# Chapter 4

# Analyzing Cancer Samples with SNP Arrays

**Peter Van Loo, Gro Nilsen, Silje H. Nordgard, Hans Kristian Moen Vollan, Anne-Lise Børresen-Dale, Vessela N. Kristensen, and Ole Christian Lingjærde**

## Abstract

Single nucleotide polymorphism (SNP) arrays are powerful tools to delineate genomic aberrations in cancer genomes. However, the analysis of these SNP array data of cancer samples is complicated by three phenomena: (a) aneuploidy: due to massive aberrations, the total DNA content of a cancer cell can differ significantly from its normal two copies; (b) nonaberrant cell admixture: samples from solid tumors do not exclusively contain aberrant tumor cells, but always contain some portion of nonaberrant cells; (c) intratumor heterogeneity: different cells in the tumor sample may have different aberrations. We describe here how these phenomena impact the SNP array profile, and how these can be accounted for in the analysis. In an extended practical example, we apply our recently developed and further improved ASCAT (allele-specific copy number analysis of tumors) suite of tools to analyze SNP array data using data from a series of breast carcinomas as an example. We first describe the structure of the data, how it can be plotted and interpreted, and how it can be segmented. The core ASCAT algorithm next determines the fraction of nonaberrant cells and the tumor ploidy (the average number of DNA copies), and calculates an ASCAT profile. We describe how these ASCAT profiles visualize both copy number aberrations as well as copy-number-neutral events. Finally, we touch upon regions showing intratumor heterogeneity, and how they can be detected in ASCAT profiles. All source code and data described here can be found at our ASCAT Web site (http://www.ifi.uio.no/forskning/grupper/bioinf/Projects/ASCAT/).

**Key words:** Cancer, Tumor, SNP arrays, ASCAT, Allelic bias, Aneuploidy, Intratumor heterogeneity

## 1. Introduction

Single nucleotide polymorphism (SNP)-based DNA microarrays represent a powerful technology, allowing simultaneous measurement of the allele-specific copy number at many different single nucleotide polymorphic loci in the genome. A SNP is a single base

Junbai Wang et al. (eds.), *Next Generation Microarray Bioinformatics: Methods and Protocols*,
Methods in Molecular Biology, vol. 802, DOI 10.1007/978-1-61779-400-1_4, 

locus in the genome that occurs in the population in two different variants, for example, some individuals can have a cytosine base (C) at that locus, while other individuals have a guanine base (G). Calling one of the allelic variants as A and the other as B, the fact that our DNA contains one paternal and one maternal copy means we may obtain genotypes AA (homozygous A), AB (heterozygous), or BB (homozygous B) for any given SNP locus. By measuring thousands or even millions of such SNP loci, a considerable part of the genome that is variable in the population can effectively be arrayed. At present, SNP array platforms are available from Affymetrix (1) and Illumina (2). Current Affymetrix SNP array technology is based on hybridization to oligonucleotides, arrayed in a regular and predefined pattern on glass slides, while Illumina technology is based on in situ single nucleotide extension reactions on bead arrays. However, despite these substantial technological differences, the resulting data show that similar properties and techniques developed on one technology are in general applicable to the other technology, after an appropriate data transformation.

Cancer genomes often show numerous DNA sequence changes, ranging in size from single nucleotide mutations to gains, amplifications, insertions or deletions of large chromosomal fragments, and even whole-genome duplications (3, 4). For this reason, genotypes in cancer are no longer limited to AA, AB, or BB, but can also be, e.g., A, BBB, AAB, or ABBB. The SNP array data contain in principle all the necessary information to deduce these more complex genotypes, but three phenomena can complicate the analysis in practice:

*Aneuploidy*: Owing to a multitude of aberrations, the total amount of DNA in a tumor cell can differ significantly from the normal state of two copies of each chromosome. This is called aneuploidy (compared to the normal state of diploidy). Aneuploidy makes it difficult to determine the normal reference state, as the average signal strength does not necessarily correspond to two copies, as in noncancer genomes. Hence, aneuploidy should be explicitly accounted for in the data analysis.

*Nonaberrant cell admixture*: A cancer biopsy always contains some nonaberrant cells. These nonaberrant cells can be nontumoral cells in the tumor microenvironment (e.g., fibroblasts, endothelial cells, infiltrating immune cells) (5), normal cells in nontumoral regions of the biopsy, or possibly a subpopulation of tumor cells with no visible aberrations. The measured signal will therefore reflect a combination of aberrant and nonaberrant cells and will be more similar to the signal of a normal sample than would have been the case for a homogeneous sample of tumor cells. The amount of nonaberrant cell admixture may differ significantly between cancer samples (from less than 10% to more than 80%), necessitating separate calculation of the fraction of nonaberrant cells for each assayed sample.

*Intratumor heterogeneity*: Different cells in a cancer biopsy may harbor different aberrations. In a recent study (6), multiple separable populations of breast cancer cells were found in more than half of the breast carcinomas, but the major cancer cell populations within any given tumor were limited to one, two, or three different subclones. These typically shared many aberrations, indicating that they had a common ancestor. As a result of this intratumor heterogeneity, for some loci, unambiguous genotypes cannot be obtained, even when accounting for nonaberrant cell admixture and aneuploidy.

Numerous data analysis tools for SNP array data exist, including many tools specifically aimed at analyzing cancer samples. Examples of automated SNP array data analysis methods that account for nonaberrant cell admixture in tumor samples are genoCNA (7) and BAFsegmentation (8). Two tools that take tumor aneuploidy into account are OverUnder (9) and PICNIC (10). Methods that automatically account for both tumor aneuploidy and nonaberrant cell admixture are GAP (genome alteration print) (11) and ASCAT (allele-specific copy number analysis of tumors) (12). These methods match the data from one sample to discrete allele-specific copy number states, thus determining tumor ploidy and aberrant tumor cell fraction, as well as copy numbers and genotypes across the genome. GAP uses pattern recognition on copy number and allelic imbalance profiles, while ASCAT directly models allele-specific copy number as a function of the SNP data, the tumor ploidy, and the aberrant cell fraction, and subsequently selects the solution that is closest to nonnegative integer copies at all assayed loci in the genome. Finally, regions subject to intratumor heterogeneity can be predicted from the output of both methods as outlier regions after the optimal genome-wide fit has been obtained.

