Published in final edited form as:

Methods Mol Biol. 2014 ; 1176: 243–259. doi:10.1007/978-1-4939-0992-6_20.

Computational methods for DNA copy-number analysis of tumors

Jude Kendall and Alexander Krasnitz

Summary

Study of DNA copy-number variation is a key part of cancer genomics. With the help of a comprehensive multi-step computational procedure described here copy-number profiles of tumor tissues or individual tumor cells may be generated and interpreted, starting with data acquired by next-generation sequencing. Several of the methods presented are specifically designed to handle cancer-related copy-number profiles. These include accounting for variation of ploidy and distilling somatic copy number alterations from the inherited background.

Keywords

DNA copy number; next-generation sequencing; cancer; normalization; segmentation; single-cell analysis

1. Introduction

Somatic DNA copy number variation is a ubiquitous feature of cancer genomes (1). The spectrum of alterations is broad and includes both amplifications and deletions, with the genomic extent ranging from less than a single gene to an entire chromosome. Research interest in genome-wide study of these alterations is motivated by their relation to the onset and progression of the disease; by their value as markers for clinical outcome and sensitivity to treatment (2); and, more recently, by their utility in reconstructing the genealogy of tumor cells (3). In addition to comparative genomic hybridization, next-generation sequencing (NGS) is being increasingly employed as a method for copy-number data acquisition. The obvious advantage of NGS over array-based technology is its ability to provide truly genome-wide coverage, not restricted to array probes.

Here we summarize methodologies we have been using for in silico copy-number analysis, with the focus on the analysis of NGS-derived data. Starting with a set of sequencing reads, a succession of computational tools produces first noisy, then piecewise-constant estimates of copy number profiles of tumor tissues or individual tumor cells. These are further interpreted to identify copy-number alteration events.

2. Methods

There are many methods for estimating DNA copy number genome-wide. Recent large-scale studies have used various microarray and high throughput sequencing methods (1, 3). In all cases the idea is to measure a quantity that correlates predictably with DNA copy number. In the case of microarrays the fluorescent intensity of spots on the array is higher for higher copy number. For high throughput sequencing the depth of coverage correlates with copy number. Also, paired end or mate pair sequencing can be used to find deletions and amplifications. We describe here data analysis methodology to infer DNA copy number alterations from high throughput sequence data. Although customized for processing data generated by Illumina sequencing machines, this methodology can be readily adapted to handle output from other sequencing platforms. A reader interested in the details of sequencing technology should consult the Illumina web site (4).

Important aspects of our methodology depend on whether DNA being analyzed originated in a single cell or in bulk tissue. In case of the former, we take advantage of the fact that copy number is integer everywhere in the single-cell genome. In case of the latter, the observed copy number profile is an average over cells representing a variety of clonal populations in the tumor, admixed with normal tissues, and we must account for this additional source of variability.

Our focus here is exclusively on in silico analysis of individual copy number profiles. Experimental techniques underlying this analysis, in particular DNA profiling of single cells, are described in detail in separate publications (3, 5). On the other, hand single-profile analysis often is followed up by examination of multiple copy-number profiles, to infer regions of copy number aberration recurrent in tumors from multiple patients (6) or to infer clonal structure and tumor evolution in the case of single cells from one tumor or patient (3). Methodology of this multi-profile analysis is also presented in detail in a separate publication (7).

The main idea used here in inferring DNA copy number from high throughput sequence data is that the read depth in a region of the genome is roughly proportional to the DNA copy number. Deviations from this proportionality are mainly due to two factors. First, Illumina sequencing machines produce short read data, generally 50 to 150 bases long. A sizable fraction (10–15%) of these short reads cannot be uniquely localized in the genome. Secondly, there are sequence-dependent biases affecting the read depth, including, but likely not limited to, those related to PCR amplification, adapter ligation and spot forming efficiency on the Illumina flow cell surface. Local GC content of the genome affects all these steps and is correlated with large biases, variable from sample to sample, but can be easily normalized. (see Note 1).

