_\ell^2}) \prod_{i}^{n} \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{(X_{ijk} - \mu_{jk})^2}{2\sigma_k^2}) \\
= \; & \frac{1}{(2\pi)^{(n+1)/2} \tau_\ell \sigma_k^n} \exp\left(-\frac{\mu_{jk}^2 - 2\gamma_\ell \mu_{jk} + \gamma_\ell^2}{2\tau_\ell^2} - \frac{n\mu_{jk}^2 - 2(\sum_i X_{ijk})\mu_{jk} + \sum_i X_{ijk}^2}{2\sigma_k^2} \right)
\end{aligned} \tag{3}$$

Write (3) as

$$\frac{1}{(\pi B)^{1/2}} \exp(-(\mu_{jk} - A)^2/B) * C * \exp(D), \tag{4}$$

where A and B are found by equating the linear and quadratic μ_{jk} terms to those in (3) and then C and D are found by equating the constants outside and inside the exponential to those in (3). Let $v_k = 2\sigma_k^2$ and $t_\ell = 2\tau_\ell^2$. Then

$$\begin{aligned} A &= (\sum_i X_{ijk} * t_\ell + \gamma_\ell * v_k)/(n * t_\ell + v_k) \\ B &= v_k * t_\ell/(n * t_\ell + v_k) \\ C &= B^{1/2} * (t_\ell^{1/2}(2\pi)^{n/2}\sigma_k^n)^{-1} \\ D &= -(\sum_i X_{ijk}^2/v_k + \gamma_\ell^2/t_\ell - A^2/B) \end{aligned}$$

Finally, note that the first part in (4) is a Normal pdf and C and D do not depend on μ_{jk}. Hence, after substitution into (2) the integral results in $C * \exp(D)$ which, given the parameters, depends only on $\sum_{i=1}^{I_{jk}} X_{ijk}$ and $\sum_{i=1}^{I_{jk}} X_{ijk}^2$.

1.3 EM algorithm

The EM algorithm [2] is a standard algorithm for fitting mixture models. Here, the mixture model is formulated on the level of segments, so we force clones in the same segment to share the underlying hidden state. Also, the second level of model (1), the Normal clone variation within a segment, needs to be incorporated. Below, the log-likelihood function is presented, which is a central entity for the EM algorithm. Denote the vector of unknown parameters by $\Theta = (p_1, \ldots, p_6, \gamma_1, \ldots, \gamma_6, \tau_1, \ldots, \tau_6)$. Three-dimensional array $\{X_{ijk}\}$ contains all clone data X_{ijk} and denote the hidden state for segment j in profile k by Y_{jk}. We assume independency between segments and conditional independence between clones within a segment (given the hidden state). Then, the log-likelihood with respect to the data $\{X_{ijk}\}$ and the vector of hidden states $\{Y_{jk}\}$ (see above for its derivation) is

$$\mathcal{L}(\Theta|\{X_{ijk}\}, \{Y_{jk}\}) = \sum_{k=1}^{K} \sum_{j=1}^{J_k} \left(\log p_{Y_{jk}} + \log h(\{X_{ijk}\}|\gamma_{Y_{jk}}, \tau_{Y_{jk}}, \sigma_k) \right), \tag{5}$$

where J_k is the number of segments in profile k and h is an explicit function, which, given the parameters, depends only on $\sum_{i=1}^{I_{jk}} X_{ijk}$ and $\sum_{i=1}^{I_{jk}} X_{ijk}^2$, where I_{jk} is the number of clones within the segment j of tumor k. Using log-likelihood $\mathcal{L}(\Theta|\{X_{ijk}\}, \{Y_{jk}\})$ the EM algorithm is applied: initialize $\Theta = \Theta_0$, compute membership probabilities $P(Y_{jk} = \ell|\{X_{ijk}\}, \Theta_0)$ by applying the Bayes' rule, compute the expectation of $\mathcal{L}(\Theta|\{X_{ijk}\}, \{Y_{jk}\})$ with respect to $\{Y_{jk}\}$, maximize it with respect to the Normal parameters and the mixture proportions and iterate until the unknown parameters hardly change. After convergence, the main quantity of interest is the probability that segment j in profile k belongs to state ℓ given the data and the estimate $\hat{\Theta}$: $P_{jk\ell} = P(Y_{jk} = \ell|\{X_{ijk}\}, \hat{\Theta})$. Initialization is crucial for EM, because the algorithm is known to converge to a local optimum, which may be wrong if initial parameter values deviate several factors from the optimal ones. Fortunately, the location of the deletion, normal and gain modes $(\gamma_2, \gamma_3, \gamma_4)$ can reasonably approximated as follows. After proper