Here, we focus on the analysis of SNP array data of cancer samples using ASCAT. We first introduce the structure of SNP array data, and explain how nonaberrant cell admixture and tumor aneuploidy influence the signal. Next, a breast cancer example dataset is analyzed using ASCAT. The data is subsequently visualized, filtered for germline heterozygous loci, and segmented. Finally, the actual ASCAT algorithm is applied and the output is discussed.

## 2. Materials

All source code and data described here can be found at our ASCAT Web site (13) (see Note 1). *R* is required for application of the ASCAT algorithm. ASCAT version 2.0 is used.

## 3. Methods

### 3.1. SNP Array Data of Cancer Samples

SNP array data consist of two data tracks (Fig. 1a): the total signal intensity and the allelic contrast. The total signal intensity is represented by Log *R* and shows the total copy number on a

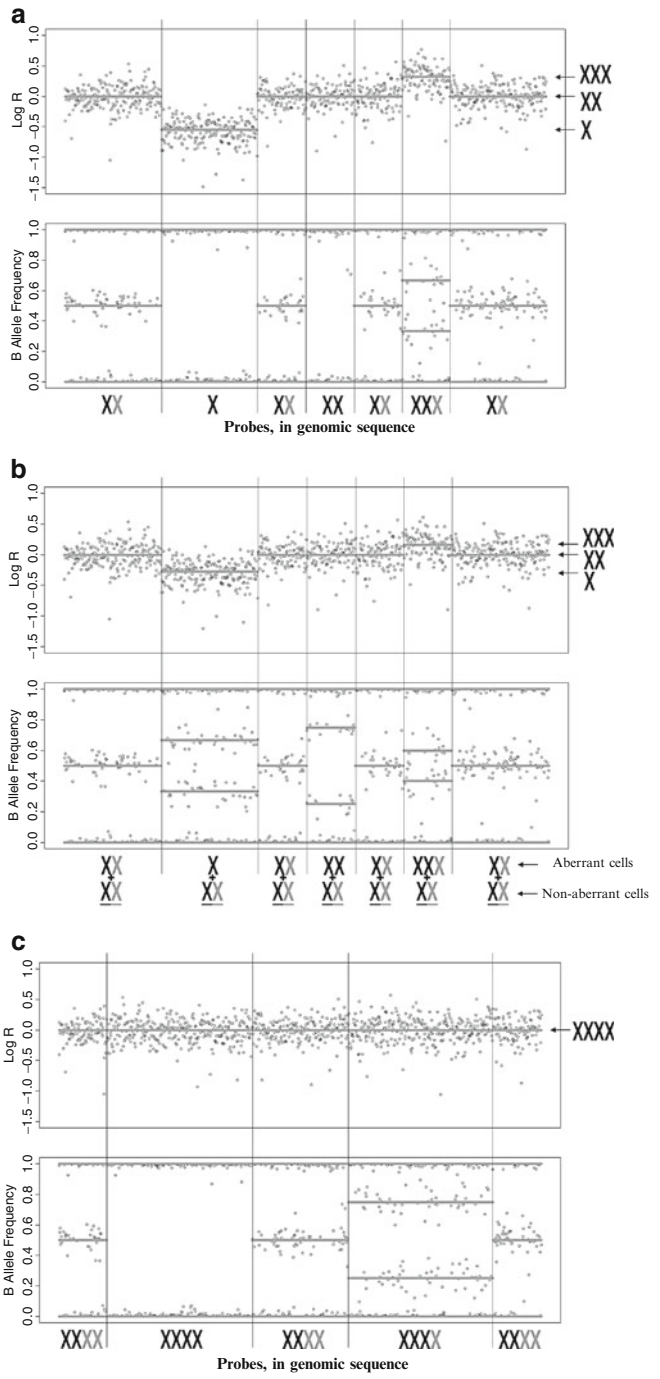


Fig. 1. The structure of SNP array data. (**a**) Log *R* (*top*) and BAF data (*bottom*). The Log *R* data track shows the copy number, with the lines close to 0 corresponding to normal

logarithmic scale. The allelic contrast is represented by the B allele frequency (BAF) and shows the relative presence of each of the two alternative nucleotides at each SNP locus profiled (see Note 2). In a diploid sample, a locus with two identical copies will appear with a Log *R* value close to 0, and a BAF value either close to 0 (genotype AA) or close to 1 (genotype BB). A heterozygous locus (genotype AB) will appear as a BAF close to 0.5. From these SNP array data, different genomic aberrations (gains, losses, copy-number-neutral events) can be delineated, as exemplified in Fig. 1a.

Most cancers show evidence of nonaberrant cell admixture (Fig. 1b). This is most evident in the BAF track, where it can be most clearly illustrated in regions with deletions. In case of a

