Framework for Identifying Common Aberrations in DNA Copy Number Data

Amir Ben-Dor[1,*], Doron Lipson[2], Anya Tsalenko[1], Mark Reimers[3],
Lars O. Baumbusch[4], Michael T. Barrett[1,5], John N. Weinstein[3],
Anne-Lise Borresen-Dale[4], and Zohar Yakhini[1,2]

[1] Agilent Laboratories, Santa-Clara, CA
[2] Computer Science Dept., Technion, Haifa
[3] National Cancer Institute, Bethesda, MD
[4] Department of Genetics, Institute for Cancer Research,
Rikshospitalet-Radiumhospitalet Medical Center
[5] Translational Genomics Research Institute, Phoenix, AZ
amir_ben-dor@agilent.com

Abstract. High-resolution array comparative genomic hybridization (aCGH) provides exon-level mapping of DNA aberrations in cells or tissues. Such aberrations are central to carcinogenesis and, in many cases, central to targeted therapy of the cancers. Some of the aberrations are sporadic, one-of-a-kind changes in particular tumor samples; others occur frequently and reflect common themes in cancer biology that have interpretable, causal ramifications. Hence, the difficult task of identifying and mapping common, overlapping genomic aberrations (including amplifications and deletions) across a sample set is an important one; it can provide insight for the discovery of oncogenes, tumor suppressors, and the mechanisms by which they drive cancer development.

In this paper we present an efficient computational framework for identification and statistical characterization of genomic aberrations that are common to multiple cancer samples in a CGH data set. We present and compare three different algorithmic approaches within the context of that framework. Finally, we apply our methods to two datasets – a collection of 20 breast cancer samples and a panel of 60 diverse human tumor cell lines (the NCI-60). Those analyses identified both known and novel common aberrations containing cancer-related genes. The potential impact of the analytical methods is well demonstrated by new insights into the patterns of deletion of CDKN2A (p16), a tumor suppressor gene crucial for the genesis of many types of cancer.

Keywords: CGH, cancer, microarray data analysis, common aberrations, breast cancer, NCI-60.

1 Introduction

Alterations in DNA copy number are characteristic of many cancer types and drive some cancer pathogenesis processes as well as several developmental disorders.

---

[*] Corresponding author.

These alterations include large chromosomal gains and losses as well as smaller scale amplifications and deletions. Genomic instability can often trigger the over-expression or activation of oncogenes and the silencing of tumor suppressors. Mapping regions of common genomic aberrations can therefore provide insight to cancer pathogenesis and lead to discovery of cancer-related genes and the mechanisms by which they drive the disease. Genomic aberrations are also routinely used for diagnosis and clinical practice. For example, Erbb2 amplification is a strong predictor of Herceptin activity in breast cancer patients [1]. Similarly, amplifications of MDM2 and CDK4 genes on chromosome 12q13-15 are useful in distinguishing well-differentiated liposarcomas from benign adipose tumors [2].

Technologies for measuring alterations in DNA copy number include local fluorescence in situ hybridization-based techniques, Comparative Genomic Hybridization (CGH [3,4,5]) and the advanced method termed array CGH (aCGH). In aCGH differentially labeled tumor and normal DNA are co-hybridized to a microarray of thousands to hundreds of thousands of genomic BAC clones, cDNA or oligonucleotide probes [6,7,8,9,10,11,12]. The use of oligonucleotide aCGH allows the determination of changes in DNA copy number of relatively small chromosomal regions. Using high density arrays allows very high DNA copy number resolution, in terms of genomic distances, down to single Kb and less.

A common first step in analyzing DNA copy number alterations (CNAs) data consists of identifying aberrant (amplified or deleted) regions in each individual sample. Aberration calling is the subject of extensive literature [13,14,15,16,17]. We briefly address this step of the process in Section 2.1.

To realize the full power of multi-sample, high-resolution oligo-aCGH studies, we are interested in efficient computational methods that enable the automatic elucidation of more complex structures. The focus of this paper is the discovery of common genomic aberrations, either in a fixed set of samples or in a significant subset of the samples. To date, little attention has been given in the literature to formal treatments of this task. Two important exceptions are the work of Disking et al [18] and Rouveirol et al [19]. In [18] the authors developed a method called Significance Testing for Aberrant Copy number (STAC) to address the detection of DNA copy number aberrations across multiple aCGH experiments. STAC uses two complementary statistical scores in combination with a heuristic search strategy. The significance of both statistics is assessed, and p-values are assigned to each location in the genome by using a permutation approach. In the work of Rouveirol et al [19] the authors propose a formal framework for the task of detecting commonly aberrant regions in CGH data, and present two algorithms (MAR and CMAR) for this task. The framework requires, however, a segmentation algorithm that categorize each data point as being gained/lost/normal. Therefore, this approach requires setting an arbitrary threshold for the discretization step, and is not sensitive to the actual copy number change. In addition, the methods of Lipson et al [20], based on optimizing a statistically motivated score function for genomic intervals can be adapted to automatic identification of aberrations that are common in subsets of the sample set. Despite the lack of formal approaches to identifying common aberrations

many studies do report common aberrations and their locations. Typically these aberrations are determined by counting and applying human judgment to single sample calls.