normalization γ_3 should be close to 0. It is initialized as the median of all segment-wise means between -0.15 and 0.15. The largest proportion of deletion data from all profiles most often corresponds to a single deletion, so γ_2 is initialized as the median of all segment-wise means between -1.0 and -0.15. The single gain mode (γ_4) is initialized analogously. With the same intervals, we estimated the corresponding standard deviations (τ_2, τ_3, τ_4) by Median Absolute Deviations (MAD) corrected for Normal distributions. Similar considerations where applied to initialize the other Normal parameters. The proportion parameters $p_\ell, \ell = 1, \ldots, 6$, were initialized as $(0.01, 0.09, 0.8, 0.08, 0.01, 0.01)$; a uniform configuration gave similar results, but with slightly more EM-iterations, in general.

Special care was taken to compute the membership probabilities $P_{jk\ell}$. Applying Bayes' rule gives:

$$P_{jk\ell} = P(Y_{jk} = \ell | \{X_{ijk}\}, \hat{\Theta}) = \frac{\exp\left(\mathcal{L}(\hat{\Theta}|\{X_{ijk}\}, Y_{jk} = \ell)\right) p_\ell}{\sum_{m=1}^{6} \exp\left(\mathcal{L}(\hat{\Theta}|\{X_{ijk}\}, Y_{jk} = m)\right) p_m}. \tag{6}$$

So, computation of $P_{jk\ell}$ requires the likelihood per segment given hidden state Y_{jk}, which equals the exponential of the the last summand in (5). However, when the number of clones in the segment I_{jk} is large, the result may become extremely small, since one basically multiplies many density functions evaluated at the data. Since $P_{jk\ell}$ is computed with the Bayes' rule and the denominator is a weighted sum of these likelihoods this can cause numerical instability. To avoid this, write $\exp(\mathcal{L}(\hat{\Theta}|\{X_{ijk}\}, Y_{jk} = \ell)) = \exp(-L_\ell)$. Then, we factorize the segment likelihood as $\exp(-L_\ell) = \exp(-M) * \exp(-(L_\ell - M))$, where $M = \min(L_1, \ldots, L_6)$. Then, the factor $\exp(-M)$ cancels out in the computation of $P_{jk\ell}$ using equation (6). Since $\exp(-(L_\ell - M))$ equals 1 for at least one ℓ, say $\ell = q$, the denominator in (6) is larger or equal to mixture proportion p_q.

1.4 Classification rules

The six states in the model are used to reflect the nature of the data as good as possible. However, final classification is into three or four classes (user specified). The posterior probabilities $P_{jk\ell}, \ell = 1, \ldots, 6$, are used as follows for classification into four classes. Segment j in profile k is assigned to:

- class 'loss' (-1) if $P_{jk1} + P_{jk2} > 0.5$
- class 'normal' (0) if $P_{jk3} \geq 0.5$
- class 'gain' (1) if $P_{jk4} + P_{jk5} + P_{jk6} > 0.5$ and $P_{jk6} < 0.5$
- class 'amplification' (2) if $P_{jk6} \geq 0.5$.

In case of three class classification, the 'amplification' class is dropped, as is the second condition for the 'gain' class.

1.5 Likelihood for the model with chromosome arm mixture proportions

Model (1) becomes

$$\begin{aligned} X_{ijk} &\sim N(\mu_{jk}, \sigma_k^2) \\ \mu_{jk} &\sim \sum_{\ell=1}^{6} p_{\ell,a_{jk}} N(\gamma_\ell, \tau_\ell^2), \end{aligned} \tag{7}$$

where a_{jk} is the chromosome arm to which segment j in profile k belongs. This model may be re-written as