Fig. 1. (Continued) (copy number 2), the decrease to −0.55 corresponding to a deletion (copy number 1) and the increase to 0.4 to a duplication (copy number 3). Both the raw data and the data after application of a segmentation algorithm are shown. The BAF data track shows three bands for normal regions (genotypes AA, AB, and BB with BAF of 0, 0.5, and 1, respectively). In these regions, 1 copy from each parent is inherited (shown at the *bottom*). In the deleted region, only A and B genotypes occur (BAF of 0 and 1, respectively), and in the duplicated region, the four bands correspond to AAA (BAF = 0), AAB (BAF = 0.33), ABB (BAF = 0.67), and BBB (BAF = 1) genotypes. Finally, the middle region shows copy-number-neutral loss-of-heterozygosity (LOH): only AA and BB genotypes are found and hence both copies of this region originate from the same parent (also called uniparental disomy). (**b**) Toy example of SNP array data of a cancer sample showing 50% nonaberrant cell admixture (compare to (**a**), which shows the same example without nonaberrant cell admixture). Notice the lower range of the Log *R* track and the particular differences in the BAF track. In the region deleted in the tumor cells, two extra bands are observed, corresponding to mixture of A genotypes in the tumor cells, admixed with nonaberrant cells with an AB genotype (BAF = 0.33) and B genotypes in the tumor cells, admixed with nonaberrant cells with AB genotype (BAF = 0.67). Similarly, the region showing copy-number-neutral LOH also shows two extra bands (AA mixed with AB at BAF = 0.25 and BB mixed with AB at BAF = 0.75). Finally, in the duplicated region, the bands are shifted compared to the homogeneous case shown in (**a**). (**c**) Toy example of SNP array data of an aneuploid sample. Based on the Log *R* track, the entire stretch of DNA shown has an identical copy number. However, the BAF track shows clear differences in allelic contrast. Three regions show an allelic balance (two homozygous bands at BAF = 0 and BAF = 1, and one heterozygous band at BAF = 0.5), one region shows complete LOH (only the two homozygous bands at BAF = 0 and BAF = 1 are present), and one region shows partial LOH (two "homozygous" bands at BAF = 0 and BAF = 1, and two partially heterozygous bands at BAF = 0.25 and BAF = 0.75). These data cannot be explained under a hypothesis of copy numbers 1, 2, or 3 and hence, this entire region is most likely copy number 4. The regions showing allelic balance have two copies from each parent, the region showing complete LOH has four identical copies, and the region showing partial LOH has three copies from one parent and one copy from the other parent. The two partially heterozygous bands correspond to AAAB (BAF = 0.25) and ABBB (BAF = 0.75) genotypes.

hemizygous deletion (one of the copies is lost) in a homogeneous (and diploid) sample, only two bands are expected in the BAF track: one at 0, corresponding to A genotypes, and one at 1, corresponding to B genotypes. In tumor samples, two extra bands are observed (Fig. 1b), corresponding to an AB genotype in the host, where A (top line) or B (bottom line) has been lost in the tumor. This results in a mixture of tumor cells with B genotypes and admixed nonaberrant cells with AB genotypes (top line) and a mixture of tumor cells with A genotypes and admixed nonaberrant cells with AB genotypes (bottom line). The closer both lines are, the higher the relative signal of nonaberrant cells. In the Log *R* track, nonaberrant cell admixture is visible as an "inflation" of the signals: while in a homogeneous sample, Log *R* drops considerably in case of a hemizygous deletion (to −0.55 in case of Illumina SNP arrays (2)), this drop is lower when nonaberrant cell admixture is observed (Fig. 1b and Table 1). Also for other aberrations, an influence of nonaberrant cell admixture can be seen. For example, for duplications, Log *R* is lower and BAF for "ABB" and "AAB" genotypes is closer together than for homogeneous samples. In addition, many cancers show aneuploidy, resulting in a shift of the Log *R* track compared to diploid samples, while the BAF track is not affected (Fig. 1c, Table 1). In the next sections, we will apply our ASCAT suite of tools (12) (version 2.0, see Note 3) to an example series of breast carcinomas. The added value of using a tool like ASCAT for the analysis of cancer SNP array data is illustrated in Fig. 2. ASCAT calculates the tumor ploidy and the aberrant cell fraction, and subsequently outputs an ASCAT profile, containing the allele-specific copy numbers across the genome, calculated specifically for the aberrant tumor cells and correcting for both aneuploidy and nonaberrant cell infiltration (Fig. 2).

### 3.2. Data Loading and Visualization

The example SNP array data consists of four files, containing Log *R* and BAF data derived from tumor samples and matched germline samples. Each is a tab-separated file, containing one data column for each sample, a header containing sample names and three columns describing the SNP loci [containing an identifier (in this case, the RS identifier of the SNP) and the genomic location (chromosome and base pair position on the chromosome)]. BAF data has by definition a range between 0 and 1, while Log *R* can in theory range between $-\infty$ and $+\infty$ (although the large majority of the values will be between −1 and 1). Both data tracks may contain NA values (see also Note 4).

First, the ASCAT libraries must be loaded (in *R*):

```
source(ascat.R)
```

**Table 1**
**Influence of infiltration of nonaberrant cells and of aneuploidy of the aberrant tumor cells on Log *R* and BAF data from Illumina SNP arrays**

| | | | Genotype tumor (BAF) | | |
|---|---|---|---|---|---|
| | | Log *R* | host: AA | host: AB | host: BB |
| No infiltration of nonaberrant cells, aberrant cells diploid | Normal, 2 copies | 0 | AA (0) | AB (0.5) | BB (1) |
| | Deletion, 1 copy | −0.55 | A (0) | A (0)<br>B (1) | B (1) |
| | Duplication, 3 copies | 0.4 | AAA (0) | AAB (0.33)<br>ABB (0.67) | BBB (1) |
| Infiltration of nonaberrant cells | Normal, 2 copies | 0 | AA (0) | AB (0.5) | BB (1) |
| | Deletion, 1 copy | >−0.55 | A (0) | A ($0 < x < 0.5$)<br>B ($0.5 < x < 1$) | B (1) |
| | Duplication, 3 copies | <0.4 | AAA (0) | AAB ($0.33 < x < 0.5$)<br>ABB ($0.5 < x < 0.67$) | BBB (1) |
| Aberrant cells aneuploid (>2 copies per cell) | Normal, 2 copies | <0 | AA (0) | AB (0.5) | BB (1) |
| | Deletion, 1 copy | <−0.55 | A (0) | A (0)<br>B (1) | B (1) |
| | Duplication, 3 copies | <0.4 | AAA (0) | AAB (0.33)<br>ABB (0.67) | BBB (1) |
| Infiltration of nonaberrant cells and aberrant cells aneuploid (>2 copies per cell) | Normal, 2 copies | <0 | AA (0) | AB (0.5) | BB (1) |
| | Deletion, 1 copy | <0 | A (0) | A ($0 < x < 0.5$)<br>B ($0.5 < x < 1$) | B (1) |
| | Duplication, 3 copies | <0.4 | AAA (0) | AAB ($0.33 < x < 0.5$)<br>ABB ($0.5 < x < 0.67$) | BBB (1) |