In this paper we present an efficient computational framework for identification and statistical characterization of common genomic aberrations. In Section 2 we start with a description of the overall structure of the framework. The first step, aimed at per-sample aberration calling is described in Section 2.1. The rest of Section 2 is devoted to detailed description of three specific approaches for detecting common aberrations. In Section 2.2 we present the commonly used penetrance method, and its weighted version. We introduce a context-corrected version of penetrance in Section 2.3. We conclude the methods section in Section 2.4 with the CoCoA algorithm, that extends the context-corrected statistical approach to multi-probe intervals.

In Section 3 we apply our methods to two DNA copy number cancer datasets, one derived from a collection of 20 breast cancer samples, and the other a set of 60 cell lines. We compare the results of the three approached using the breast cancer dataset, and highlight several interesting significant aberrations that contain cancer related genes.

2   Framework

In this section we describe a framework for identifying and statistically scoring aberrations that are reoccurring in multiple samples. In a nutshell, the process consists of four steps.

1. **Aberration Calling** – Each of the samples' data vector is analyzed independently, and a set of aberrations (amplifications and deletions) is identified.
2. **Listing candidate intervals** - Given the collection of aberration sets called for all samples, we construct a list of genomic intervals that will be evaluated. We refer to these intervals as *candidate intervals*.
3. **Scoring ⟨candidate interval, sample⟩** – In this step, we calculate a statistical significance score for each candidate interval with respect to each sample.
4. **Scoring candidate intervals** – For each candidate interval, we combine the per-sample scores derived in the previous step into a comprehensive score for the candidate interval and estimate its statistical significance. In addition, we also identify for each candidate interval the set of samples that supports it.

At the end of the process, we list the top-scoring candidate intervals together with their support sets.

The framework is modular in nature, in the sense that different algorithms and statistical models and methods can be used in each of the different steps. For example, alternative algorithms can be used to call aberrations in the first step. Similarly, alternative approaches may be employed to define candidate intervals and interval scores.

In the rest of this section we will describe several specific embodiments of the framework. We begin (Section 2.1) by discussing single sample aberration calling, which may be viewed as the input of the common aberration analysis procedure. In Sections 2.2-2.4 we describe three different algorithms based on the framework. For simplicity, we will describe only scores related to common amplifications, although it is clear that symmetric scores apply to common deletions.

2.1   Single Sample Aberration Calling

The starting point of the procedure of identifying statistically significant common aberrations is a set of aberrant segments for each sample. In this paper we assume that, independent of the particular aberration-calling algorithm, the set of aberration calls for a particular sample and a particular chromosome can be represented by a *step-function*. The latter consists of discrete segments parallel to the x-axis, that together span the entire chromosome. Formally, for a sample s, denote the length (in Mb) of the chromosome by ℓ. A step-function $\mathcal{F}_s : [0, \ell] \longrightarrow \mathcal{R}$ contains a segment for each aberration call (with the appropriate boundaries and height). In addition, segments of height zero are used to represent non-aberrant regions of the chromosome. See Figure 1 for an example of a step-function.
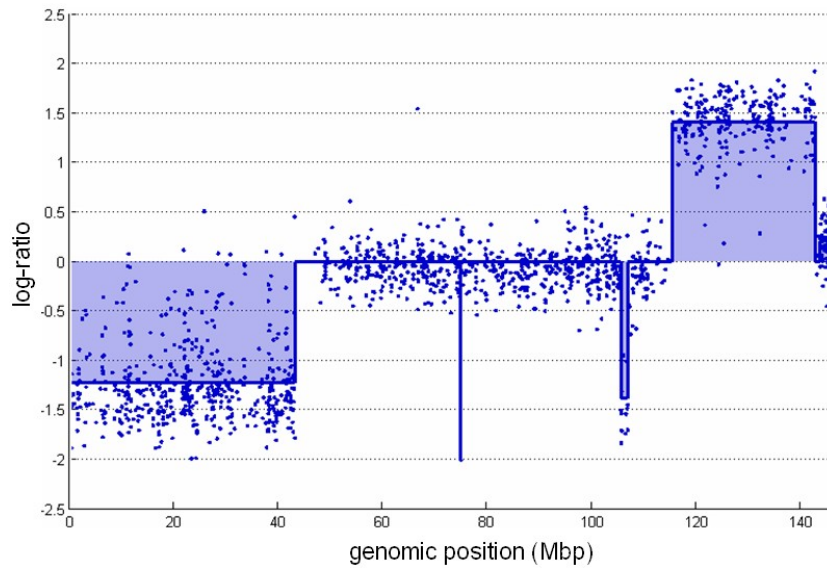

Fig. 1. Step-function derived from chromosome 8 data for colon carcinoma cell line HT29, data from Agilent 44K aCGH array. Solid blue line shows the step-function $\mathcal{F}_s$.

In this study we used the *StepGram* algorithm for single sample aberration calling. StepGram runs in subquadratic time in terms of the number of probes on the chromosome. That translates to < 1 sec for 44K probes, and 3 sec for

185K probes with current the implementation. StepGram is therefore particularly suitable for analysis of large datasets and useful in the context of looking for common aberrations. The details of StepGram were described previously by Lipson et al [20], and an overview is provided here for completeness.