$$\begin{aligned} \Theta &= (p_{1,1}, p_{2,1}, \ldots, p_{6,1}, \ldots, p_{1,A}, p_{2,A}, \ldots, p_{6,A}, \gamma_1, \ldots, \gamma_6, \tau_1, \ldots, \tau_6) \\ (Y_{jk}|\Theta) &\sim \mathcal{M}(p_{1,a_{jk}}, \ldots, p_{6,a_{jk}}) \\ (\mu_{jk}|Y_{jk}, \Theta) &\sim N(\gamma_{Y_{jk}}, \tau_{Y_{jk}}^2) \\ (X_{ijk}|\mu_{jk}) &\sim N(\mu_{jk}, \sigma_k^2), \end{aligned}$$

where A is the number of chromosome arms. Then the Log-Likelihood is

$$\mathcal{L}(\Theta|\{X_{ijk}\}, \{Y_{jk}\}) = \sum_{k=1}^{K} \sum_{j=1}^{J_k} \left(\log p_{Y_{jk},a_{jk}} + \log h(\{X_{ijk}\}|\gamma_{Y_{jk}}, \tau_{Y_{jk}}, \sigma_k) \right). \tag{8}$$

Estimation of the unknown parameters is completely analogous to that of the parameters of the standard model. The membership probabilities for Y_{jk} are computed as in (6), with the log-likelihood as in (8) and p_ℓ replaced by $p_{\ell,a_{jk}}$.

1.6 Log-ratios corrected for contamination by normal cells

As a caution, we note that samples with very different proportions of tumor cells should not be called simultaneously, because the difference in proportions impacts the aberration levels [4]. This problem is absent when micro-dissection techniques have been used to guarantee pure tissue samples. Alternatively, the proportion of tumor cells can be estimated and may be corrected for as follows.

Assume a signal proportional with factor f to the number of copies and assume 2 copies for the normal tissue, then we have with proportion of tumor cells c, signal Sf for the contaminated tissue and $\log_2(A/2)$ the copy number log-ratio of interest $\log_2\big(Sf/(2f)\big) = \log_2\big((cAf + 2(1-c)f)/2f\big)$, so

$$\log_2(A/2) = \log_2\big(S/(2c) - (1-c)/c\big) = \log_2\big(R/c - (1-c)/c\big),$$

where R is the measured signal ratio.

2 Results

2.1 Simulation tests

To assess our method we performed a simulation study. Therefore, array CGH data are simulated as by Willenbrock and Fridlyand [11]. These were claimed to be more realistic than

those in previous studies, because real data were used to emulate variable copy numbers and segment lengths. Even though CGHcall assumes that tumor samples have similar proportions of tumor cells or that these are corrected for, the proportions are allowed to vary between 0.5 and 0.7 (hence not particularly high) for the simulated profiles, which may reflect the uncertainty in the estimated proportions. The fractions corresponding to the six states used in [11] are 0.02, 0.12, 0.70, 0.10, 0.04 and 0.02, so 70% of the segments is normal. Since normal segments were longer on average, 85% of the simulated clones is normal. Fifty samples with 1000 data points were generated and CGHcall was applied to these simulated data. Classification accuracy, defined as the percentage of clones for which the classification agreed with the truth, was used as a performance metric. CGHcall attained 99.6% classification accuracy when data were classified into three classes. Of the true normals (clones included in a segment simulated from the normal state), 99.8% was classified as such, while 98.5% of the true aberrations was correctly classified. A mis-prediction is usually an individual clone which the segmentation algorithm assigned to a segment neighboring the true segment containing that clone. So, the classification accuracy of CGHcall in this simulation set-up is extremely high. Therefore, we decided to study also a more challenging situation before comparing CGHcall to other methods.

While the previously simulated data may be realistic for high quality DNA data sets, our experience is that profiles regularly contain more segment levels which are less distinct. This may be due to the type of tumor or the use of DNA isolated from paraffin embedded material which is usually of lower quality than fresh frozen DNA material [9]. Therefore, another simulation was run with segment-wise error added to the signals. Standard deviations for the six states used in [11] were set to $(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6) = (0.25, 0.10, 0.07, 0.11, 0.16, 0.3)$. The magnitude of these parameters is based on estimates from several real data sets. For these more 'diffused' data, we compared CGHcall with DNACopy-MergeLevels [11], CGHclassify [3], CLAC [10] and the '2sd-rule' [5]. The results are displayed in Table 1. CGHcall achieved a classification accuracy of 98.6%, higher than that of any of the other methods. Of the true normal clones, 99.3% was classified as such; while CGHclassify, CLAC and 2sd achieved similar levels of the true negative rate, DNAcopy-MergeLevels achieved less than 90% (hence a high false negative rate). On the other hand, only DNAcopy-MergeLevels achieved a similar true positive rate as the 95.8% attained by CGHcall; all the other methods attained much lower levels of the true positive rate. Hence, for this simulation set-up, CGHcall outperformed the other methods. CGHcall plots for the data with and without additional segment-wise error are accessible via the website.