Typical values of Log *R* and BAF, as well as genotypes, are shown under different scenarios, each time for regions with normal copy number (two copies), deleted regions (one copy), and duplicated regions (three copies)

Next, the data described above can be loaded into ASCAT:

```
ascat.bc = ascat.loadData(Tumor_LogR.
   txt,Tumor_BAF.txt,
Germline_LogR.txt,Germline_BAF.txt)
```

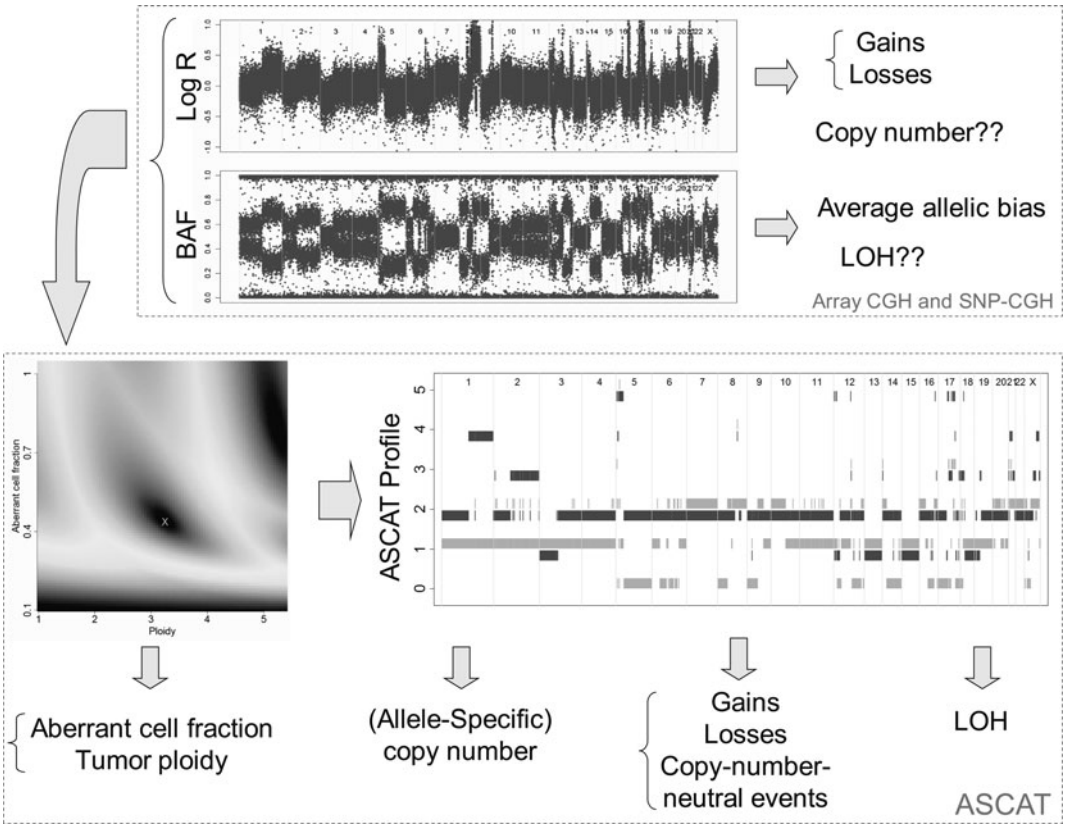


Fig. 2. The principle of data analysis using ASCAT. The result of an array-CGH experiment is a genome-wide measure of total copy number. This allows derivation of gains and losses, but in cancer samples, copy-number estimates are difficult, due to nonaberrant cell infiltration and tumor aneuploidy. SNP-CGH in addition delivers a measure of allelic contrast (BAF). From BAF, allelic bias can be derived, but, e.g., LOH is difficult to determine (due to the nonaberrant cell admixture). ASCAT calculates genome-wide allele-specific copy number profiles for tumor samples, taking into account tumor ploidy and nonaberrant cell admixture. The algorithm first determines the ploidy of the tumor cells and the fraction of aberrant cells ("sunrise plot," *bottom left*). This procedure evaluates the goodness-of-fit for a grid of possible values for both parameters. The optimal solution of tumor ploidy and percentage of aberrant tumor cells is shown by the cross. Next, an "ASCAT profile" is calculated, containing the allele-specific copy-number of all assayed loci (copy-number on the *Y*-axis vs. the genomic location on the *X*-axis; for illustrative purposes only, both lines are slightly shifted such that they do not overlap; only probes heterozygous in the germline are shown). These ASCAT profiles allow accurate derivation of gains (which can be further subdivided into, e.g., duplications, triplications, and amplifications), losses (of one or more copies), copy-number-neutral events, and LOH.

This will create a data structure containing the Log *R* and BAF data for both tumor and germline, as well as some supporting information, such as the position of each probe on the array and a list of the samples.

Next, the data can be plotted:

```
ascat.plotRawData(ascat.bc)
```

These plots are informative to evaluate the quality of the data and to double check if germline samples have not been contaminated with tumor tissue (Fig. 3).