The data analysis consists of the following steps,

¹For single cell analyses in our lab we use the Sigma Genomplex whole genome amplification (WGA) kit. Rubicon's Picoplex WGA kit has different biases and must therefore be normalized using different methods to get optimal results. A similar adaptation is required to any other method of DNA sample preparation.

1. Partition the genome into bins with an expected equal number of mappable positions.
2. Count the number of reads mapping to each genome bin.
3. Remove from subsequent analysis bins that have consistently high read counts.
4. Normalize the read counts based on GC content of each bin.
5. Estimate the underlying piecewise-constant profile by segmentation.
6. If analyzing single cell data, for each segment estimate an integer copy number.
7. Center the profiles and determine segments with altered copy number.
8. If analyzing bulk DNA, determine segments with altered copy number by comparison with profiles of cancer-free tissues.
9. Mask inherited copy number alterations.
10. Identify copy number altering events.

These individual steps are organized into a workflow, as shown in Figure 1. Software code samples for steps 1–5 are given in (5). Software packages for steps 7–9 are publicly available from the Comprehensive R Archive Network (CRAN) at the following web locations: <http://cran.us.r-project.org/web/packages/ParDNACopy/> and <http://cran.us.r-project.org/web/packages/CNprep/>.

2.1 Partition the genome into bins with an expected equal number of mappable positions

The Illumina sequencing machines produce 50–150 bases of sequence data from the ends of DNA molecules. This can be done by sequencing just one end of each molecule, called single read sequencing, or by sequencing both ends of each molecule, called paired end sequencing. We count the number of molecules that map to each genome bin, so we typically use single read sequencing. The sequencing machines output a data file in FASTQ format with four lines for each read, illustrated in Figure 2(A). Line 1 is a read ID, line 2 is a sequence consisting of A, C, G, T or N, line 3 is a plus sign, line 4 is a quality score for each base (8). The quality score is an ASCII-encoded phred quality score (9). Mapping these reads to a reference genome consists in finding the 100 base long sequence in the reference genome that most closely matches the sequence read. There are usually base calling errors within each read, especially in the low quality base calls occurring at the end of the read. An alignment with the fewest mismatches to the reference genome is found, and if no other location in the genome aligns with the read with the same number of mismatches, then the read is considered to be uniquely mapped to the reference genome. There are many software packages available for short read alignment. In this analysis we use Bowtie (10). Bowtie takes as input the sequence reads in FASTQ format and a reference genome sequence and outputs alignments in SAM file format (11). Figure 2(B) is an example of an alignment in SAM format. There is one tab-delimited line for each alignment. The fields of interest to us are: field 1, the read ID; field 2, a mapping code, which for single read mappings is 0 for an

alignment on the plus strand and 16 for an alignment on the minus strand; and fields 3 and 4, the chromosome and chromosome position to which the read aligns.

Not every 50-base sequence in the human genome is unique to just one location. (see Note 2). For example, the sequence consisting of 50 A-s matches over 5000 places in the GRCh37 reference human genome. For such a read we cannot know from where in the genome it was sampled. In a typical sequencing protocol DNA is extracted from one or many cells and in effect randomly sampled and sequenced. The DNA library will consist of billions of short DNA fragments, typically about 300 bases long. Output from a sequencing run can contain from 10 to 300 million reads. Samples can be indexed and sequenced together in a library giving as few as one million molecules per sample (see Note 3). The simplest read depth data analysis could consist of partitioning the genome into for example 100,000 bins of equal length of about 30,000 bases in the case of the human genome. But due to the existence of a large amount of repetitive sequence in the human genome, from transposons, microsatellites, and segmental duplications, some of these bins will have very few 50 base sequences that are unique in the genome.