StepGram. Given a vector of real values $V = (v_1, \ldots, v_n)$ (corresponding to normalized log-ratio measurements along a particular chromosome) the optimization problem solved by StepGram involves identifying the interval $I \subseteq [1, n]$ that maximizes the score $|\sum_{i \in I} v_i / \sqrt{I}|$. A branch-and-bound approach allows solving this problem in $O(n^{1.5})$ time complexity in practice. Following identification of the maximal scoring interval the analysis is repeated by recursion to the left, to the right, and within the identified interval until some lower threshold score is attained. A stand-alone implementation of the StepGram algorithm is publicly available at `http://bioinfo.cs.technion.ac.il/stepgram/`.

Other aberration-calling algorithms. Several other algorithms for identifying aberrations in DNA copy number data have been described. These include CBS [15] based on binary segmentation, CLAC [16] based on clustering, aCGH [13] based on HMM, ACE [21] based on FDR, and others. Comparison studies of several of these algorithms were conducted by Lai et al [14] and by Willenbrock et al [17]. Note that many of them are *segmentation* algorithms in the sense that they partition the chromosome into segments of equal copy number but do not attempt to identify which of those segments are aberrant. For the purpose of identifying common aberrations the segmentation output is typically sufficient.

2.2   Weighted and Unweighted Penetrance

We begin by describing the commonly-used penetrance score and its role within the common aberrations analysis framework. Although the penetrance score is not a measure of statistical significance, it does exemplify the different steps of the process.

Candidate intervals. In the case of the penetrance score, the candidate intervals are defined simply as the positions of the probes in the aCGH array. Similar definitions, such as uniformly-spaced pseudo-probes, are also possible. In either case, for a particular chromosome the candidate intervals can be formally defined as a set of non-overlapping intervals $\mathcal{I} = \{[x_i - \epsilon, x_i + \epsilon]\}$. Here ϵ is an arbitrary constant smaller then the minimum distance between any two probes on the array.

Scoring ⟨interval, sample⟩. For a given interval $I = [x_i - \epsilon, x_i + \epsilon]$ and sample s the unweighted amplification penetrance score is defined as a binary score $\alpha(I, s) = 1_{\mathcal{F}_s(x_i) > t}$ for some threshold t. The weighted penetrance scores take into account also the height of the aberration: $\alpha'(I, s) = 1_{\mathcal{F}_s(x_i) > t} \cdot \mathcal{F}_s(x_i)$.

Scoring candidate intervals. The overall penetrance score for a given candidate interval I is defined simply as $\alpha(I) = \sum_s \alpha(I, s)$. As noted before, this score does not reflect any measure of statistical significance.

2.3   Context-Corrected Penetrance

A variant of the penetrance score provides a measure of statistical significance of the common aberration at the specified probe. The significance is defined with respect to the genomic background of each sample, as represented by the pattern of aberrations over each of the samples. In other words, given the specific set of aberration calls for each sample, we wish to describe our "surprise" at seeing a specific set of aberrations co-localized at the same genomic position. Note that the context provided for the score may be either genomic or chromosomal.

Candidate intervals. As was in the case of the penetrance score, the candidate intervals are defined as as a set of non-overlapping intervals at specific genomic positions: $\mathcal{I} = \{[x_i - \epsilon, x_i + \epsilon]\}$.

Scoring ⟨interval, sample⟩. For the context-corrected score, we wish the score of a given interval $I = [x_i - \epsilon, x_i + \epsilon]$ and sample s to reflect the probability of finding an interval of similar (or higher) amplitude given the context of the sample. The score is therefore defined as

$$p(I, s) = \frac{|\{x_j \in \mathcal{I} : \mathcal{F}_s(x_j) \geq \mathcal{F}_s(x_i)\}|}{|\mathcal{I}|}.$$

Scoring candidate intervals. Let S be the set of samples, with $m = |S|$. For a given interval I we now have m scores. Note that the interval I might be aberrant in only a subset of the samples, we therefore seek the subset of samples that will provide maximal significance. Assume, w.l.o.g., that $p(I, 1) \leq p(I, 2) \leq \ldots \leq p(I, m)$. Looking at the first k samples, the probability of concurrently observing k or more scores of probability $p = p(I, k)$ or lower is provided by the Binomial distribution:

$$\rho_k(I) = \text{Binom}(k, m, p) = \sum_{i=k}^{m} \binom{m}{i} p^i (1-p)^{m-i}$$

Since we are interested in identifying aberrations that occur in at least two samples, and to address multiple testing concerns, we define a more conservative score that ignores the first success in the computation,

$$\rho_k'(I) = \text{Binom}(k-1, m-1, p)$$

We define the score of I, to be the minimum of these scores over all values of k, namely,

$$\rho(I) = \min_{k=1,\ldots,m-1} \rho_k'(I).$$

2.4   Context-Corrected Common Aberrations (CoCoA)

Although the context-corrected penetrance algorithm will clearly detect statistically significant common aberrations that are affecting a single probe, its ability

to detect larger significant aberrations is not guaranteed. In some cases, a multi-probe common aberration may be significant as a whole, although the score of each single probe contained in the aberration may not show statistical significance. For example, consider the case in which each of many samples contains many random high-amplitude single-probe amplifications and a common large moderate-amplitude amplification. In that case, the size of the aberration may help us to determine its significance, since not many random aberrations of the same size will be detected in the background.