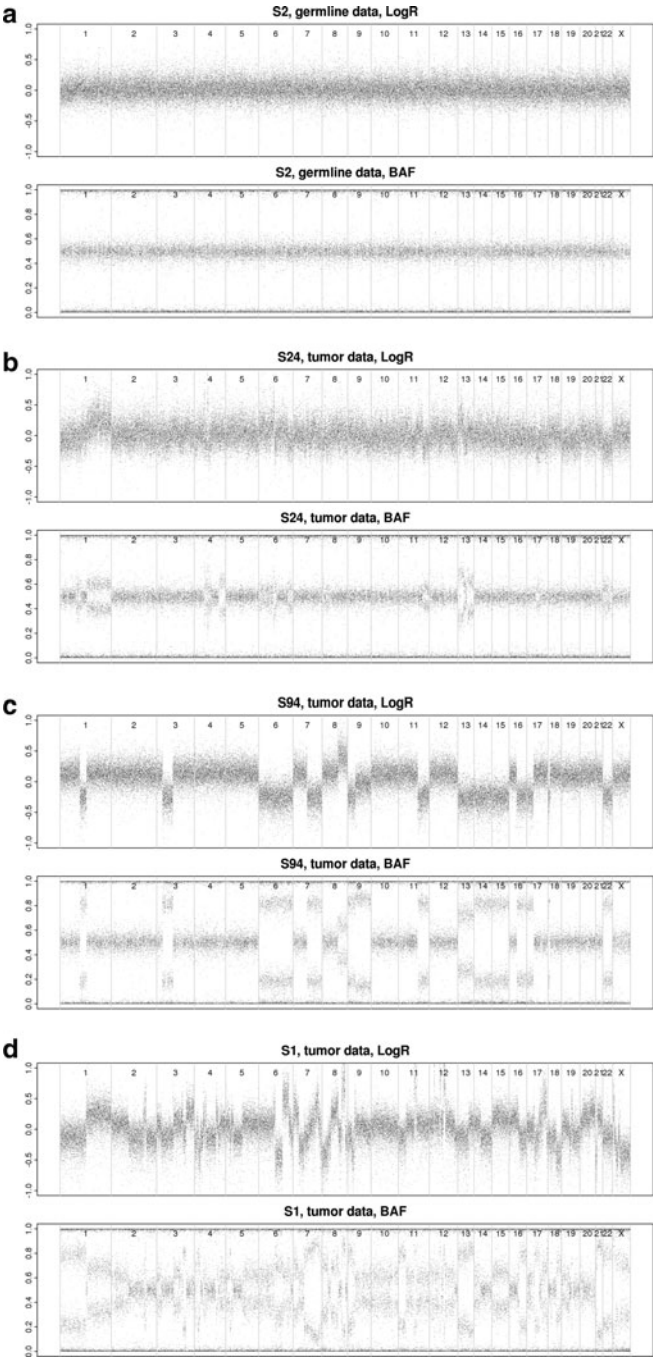


Fig. 3. Example plots of germline and tumor SNP array data. (**a**) A germline sample clearly showing a flat Log *R* profile and three bands in BAF, corresponding to AA (BAF = 0), AB (BAF = 0.5), and BB (BAF = 1) genotypes. (**b**–**d**) Three tumor samples showing a low aberrant cell fraction (limited range of Log *R* and BAF) (**b**), a higher aberrant cell fraction (**c**), and a higher aberrant cell fraction with extensive aberrations (**d**).

### 3.3. Segmentation of SNP Array Data

The loaded SNP array data can subsequently be segmented by the allele-specific piecewise constant fitting (ASPCF) algorithm (see Note 5):

```
ascat.bc = ascat.aspcf(ascat.bc)
```

In a first step, this uses the germline data to determine which SNP array probes are germline homozygous (germline genotypes AA or BB) (see Note 6). For these germline homozygous probes, the BAF data track from the tumor is uninformative for copy number determination, as, e.g., germline genotype AA cannot result in tumor genotypes containing B alleles (e.g., A, AA, AAA are possible, but, e.g., genotype AAB is not), and hence, BAF will always be close to 0. Similarly, germline genotypes BB will result in BAF close to 1 for the tumor data. In a second step, the data is segmented by the ASPCF segmentation algorithm (note that this requires aspcf.R), and the results are added to the ASCAT data structure (see Note 7).

The segmented data can subsequently be plotted, using:

```
ascat.plotSegmentedData(ascat.bc)
```

From these plots (Fig. 4), the quality of the data can be further evaluated (e.g., on samples with a serious wave artifact (14) in Log $R$, ASCAT may subsequently fail (12)).

### 3.4. Running the ASCAT Algorithm

The ASCAT algorithm is next applied to the segmented data:

```
ascat.output = ascat.runAscat(ascat.bc)
```

This output is saved in a data structure, and three figures are made for each tumor. The output data structure contains the aberrant cell fraction and the ploidy, and the copy numbers across the whole genome for both alleles, for each sample. In addition, a list of samples on which ASCAT analysis failed are included [this is often caused by problems with the input data, which can be traced back using the figures generated in the previous sections (Figs. 3 and 4)]. The figures include a "sunrise plot," an ASCAT profile, and a raw copy number profile, for each sample. The sunrise plot is used to determine the optimal aberrant cell fraction and ploidy of the tumor sample and contains a landscape of aberrant cell fraction and ploidy values on which the optimal solution is annotated (Fig. 5). The ASCAT profile contains the estimated allele-specific copy numbers across the genome and can be considered the key output of ASCAT analysis. From these plots, all gains and losses are visualized, as well as copy-number-neutral aberrations and loss-of-heterozygosity (LOH) (Fig. 6). An aberration reliability score for each aberration is also shown in this plot (Fig. 6). The raw copy number profile contains the total copy number, as well as the copy number of the minor allele (the allele with the lowest copy number), without rounding to nonnegative integers (Fig. 7). This plot can be used to evaluate the solution reported