Method	CGHcall	DNAcopy-MergeLevels*	CGHclassify	CLAC**	2sd
Overall classification rate	98.6	90.8	95.1	88.0	87.4
True positive rate	95.8	96.2	75.9	46.2	36.3
True negative rate	99.3	89.2	99.7	98.3	99.9

Table 1: *: for DNAcopy-MergeLevels, we used cut-off parameters: Wilcoxon $p = 0.1$, Ansari $p = 0.01$. **: for CLAC, three 'Normal' samples were generated analogously to the Normal state segments in the simulated Tumor samples; we used default setting FDR = 0.01.

2.2 Validation of CGHcall for oligonucleotide data

Oligonucleotide arrays contain smaller elements than BAC arrays, which makes this platform more suitable to detect small amplifications and deletions [13]. Amplifications (or high copy number gains) are biologically particularly important as was, for example, demonstrated for expression of the ERBB2 gene [8]. As opposed to BAC clones the elements of oligonucleotide arrays are synthetic and usually referred to as oligonucleotides or oligos rather than clones. The variation between oligos is larger compared to that between clones on BAC arrays. Nevertheless, CGHcall performs well for oligonucleotide array CGH data too. Amplifications show up as (usually short) sequences of outliers in the data. It was shown that mixture models can be used to model such outliers [1]. CGHcall can automatically classify the segments into four classes instead of the more conventional three, thereby discriminating candidate amplifications from single and double copy gains. To detect an amplification the segmentation algorithm needs to assign a segment to the sequence of clones forming the amplified region and the calling should classify that segment to the last mode in the model.

Figure 1(a) shows the results of CGHcall for the BT474 breast cancer cell line data discussed in detail in [9]. Several candidate amplifications are detected, in particular on chromosome 17, which is magnified in Figure 2. Moreover, despite the rather 'wild' profile, the algorithm detects several large blocks of gains and losses. As a comparison, Figure 1(b) displays the result of CGHclassify [3] on exactly the same data. Due to the discontinuous result the plot is spiky suggesting unrealistically many breakpoints. The figure illustrates that the segmentation used by CGHcall captures the dependency structure better than the short-range structures used in [3] for this type of oligonucleotide data.

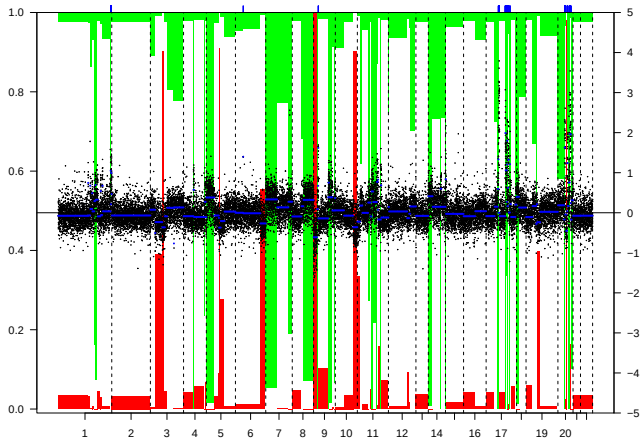
Finally, the capability of CGHcall to automatically detect small deletions is illustrated. The cell line SKBR7 was discussed in [9], in which a small deletion on chromosome 12 could only be visually identified. This deletion was biologically and technically validated using FISH. In Figure 3 it is observed that CGHcall detects this small deletion.