The third, most sophisticated, algorithm for identifying significant common aberrations expands the concept of a context-corrected significance score to intervals that are larger than a single probe.

Candidate intervals. Consider a particular chromosome, c, and denote by $T = \{[b_1, e_1], \ldots, [b_k, e_k]\}$ the set of all genomic intervals in c that are called as aberrant in any of the samples. The set of candidate intervals in c is defined to be all genomic intervals that starts at the left side of one interval from T and end at the right side of another. That is, $\mathcal{I} = \{[b_i, e_j] : 1 \leq i, j \leq k, and\ b_i \leq e_j\}$. Note that the size of $\mathcal{I}$ is quadratic in k, the number of called aberrations. A smaller list of candidate intervals can be constructed by considering only intervals in T and intersections thereof. that is $\mathcal{I} = T \cup \{t \cap s : t, s \in T\}$. The size of $\mathcal{I}$ is typically $o(k^2)$, and can be constructed in linear time (proof omitted).

Scoring ⟨interval, sample⟩. Applying the same reasoning as for the Context-Corrected Penetrance, we wish the score of a given interval $I = [b, e]$ and sample s to reflect the probability of finding an interval of the same length with a similar (or higher) amplitude given the context of the sample. More specifically, assume we pick a random interval J of the same size as I in the context (that is, in the same chromosome, or in the entire genome). The score is defined as the probability that the average height of J would be as high (or higher) as the height of I,

$$p(I, s) = \mathrm{Pr}_{J:|J|=|I|}\left(h_s(J) \geq h_s(I)\right).$$

where $|I|$ denotes the genomic size I, and $h_s(J)$ denotes the average height of the step-function $\mathcal{F}_s$ over the interval J. We outline now how to computed $p(I, s)$ efficiently (in linear time). Denote by $\mathcal{F}_{s,\ell}(\cdot)$ the ℓ-window moving average of $\mathcal{F}_s$. The score $p(I, s)$, can now be expressed as a function of $\mathcal{F}_{s,\ell}$,

$$p(I, s) = \frac{|x : \mathcal{F}_{s,\ell}(x) \geq h_s(I)|}{c - \ell}.$$

where c denoted the length of the chromosome. Since $\mathcal{F}_s$ is a step-function, its moving average $\mathcal{F}_{s,\ell}$ is a piecewise-linear function. Thus, we can efficiently identify the regions where $\mathcal{F}_{s,\ell}(x) \geq h_s(i)$, and compute $p(I, s)$.

Scoring candidate intervals. After computing context-corrected per-sample scores for I, we combine them into a statistical score for I using the same binomial distribution calculation as detailed in Section 2.3.

3   Results

In this section we demonstrate the application of the above methods to DNA copy number data from two datasets, both measured using Agilent 44K aCGH arrays:

1. A set of 20 primary breast tumor samples were included in this study. These samples are part of a larger patient cohort consisting of 920 breast cancer patients stage I and II referred for surgical treatment and where detection of isolated tumor cells in bone marrow was performed (The Oslo Micrometastases Study) [22]. Tumor material were fresh frozen immediately after surgery and stored at -80C until use. Although the sample set includes several distinct subtypes that had previously been characterized [23,24], due to its relative homogeneity, we expected to encounter common aberrations typical of breast cancer.

2. A diverse set of 60 cancer cell-lines known as the NCI-60 cell line panel [25]. The NCI-60 panel has been used by the Developmental Therapeutics Program (DTP) of the U.S. National Cancer Institute (NCI) to screen $> 100,000$ chemical compounds and natural product extracts for anticancer activity since 1990 [26,27,25]. The NCI-60 panel is comprised of cell lines from diverse human cancers, including leukemias, melanomas, and cancers of renal, ovarian, lung, colon, breast, prostate, and central nervous system origin. The NCI-60 have been profiled more comprehensively at the DNA, RNA, protein, and functional levels than any other set of cells in existence. The resulting information on molecular characteristics and their relationship to patterns of drug activity have proven fruitful for studies of drug mechanisms of action and resistance [28,29,30,31,25]. Because of its diversity, we expected to find mostly aberrations common only to specific tissue of origin, and possibly some that were found more generally in the panel.