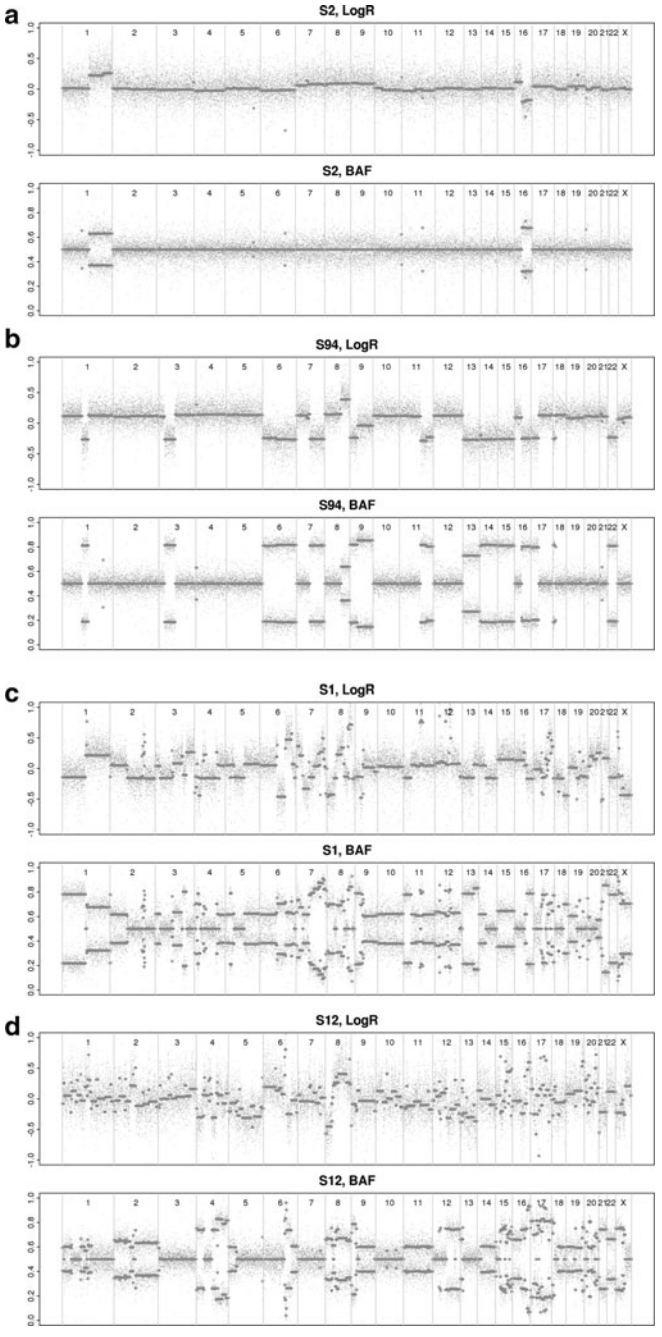


Fig. 4. Example plots of tumor SNP array data, after segmentation. The raw data is plotted, as well as the data after application of the ASPCF segmentation algorithm. (**a**) A sample with few aberrations. (**b**) A sample with more aberrations. (**c**) A highly complex sample. (**d**) A sample showing a clear wave artifact in the Log *R* data track. This is most clearly visible in segments with constant BAF but fluctuating Log *R* (which is not eliminated by the segmentation). In case of such problems, ASCAT may be unable to obtain a solution.

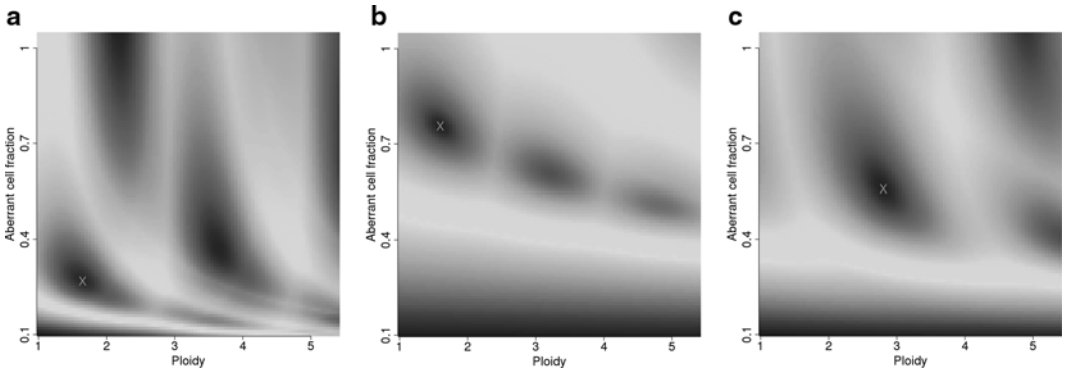


Fig. 5. Example sunrise plots from ASCAT. These plots evaluate different options for the tumor ploidy (*X*-axis) and the aberrant cell fraction (*Y*-axis). For each value plotted, the resulting copy-number profile is evaluated. When the copy-number profile matches whole numbers closely, a good match is obtained (see ref. 12 for details). The optimal match is annotated by a cross. (**a**) A near-diploid sample with a very low aberrant cell fraction (high nonaberrant cell admixture). This is the sample shown in Fig. 3b. (**b**) A near-diploid sample with a high aberrant cell fraction. This is the sample shown in Fig. 4b. (**c**) A near-triploid sample with intermediate aberrant cell fraction. This is the sample shown in Fig. 4c.

by ASCAT. In addition, by scanning for regions that do not fit the whole-number solution, one can gain insight into intratumor heterogeneity (Fig. 7c).

## 4. Notes

1. Owing to privacy issues with genome-wide genotyping data, data access is often limited. For the data used as an example here, a material transfer agreement is in place. For the purpose of reproducing the procedures outlined here, data access will always be granted.
2. The standard output from Illumina SNP array data is Log *R* and BAF, the latter corresponding to $n_B/(n_A + n_B)$, where $n_A$ is the copy number of the A allele and $n_B$ is the copy number of the B allele. The standard output from Affymetrix SNP array data is Log *R* and an Allelic Difference score that corresponds to $\log_2(n_A/n_B)$. Apart from a rescaling/transformation of this measure of allelic contrast, the choice of BAF over Allelic Difference is arbitrary. However, methods exist to directly calculate Log *R* and BAF from Affymetrix CEL files, such as PennCNV (15) and the aroma.affymetrix R package (16) as well as some commercial packages.
3. ASCAT 2.0 has evolved considerably since its inception (12). The ASPCF segmentation algorithm has been ported from MATLAB to R and now has a faster implementation that scales linearly with array density, making it highly suitable also for high-density platforms. ASCAT 2.0 is applicable to SNP array data from both Illumina and Affymetrix (see also Note 2).