2.3 Summary Plot

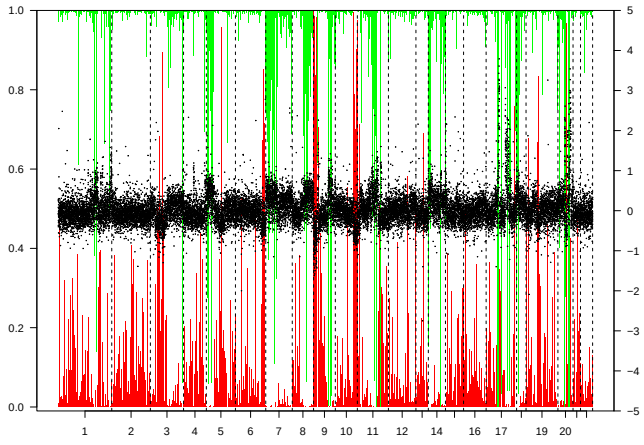
The CGHcall package contains a simple summary plot function, which generates a quick overview to determine recurrently aberrated chromosomal regions. It is slightly more sophisticated than a frequency plot, because it weighs each call with its posterior probability (Figure 4). Comparing it to frequency plot Figure 1b in [7] we observe a confirmative similar pattern. Striking, however, is the more realistic appearance of Figure 4 with respect to Figure 1b in [7]. The latter is much more discontinuous: frequencies fluctuate heavily from clone to clone, which is caused by the clone-wise aberration calling in [7] as opposed to our segment-wise calling.

2.4 Computing Times

Computing times depend mostly on the number of segments in the data and the number of iterations of the EM algorithm (usually between 1 and 6). Hence, computing times for 30-40K oligonucleotide data are not necessarily much longer than those for 3-6K BAC data. As an indication, we mention here a few examples using a PC with 1.73Ghz processor and 1Gb



(a) CGHcall



(b) CGHclassify

Figure 1: Calling results for BT474 cell line on oligo array [9] using CGHcall (a) and CGH-classify (b)

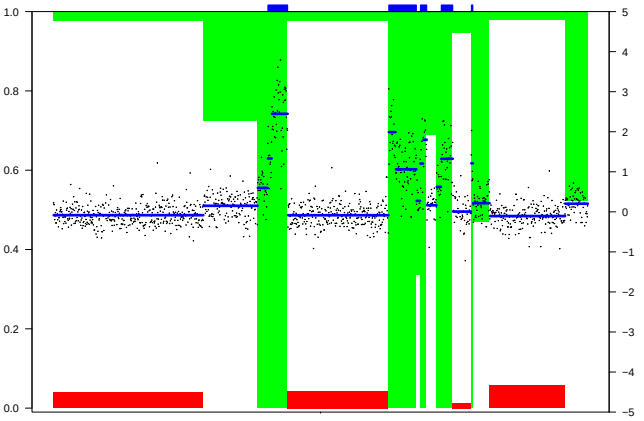


Figure 2: Chromosome 17 calling results for BT474 cell line on oligo array [9] using CGHcall

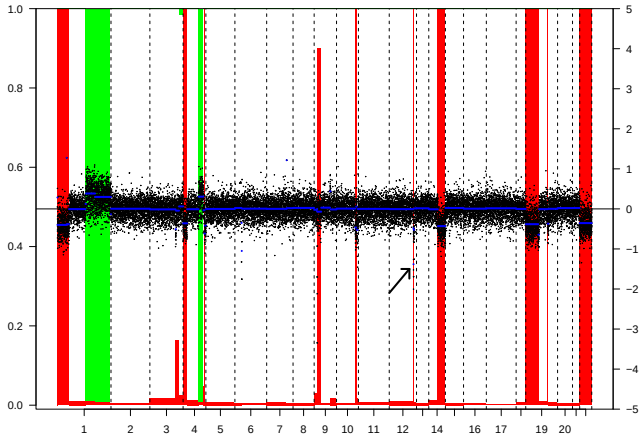


Figure 3: Calling results for SKBR7 cell line on oligo array [9] using CGHcall

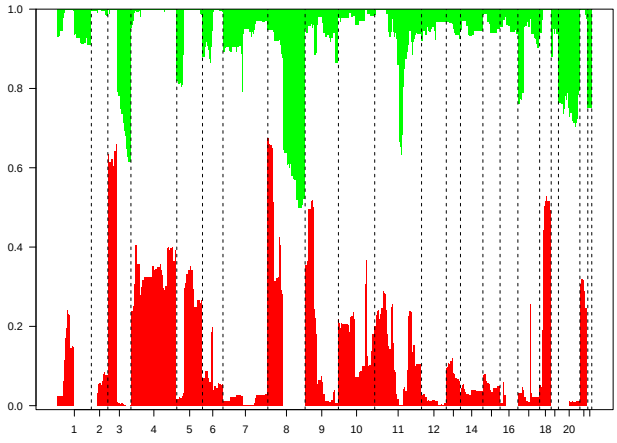


Figure 4: Aberration summary plot for 89 oral squamous cell carcinomas [7]