In order to compensate for these biases we map every 50 base long sequence (50-mer) in the reference genome. For each of these sequences we know the chromosome and chromosome position. Start positions for 50-mers that are unique in the genome are called uniquely mapping positions. We then partition the genome into 20,000 bins each with the same number of uniquely mapping positions, plus or minus 1 to handle rounding to integers (see Note 4). For 20,000 bins and the GRCh37 version of the reference genome the median bin size is about 120,000 uniquely mapping positions. The base count varies from bin to bin, and some bins near centromeres may span many megabases.

2.2 Count the number of reads mapping to each genome bin

For each uniquely mapped sequence read we know the chromosome and chromosome position of its alignment. We take the starting position of each alignment, find the bin in which it is contained and increment the read count for that bin. The result of counting each read is a file as illustrated in Table 1. The first column is the bin number, the second column is the chromosome of the bin, the third column is the starting position of the bin along the chromosome (the bin ends at the position just before the start of the subsequent bin), the

²Bin boundaries are set giving bins with equal mappability of uniquely mappable 50-base sequences. Using barcoded (indexed) WGA DNA samples there are 38 bases of barcode and primer to remove from the 5' end of each read. For 76 base reads this leaves 38 mappable bases and for 100 base reads 62. We have found that bins based on mappability of 50 base reads work similarly enough for 38 base and 62 base reads that these bin boundaries can be used in both cases. For a given single study with different samples sequenced at different lengths the longer reads are truncated so that the same number of bases are mapped for all samples. Bin boundaries can be computed for the desired mapped read length.

³In our studies of tumor heterogeneity we are interested in profiling the genomes of as many cells from a tumor as we can afford, so we are trying to get a usable genome copy number profile with as few reads as possible. For aneuploid solid tumors we have found that splitting the genome into 5000 bins and generating 250,000 reads per cell is useful. This corresponds to a read depth of 1 across 5% of the human genome. This is very different from the data analyses most people are publishing. It is more typical to try to get as much coverage of the genome as possible and high read depth in order to accurately determine point mutations and SNPs.

⁴For the examples in this paper we are showing data from the human genome partitioned in to 20,000 bins. This gives good copy number profiles with about 1 million reads per cell. If the purpose of a study is to sample as many cells as possible from a tumor at low cost fewer bins can be used. This might be the goal in a study of tumor heterogeneity. In a study with some interest in finding amplified or deleted genes more sequence data can be obtained by sequencing fewer samples per flowcell lane and partitioning the genome into smaller bins. This would allow the detection of smaller regions of variation where maybe a single gene is copy number zero, or a small sub-megabase region is amplified to many copies.

fourth column is the bin count or the number of reads uniquely mapping to a position within the bin boundaries.

2.3 Remove anomalous bins

There are dispersed regions in the genome that accumulate very high read depth compared to expected read depth based on the number of reads, for example, the bins surrounding the centromere on chromosome 1. These bins consistently, across vast majority of the profiles, have 5 to 10 times as many reads as the median bin count across the genome. A succession of such bins would be mistakenly identified as a region of copy number gain. An example is shown in Figure 3 where these bins are referred to as "bad bins." Anomalously high read counts are observed in the very large bins consisting of mostly repetitive DNA sequence. For example, there are over a million Alu elements in the human genome. The older copies are mutated over time. The combination of short read mapping, base mutations in the reference genome, base-calling errors in the sequence data, and errors in the reference genome in highly repetitive regions combine to make mapping in these regions less reliable than is generally the case. We have a list of bins empirically determined to consistently accumulate an unusually high number of mappings. These bins are removed from subsequent data analysis (see Note 5).

2.4 Normalize the read counts based on the GC content of each bin

There are large biases in read count correlated with GC content. The result is that the read counts for each bin, rather than being directly proportional to DNA copy number, are distorted. This phenomenon is illustrated in Figure 4 which shows a scatter plot of bin read count against the bin GC content, defined as the share of G or C in the reference sequence of the bin. That is, bins with high GC content tend to have lower bin counts than bins with an average GC content. Importantly, the GC bias varies from sample to sample, meaning that, whereas variation of mappability across the genome could be assessed once and adjusted for by making bins of variable length, it is necessary to adjust the read counts for GC content in each sample individually. To this end we fit a Lowess curve to the GC content by bin count data and subtract from each bin read count its respective Lowess estimate. As can be seen in Figure 2, the bias is thereby dramatically reduced. This correction is analogous to using Lowess fit for the normalization of intensity-dependent dye bias in two-color microarray data (12). The original and GC-adjusted read count values are also compared for a small number of bins in Table 1.