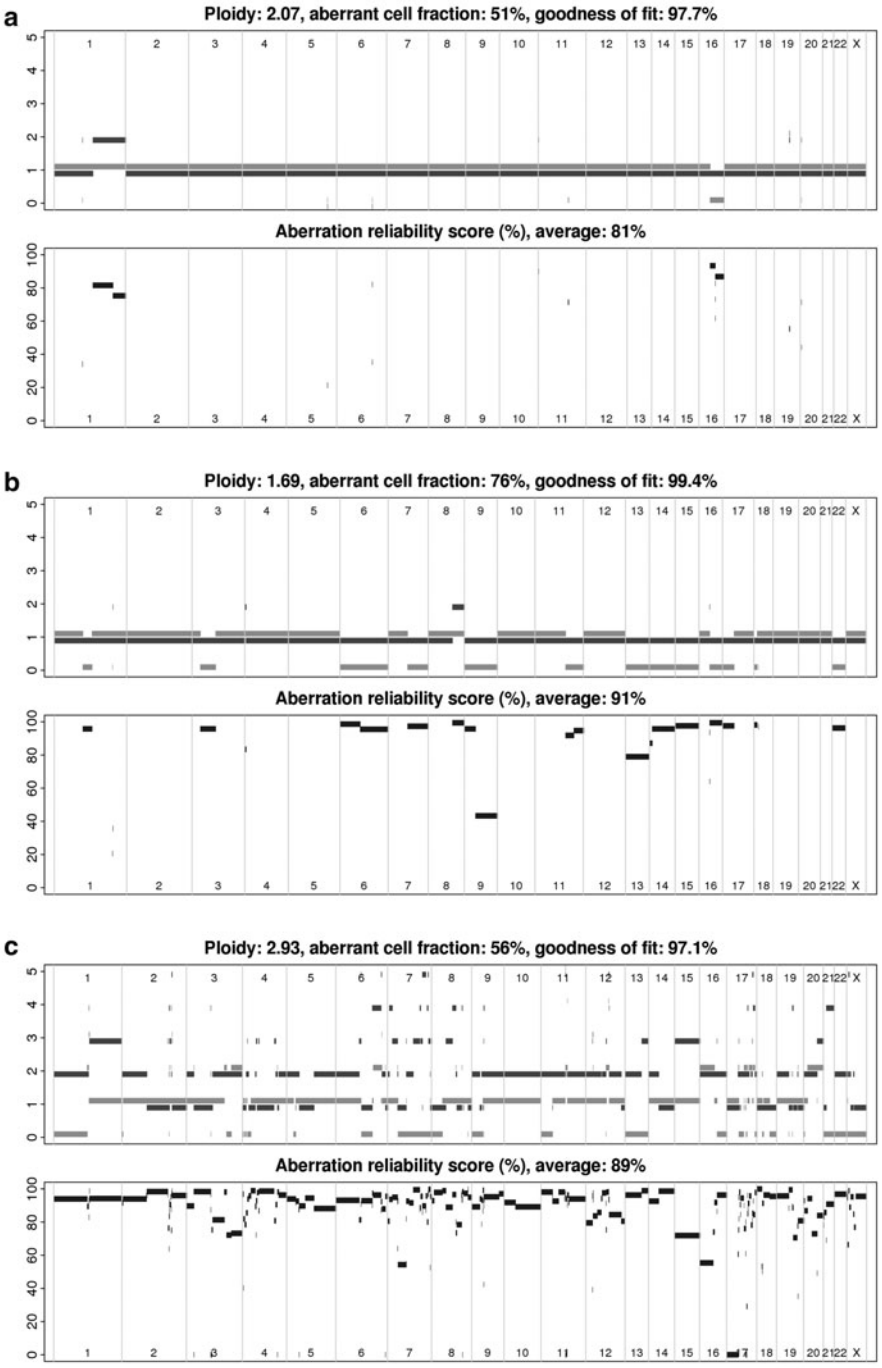


Fig. 6. Example ASCAT profiles and corresponding aberration reliability score plots. The ASCAT profiles (*top*) show the allele-specific copy number across the genome. The copy number of both alleles is shown. All estimated copy numbers are nonnegative whole numbers. Both lines are slightly shifted such that they do not overlap. The aberration reliability score plots (*bottom*) show the confidence one can have in each detected aberration, compared to the hypothesis of no aberration (see ref. 12 for details). (**a**) A sample with few aberrations (shown in Fig. 4a). A duplication of the 1q chromosome arm, as well as a hemizygous deletion (one copy lost) of 16q is immediately apparent. (**b**) A more complex sample (shown in Figs. 4b and 5b). Multiple hemizygous deletions are present, as well as a duplication at 8q. (**c**) A highly complex sample (shown in Figs. 4c and 5c). Few regions in the genome are unaffected by genomic aberrations in this sample.