We first compared three algorithms – simple unweighted penetrance, context-corrected penetrance, and CoCoA – on the breast tumor dataset. Overall, the three algorithms detected similar patterns, although the specific output contained obvious differences. In Figure 2 we show the output of the three algorithms for chromosome 9 of the breast tumor dataset. The top panel (a) depicts the aberration calls made on that set of samples, using the StepGram algorithm [20][1]. Several common aberrations, detectable by visual inspection, are indicated at the top of the panel by green and red arrows (deletions and amplifications, respectively). The lower three panels (b-d) depict the output of the three algorithms for the chromosome, aligned by genomic position along the x-axis. Output for the simple penetrance method is expressed in fraction of affected samples, whereas the output for the remaining algorithms is expressed in units of $-\log_{10} \rho(I)$. Note that while the output of the two penetrance algorithms (b,c) is simple to plot in genomic coordinates (by probe location), the output

---

[1] The data points were first centered by most common ploidy. StepGram was then applied with a threshold parameter of 5 stds.
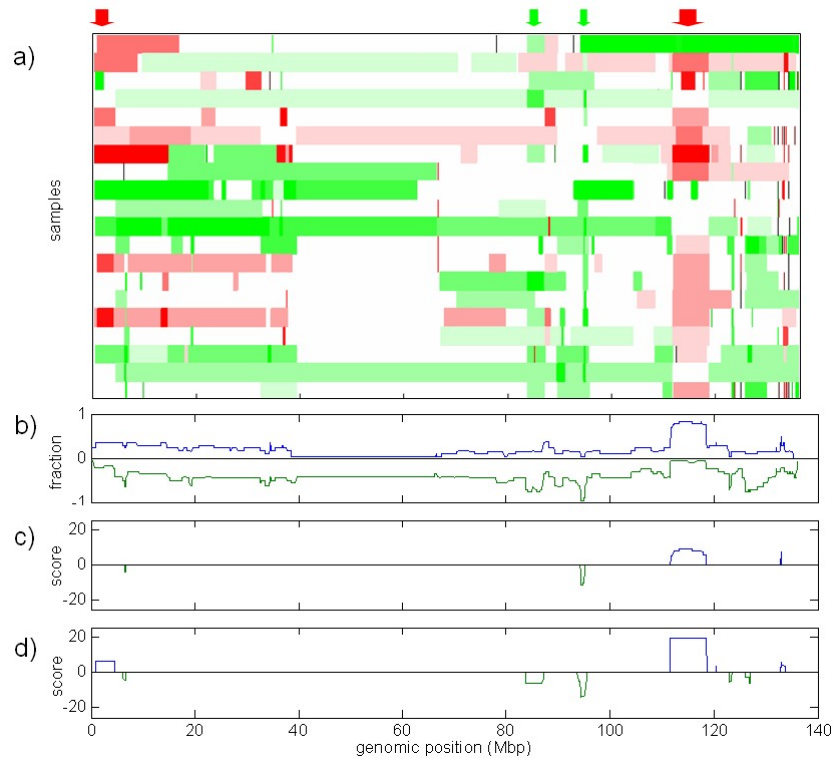
Fig. 2. Common aberrations in a panel of 20 breast tumor samples, chromosome 9: a) Aberration calls in each of the tumor samples (amplifications noted in red, deletion in green). Aberrations were called using StepGram algorithm [20] on centered data, with threshold of 5 stds; b) unweighted penetrance (fraction of samples), c) context-corrected penetrance, d) context-corrected common aberrations (CoCoA), where each probe was scored according to the maximal-scoring interval containing it. Positive values denote amplifications, negative values — deletions. Scores for last two methods are given in $-\log_{10}\rho(I)$ units, only aberrations with score $\rho(I) < 10^{-3}$ and larger than one probe are denoted. Some specific common aberrations in the data are highlighted by arrows at the top of the figure.

of the CoCoA algorithm was transformed into a genomic plot by setting the value of each probe to the score of the maximally-scoring common interval that contains it.

The most prominent common aberrations in the chromosome shown are clearly the large amplification between 110-120Mb and the smaller deletion at 95Mb, both of which were detected by all algorithms. The results of the simple penetrance method, which is a non-statistical method, can be interpreted loosely based on setting of some arbitrary threshold. It is clear that a significant part of the genome can be considered to contain common aberrations if that method is used. The context-corrected penetrance method gives improved output in the

Table 1. Number of common aberrations in the breast cancer data

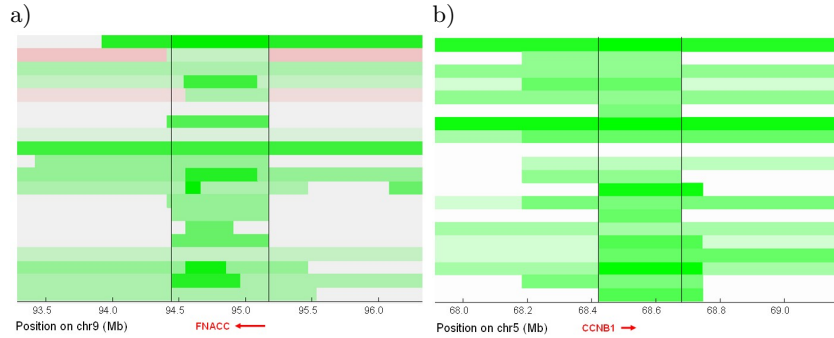|	Amplifications	Deletions	Total
< 200Kb	160	118	278
≥ 200Kb	86	32	118
Total	246	150	396

Fig. 3. Two common focal deletions identified in a panel of 20 breast tumor samples: a) Common deletion in 9q22.32 disrupting FANCC – a gene that encodes a DNA repair protein (11/20 samples, $\rho(I) = 10^{-21}$), b) Common deletion in 5q13.2 disrupting a cyclin gene CCNB1 (8/20 samples, $\rho(I) = 10^{-11.8}$)

sense that only very specific parts of the chromosome are deemed to contain common aberrations, based on a very modest threshold $\rho(I) < 10^{-3}$. Clearly, from the biological point of view, specific output of this type, a result of the correction for the chromosomal context, is highly preferable.