internal memory. These computing times include segmentation by DNAcopy. CGHcall ran 12045 sec. (∼3 hours; three EM-iterations) for a 96 sample gastric cancer data set with ∼4000 BAC clones and 6578 segments, whereas it ran 6220 sec (∼1h.40 min; two EM-iterations) on a 71 sample gastric cancer data set with ∼3900 clones, 4484 segments. Computations took 1835 sec. (∼30 min.; one EM-iteration) for a 42 sample lymphoma cancer data set with ∼4000 BAC clones and 2566 segments. For a 10 sample colon cancer data set with ∼23000 oligonucleotide clones and 428 segments, it used 917 sec. (∼15 min; two EM iterations).

2.5 Supplementary data and plots

The oral carcinomas data set [7] is available as Supplementary Table B at the following site: http://www.nature.com/onc/journal/v24/n26/suppinfo/1208601s1.html (please save as '.txt' and remove 'Target' and 'Pos' columns). The BT474 and SKBR7 cell line data [9] are available from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm. nih.gov/projects/geo/), via accession numbers GSM73557 and GSM75171, respectively.

CGHcall is implemented as a package for the statistical software environment R [6]. It is available from http://www.few.vu.nl/˜mavdwiel/CGHcall.html. As an example data set, five profiles from the cervical cancer data [12] are available at the same web-site. Links to the other data sets discussed in the paper and CGHcall plots for all data sets are also available from this web-site.

References

[1] M. Aitkin and G.T. Wilson. Mixture models, outliers, and the EM algorithm. *Technometrics*, 22:325–331, 1980.

[2] J. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report ICSI-TR-97-021, University of Berkeley, citeseer.ist.psu.edu/bilmes98gentle.html, 1997.

[3] D.A. Engler, G. Mohapatra, D.N. Louis, and R.A. Betensky. A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, 7:399–421, 2006.

[4] M. Khojasteh, W.L. Lam, R.K. Ward, and C. MacAulay. A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics*, 6:274–274, 2005.

[5] T.L. Naylor, J. Greshock, Y. Wang, T. Colligon, Q.C. Yu, V. Clemmer, T.Z. Zaks, and B.L. Weber. High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res*, 7(6):1186–1198, 2005.

[6] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0, http://www.R-project.org.

[7] A.M. Snijders, B.L. Schmidt, J. Fridlyand, N. Dekker, D. Pinkel, R.C.K. Jordan, and D.G. Albertson. Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*, 24(26):4232–4242, Jun 2005.

[8] E.H. van Beers and P.M. Nederlof. Array-CGH and breast cancer. *Breast Cancer Res*, 8(3):210, 2006.

[9] P. van den IJssel, M. Tijssen, S-F. Chin, P. Eijk, B. Carvalho, E. Hopmans, H. Holstege, D.K. Bangarusamy, J. Jonkers, G.A. Meijer, C. Caldas, and B. Ylstra. Human and mouse oligonucleotide-based array CGH. *Nucleic Acids Res*, 33(22):e192, 2005.

[10] P. Wang, Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani. A method for calling gains and losses in array CGH data. *Biostatistics*, 6:45–58, 2005.

[11] H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21:4084–4091, 2005.

[12] S.M. Wilting, P.J.F. Snijders, G.A. Meijer, B. Ylstra, P.R.L.A. van den IJssel, A.M. Snijders, D.G. Albertson, J. Coffa, J.P. Schouten, M.A. van de Wiel, C.J.L.M. Meijer, and R.D.M. Steenbergen. Increased gene copy numbers at chromosome 20q are frequent in both squamous cell carcinomas and adenocarcinomas of the cervix. *J. Pathol.*, 209:220–230, 2006.

[13] B. Ylstra, P. van den IJssel, B. Carvalho, R.H. Brakenhoff, and G.A. Meijer. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res*, 34(2):445–450, 2006.