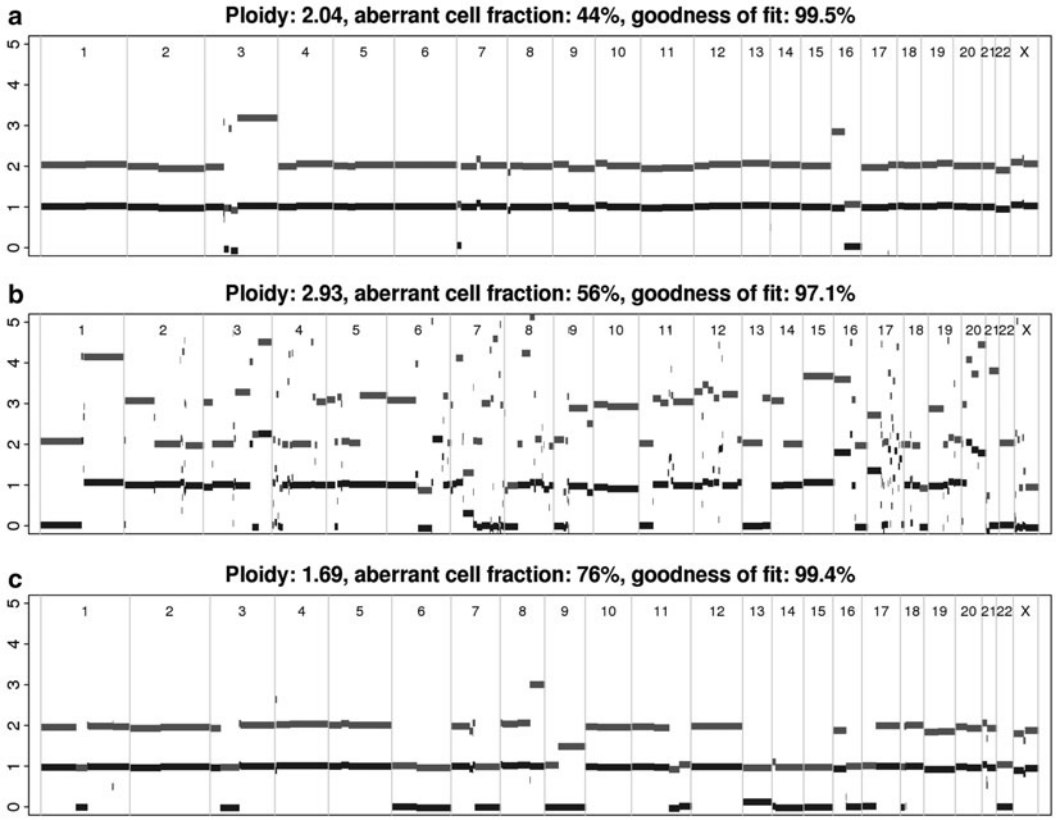


Fig. 7. Examples of raw copy number profile plots from ASCAT. These plots can be used to evaluate the solution reported by ASCAT and to gain insight into intratumor heterogeneity. The copy number of the minor allele is shown, as well as the total copy number, as directly derived from the data, without rounding to nonnegative integers. When a good solution is obtained, most or all regions should be close to whole numbers. In cases where a good global fit is obtained, yet some particular regions show copy numbers that are far from integers, these regions are likely subject to intratumor heterogeneity. (**a**) A sample showing few aberrations. All calculated copy numbers are close to integers, confirming a close fit. (**b**) A highly complex sample (shown in Figs. 4c, 5c, and 6c). Copy numbers clearly cluster close to integers. (**c**) A sample of intermediate complexity (shown in Figs. 4b, 5b, and 6b). All segments cluster close to whole numbers, except one on chromosome 9. One copy of the entire chromosome 9 has been lost in all tumor cells. In addition, the copy number of the q arm of chromosome 9 is close to 1.5, suggesting that there are two major subclones in the tumor: about 50% of the aberrant tumor cells show a gain of (the remaining copy of) 9q, while the other 50% of tumor cells do not have this gain. Due to the bad fit to whole numbers of this segment, the 9q arm also shows a clear drop in the aberration reliability score (Fig. 6b).

4. Some SNP array platforms (e.g., Affymetrix SNP 6.0) contain copy-number-only probes. These are probes in non-SNP locations. ASCAT can take these copy-number-only probes into account and calculate the total copy number at these loci. As no allelic contrast information is available, these copy-number-only probes should have NA values in their BAF data.
5. A segmentation algorithm of choice can be inserted in this step. The ASPCF segmentation algorithm, as part of the

ASCAT package, segments Log *R* and BAF simultaneously (automatically accounting for the structure and symmetry of BAF). ASPCF segment borders in Log *R* and BAF are automatically aligned (and optimized using data from both tracks). However, Log *R* and BAF can also be segmented separately using another segmentation algorithm (e.g., CBS (17)), without causing problems in later steps of the data analysis.

6. Removal of germline homozygous probes is most easily performed when matched germline samples (i.e., from the same individual) are also profiled by SNP arrays. If this material is not available, these homozygous probes can still be eliminated, e.g., by applying a threshold or by more specialized procedures. We aim to include an automated function to infer germline genotypes from tumor data in the next release of ASCAT.
7. The ASPCF segmentation algorithm is the computationally intensive step of the pipeline. However, this step can be executed in parallel, by using, e.g.,

```
ascat.bc = ascat.aspcf(ascat.bc, 1:5)
```

to segment the first five samples of a dataset. For every sample, two files are created containing the segmented Log *R* and BAF data. When these files exist upon execution of the `ascat.aspcf()` function, the results are read from disk rather than recalculating. Hence, by first splitting the segmentation over multiple processors, copying the resulting segmentation files to one directory and finally executing

```
ascat.bc = ascat.aspcf(ascat.bc)
```

this segmentation can be easily parallelized.

## References


1. McCarroll SA, Kuruvilla FG, Korn JM et al (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40:1166–1174.
2. Peiffer DA, Le JM, Steemers FJ et al (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16:1136–1148.
3. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458:719–724.
4. Balmain A, Gray J, Ponder B (2003) The genetics and genomics of cancer. Nat Genet 33 Suppl:238–244.
5. Witz IP, Levy-Nissenbaum O (2006) The tumor microenvironment in the post-PAGET era. Cancer Lett 242:1–10.
6. Navin N, Krasnitz A, Rodgers L et al (2010) Inferring tumor progression from genomic heterogeneity. Genome Res 20:68–80.
7. Sun W, Wright FA, Tang Z et al (2009) Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic Acids Res 37:5365–5377.
8. Staaf J, Lindgren D, Vallon-Christersson J et al (2008) Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol 9:R136.
9. Attiyeh EF, Diskin SJ, Attiyeh MA et al (2009) Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. Genome Res 19:276–283.

10. Greenman CD, Bignell G, Butler A et al (2010) PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics 11:164–175.
11. Popova T, Manie E, Stoppa-Lyonnet D et al (2009) Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome Biol 10:R128.
12. Van Loo P, Nordgard SH, Lingjærde OC et al (2010) Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A 107:16910–16915.
13. http://www.ifi.uio.no/bioinf/Projects/ASCAT
14. Marioni JC, Thorne NP, Valsesia A et al (2007) Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol 8:R228.
15. Wang K, Li M, Hadley D et al (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17:1665–1674.
16. Bengtsson H, Irizarry R, Carvalho B et al (2008) Estimation and assessment of raw copy numbers at the single locus level. Bioinformatics 24:759–767.
17. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23:657–663.