The superiority of the common aberrations method (CoCoA) lies in the higher significance that it gives common aberrations that are longer than one probe. This feature allows higher sensitivity for lower-amplitude common aberrations without loss of specificity. An example of the increased sensitivity is the common amplification detected between 1-5Mb. That aberration is not clearly visible in the outputs of the two methods based on single probe.

Overall, CoCoA identified 396 disjoint common aberrations with score $\rho(I) < 10^{-3}$ in the breast tumor dataset (see Table 1). The range of sizes of the common aberrations identified on the basis of more than a single probe is 1.7Kb - 60Mb. The aberrations are supported by 3-17 samples each. Two specific common focal deletions that were identified in the data set are depicted in Figure 3. The two

Table 2. Number of common aberrations in the NCI-60 data

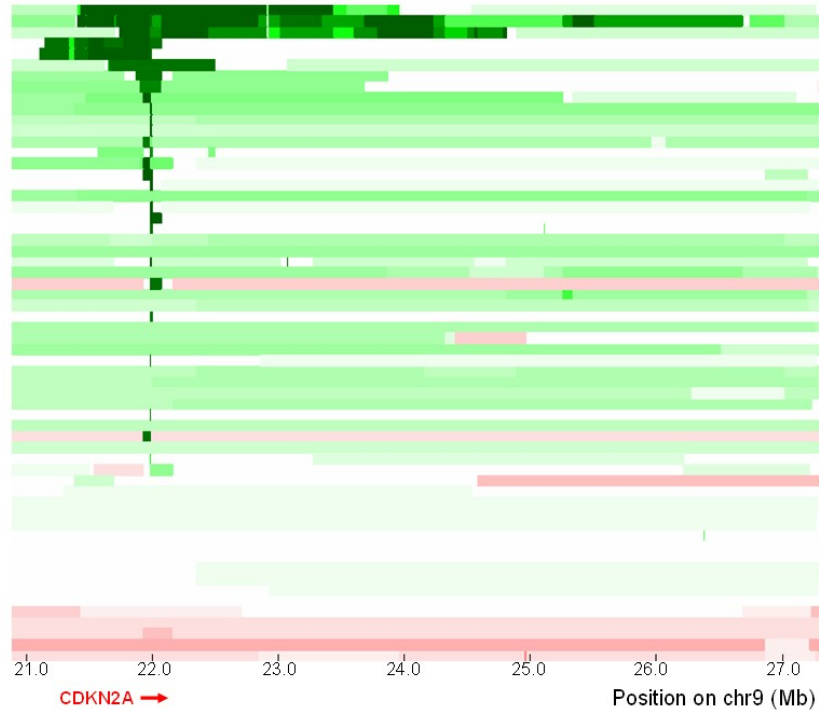|	Amplifications	Deletions	Total
< 200Kb	216	145	361
≥ 200Kb	60	50	110
Total	276	195	471

Fig. 4. A common deletion in 9p that reoccurs in a large fraction (20/60) of the cell-lines of the NCI-60 panel. Common aberration analysis points to the focus of the deletion as being the known tumor suppressor gene CDNK2A (p16), with $\rho(I) = 10^{-54}$.

deletions, identified in 5q13.2 and 9q22.32, appear to be disrupting two genes with direct involvement in tumor development – CCNB1 (a cyclin gene) and FANCC (a gene encoding a DNA repair protein), respectively. Slightly larger intervals are also aberrant in many samples. The highlighted intervals, however, have the strongest statistical significance.

In the NCI-60 cell line panel CoCoA identified 471 common aberrations (see Table 2). The range of sizes of the common aberrations identified on the basis of more than a single probe is 0.5kb - 100Mb, and aberrations are supported by 3-38 samples each.

One striking common aberration detected in the NCI-60 dataset was a deletion of CDKN2A (p16), a well-characterized tumor-suppressor gene (Figure 4). Clearly the deletion of this gene is a common feature of many of the cell-lines (20/60 of the samples), crossing the boundaries of cell-line subtype. Note also that even though some samples have deletions over larger regions, they all overlap at the genomic location of the p16 gene itself. This observation indicates that a selective pressure to delete p16 was part of the development of all 20 cell line populations and represents a very common feature of cancer development.

4 Discussion

In this paper we propose a computational framework for identifying and analyzing copy number aberrations (amplifications and deletions) that occur across multiple samples and for assessing their statistical significance. The framework allows using different aberration calling algorithms as input, independent of their statistical modeling.

Two central features of our methods are: A)When assessing the significance of a particular aberration, we use the height of the aberration, as opposed to requiring an additional threshold to discretize the aberration calls. B) The ability to address the context of the aberration structures in the individual samples. Given a candidate interval, its significance at a particular sample depends not only on the average height of the candidate interval, but also on the overall prevalence of aberrations in that sample. We describe two methods that have those important features. The CoCoA method scores intervals while the context-corrected penetrance method scores individual loci. In theory, there is a larger statistical power in considering multi-loci aberrations as both a supporting sample set and a genomic interval are identified together. Another difference between probe level and interval level analysis, is that in probe level analysis an additional thresholding step is required to determine the boundaries of the common aberrations. Note that for any single locus penetrance based method intervals with consistent high scoring can theoretically arise from aberrations in different sets of samples. In practice this is usually not the case. When scoring intervals, as CoCoA does, sample integrity is always preserved: the set of samples over which an interval is reported as a common aberration is the same for all loci spanned by said interval.

Our framework is very efficient. When run on the NCI60 sample set our process takes under 1 minute, including the first step of single sample aberration calling, using StepGram. This enables interactive data analysis that is not possible for less efficient approaches. This combined approach will scale up to larger datasets and to denser arrays that allow for much finer mapping of aberrant regions. We emphasize that this requires not only an efficient approach to common aberrations but also a very efficient aberration-calling methods.

One important previous formal treatment of calling common aberrations in CGH data is described in [18]. The method described therein, called STAC, is based on a heuristic search seeking to optimize statistical scores assigned to candidate regions of common aberrations. STAC's search is computationally intensive and performance is further limited by relying on permutations and simulations to obtain significance estimates. According to the paper's Supplementary material STAC implementation takes days to run on relatively small datasets of 42 and 47 samples, measured using a low resolution (approximately 1Mb) technology. STAC treats gains and losses as binary and does not takes into account the exact amplitude of the measured signal.

We have shown examples of applying the framework on a set of breast cancer samples that identify both known and novel cancer related genes. It is interesting to note p16 as a universal deletion in the NCI60 panel. FANCC, a gene from the Fanconi anemia group of genes (FA), which codes to a DNA repair protein is

deleted in 11 out of the 20 breast cancer samples. FA genes are known to be co-factors interacting with BRCA2 in breast cancer pathogenesis. In a recent study [32] the authors demonstrate a role for the FA pathway in interstrand cross-link repair which is independent from that of BRCA2 in the same process. This finding and our implication of FANCC as a fairly focal common breast cancer deletion together suggest an important role for FANCC under-functioning in cancer pathogenesis.

Lastly, we note that the methods herein presented can be extended to identify differential aberrations in DNA copy number data coming from several pheno-typic classes. A more detailed investigation of this application will be the topic of future work.

References

1. Kauraniemi, P., Hautaniemi, S., Autio, R., Astola, J., Monni, O., Elkahloun, A., Kallioniemi, A.: Effects of Herceptin treatment on global gene expression patterns in HER2-amplified and nonamplified breast cancer cell lines. Oncogene **23**(4) (2004) 1010–1013
2. Binh, M., Sastre-Garau, X., Guillou, L., de Pinieux, G., Terrier, P., Lagace, R., Aurias, A., Hostein, I., Coindre, J.: MDM2 and CDK4 immunostainings are useful adjuncts in diagnosing well-differentiated and dedifferentiated liposarcoma subtypes: A comparative analysis of 559 soft tissue neoplasms with genetic data. American Journal of Surgical Pathology **29**(10) (2005) 1340–1347
3. Balsara, B., Testa, J.: Chromosomal imbalances in human lung cancer. Oncogene **21**(45) (2002) 6877–83
4. Kallioniemi, O., Kallioniemi, A., Sudar, D., Rutovitz, D., Gray, J., Waldman, F., Pinkel, D.: Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors. Semin Cancer Biol **4**(1) (1993) 41–46
5. Mertens, F., Johansson, B., Hoglund, M., Mitelman, F.: Chromosomal imbalance maps of malignant solid tumors: a cytogenetic survey of 3185 neoplasms. Cancer Research **57**(13) (1997) 2765–80
6. Barrett, M., Scheffer, A., Ben-Dor, A., Sampas, N., Lipson, D., Kincaid, R., Tsang, P., Curry, B., Baird, K., Meltzer, P., Yakhini, Z., Bruhn, L., Laderman, S.: Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. PNAS **101**(51) (2004) 17765–17770
7. Bignell, G., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K., Wei, W., Stratton, M., Futreal, P., Weber, B., Shapero, M., Wooster, R.: High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Research **14**(2) (2004) 287–95
8. Brennan, C., Zhang, Y., Leo, C., Feng, B., Cauwels, C., Aguirre, A., Kim, M., Protopopov, A., Chin, L.: High-resolution global profiling of genomic alterations with long oligonucleotide microarray. Cancer Research **64**(14) (2004) 4744–8
9. Hedenfalk, I., Ringner, M., Ben-Dor, A., Yakhini, Z., Chen, Y., Chebil, G., Ach, R., Loman, N., Olsson, H., Meltzer, P., Borg, A., Trent, J.: Molecular classification of familial non-BRCA1/BRCA2 breast cancer. PNAS **100**(5) (2003) 2532–7
10. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W., Chen, C., Zhai, Y., Dairkee, S., Ljung, B., Gray, J., Albertson, D.: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics **20**(2) (1998) 207–211

11. Pollack, J., Perou, C., Alizadeh, A., Eisen, M., Pergamenschikov, A., Williams, C., Jeffrey, S., Botstein, D., Brown, P.: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nature Genetics **23**(1) (1999) 41–6

12. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T., Trask, B., Patterson, N., Zetterberg, A., Wigler, M.: Large-scale copy number polymorphism in the human genome. Science **305**(5683) (2004) 525–8

13. Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D., Jain, A.: Hidden markov models approach to the analysis of array cgh data. Journal of Multivariate Analysis **90** (2004) 132–153

14. Lai, W., Johnson, M., Kucherlapati, R., Park, P.: Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics **21**(19) (2005) 3763–70

15. Olshen, A., Venkatraman, E., Lucito, R., Wigler, M.: Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics **5** (2004) 557–72

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

17. Willenbrock, H., Fridlyand, J.: A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics **21**(22) (2005) 4084–91

18. Diskin, S., Eck, T., Greshock, J., Mosse, Y., Naylor, T., Stoeckert, C., Weber, B., Maris, J., Grant, G.: STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. Genome Research **16** (2006) 1149–1158

19. Rouveirol, C., Stransky, N., Hupe, P., Rosa, P.L., Viara, E., Barillot, E., Radvanyi, F.: Computation of reccurant minimla genomic alterations from array-cgh data. Bioinformatics (2006) 849–856

20. Lipson, D., Aumann, Y., Ben-Dor, A., Linial, N., Yakhini, Z.: Efficient calculation of interval scores for DNA copy number data analysis. Journal of Computational Biology **13**(2) (2006) 215–28

21. Lingjarde, O.C., Baumbusch, L.O., Liestol, K., Glad, I.K., Borresen-Dale, A.L.: Cgh-explorer: a program for analysis of array-cgh data. Bioinformatics **21(6)** (2005) 821–822

22. Wiedswang, G., Borgen, E., Kvalheim, R.K.G., Nesland, J., Qvist, H., Schlichting, E., Sauer, T., Janbu, J., Harbitz, T., Naume, B.: Detection of isolated tumor cells in bone marrow is an independent prognostic factor in breast cancer. Journal of Clinical Oncology **21** (2003) 3469–3478

23. Sorlie, T., Perou, C., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Thorsen, T., Quist, H., Matese, J., Brown, P., Botstein, D., Lonning, P.E., Borresen-Dale, A.: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS **98**(10) (2001) 10869–74

24. Sorlie, T., Wang, Y., Xiao, C., Johnsen, H., Naume, B., Samaha, R., Borresen-Dale, A.L.: Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: Gene expression analyses across three different platforms. BMC Genomics **7** (2006) 127

25. Weinstein, J., Myers, T., O'Connor, P., Friend, S., Jr., A.F., Kohn, K., Fojo, T., Bates, S., Rubinstein, L., Anderson, N., Buolamwini, J., van Osdol, W., Monks, A., Scudiero, D., Sausville, E., Zaharevitz, D., Bunow, B., Viswanadhan, V., Johnson, G., Wittes, R., Paull, K.: An information-intensive approach to the molecular pharmacology of cancer. Science **275**(10) (1997) 343–49

26. Monks, A., Scudiero, D., Skehan, P., Shoemaker, R., Paull, K., Vistica, D., Hose, C., Langley, J., Cronise, P., et al., A.V.W.: Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. Journal of the National Cancer Institute **83** (1991) 757–66

27. Shoemaker, R., Monks, A., Alley, M., Scudiero, D., Fine, D., McLemore, T., Abbott, B., Paull, K., Mayo, J., Boyd, M.: Development of human tumor cell line panels for use in disease-oriented drug screening. Progress in Clinical and Biological Research **276** (1988) 265–86

28. Nishizuka, S., Charboneau, L., Young, L., Major, S., Reinhold, W., Waltham, M., Kouros-Mehr, H., Bussey, K., Lee, J., Espina, V., Munson, P., 3rd, E.P., Liotta, L., Weinstein, J.: Proteomic profiling of the nci-60 cancer cell lines using new high-density reverse-phase lysate microarrays. PNAS **100** (2003) 14229–34

29. Paull, K., Shoemaker, R., Hodes, L., Monks, A., Scudiero, D., Rubinstein, L., Plowman, J., Boyd, M.: Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and compare algorithm. Journal of the National Cancer Institute **81** (1989) 1088–92

30. Shi, L., Fan, Y., Lee, J., Waltham, M., Andrews, D.T., Scherf, U., Paull, K., Weinstein, J.: Mining and visualizing large anticancer drug discovery databases. Journal of Chemical Information and Computer Sciences **40** (2000) 367–79

31. Staunton, J., Slonim, D., Coller, H., Tamayo, P., Angelo, M., Park, J., Scherf, U., Lee, J., Reinhold, W., Weinstein, J., Mesirov, J., Lander, E., Golub, T.: Chemosensitivity prediction by transcriptional profiling. PNAS **98** (2001) 10787–92

32. Kitao, H., Yamamoto, K., Matsushita, N., Ohzeki, M., Ishiai, M., Takata, M.: Functional interplay between brca2/fancd1 and fancc in dna repair. Journal of Biological Chemistry **281(30)** (2006) 21312–21320