2.5 Estimate the underlying piecewise-constant profile by segmentation

The bin read count as a function of the genomic position represents the copy number of the sample superimposed with observational noise. The copy number itself is a piecewise-constant function of the genomic position, resulting from copy number gains and losses, both inherited and caused by cancer. Regions of copy-number variation vary in length from

⁵We found that for a 20,000-bin partition of the genome 38 bins consistently had anomalously high read counts and had to be removed. The reads in these bins are not uniformly distributed. There are sparsely distributed regions of high read density of about 1000 bases in size. An alternative to removing these bins would be to find and mask these regions throughout the genome. Preliminary results show that this approach can reduce read count variability throughout the genome in addition to removing the false positive amplification segments around the centromeres. Work is in progress on this alternative method.

whole chromosomes and chromosome arms down to single genes. Our ability to find these regions depends both on the level of noise in the data and on the sampling rate, i.e., the number of bins per base pair. Usually a region cannot be reliably detected unless it spans multiple bins. A procedure for estimating the underlying piecewise-constant copy-number profile from noisy bin-count data is called segmentation. We use the DNACopy segmentation algorithm for this purpose (13). In single cells each of these segments corresponds to a region of chromosome with an integer number of copies. In DNA samples extracted from many cells the copy number in each of these regions represents an average across a population of cells. The segmented profile is compared with the original noisy data genome-wide in Figure 5 and for a small number of bins in Table 1. Figure 5 also illustrates the importance of bias removal prior to segmentation.

2.6 If analyzing single-cell data, for each segment estimate an integer copy number

As indicated above, when analyzing data from a single cell we know that the copy number must be an integer at any point in the genome. This means that if the segmented values accurately represent the copy number in each bin then the genome-wide distribution of these values should be multimodal and sharply peaked, with each mode representing an integer copy number value present in the genome. With this expectation we estimate the copy number genome-wide, assuming in addition that (a) the segmented values are approximately proportional to the copy number and (b) the genome-wide mean of the copy number lies in the 1.5–5.5 range. The estimate is obtained by least-squares fit (14): denoting the nearest integer by $[\cdot]$, the genome-wide mean by $\langle \cdot \rangle$ and the segmented values by S , we minimize the mean squared rounding error $\langle (MS - [MS])^2 \rangle$ with respect to multiplier M , subject to the condition $1.5 \leq \langle MS \rangle \leq 5.5$. The estimated copy number is then given, for each bin, by the corresponding $[MS]$ computed at the optimal value of M . The first additional assumption we make is justified a posteriori: the mean squared rounding error is small for the vast majority of single cells we have analyzed. Armed with this experience, we use this error for quality control: copy number profiles with large errors either may originate from erroneous pooling of DNA from multiple cells, or derive from parts of cells, or represent degraded DNA (see Note 6). Our second additional assumption is justified by observation: cell ploidy, as measured by flow cytometry, very rarely lies outside the 1.5–5.5 range. Indeed, we find the optimal value of $\langle MS \rangle$ to lie in the interior of this range in a vast majority of cases.

The copy number estimation step is illustrated in Figure 6. The left panel of the figure shows a segmented genome-wide profile of a single cell from the hypo-diploid cell population (3). It is evident that the genome is predominantly, and nearly evenly, distributed between two values of the copy number, one of which is roughly twice the other. These two values are estimated as described to be 2 and 1 (right panel). Estimated values of MS are also listed in Table 1.

⁶The method to convert the segmented values to a copy number estimate described in our earlier (3) and in more detail in our more recent (5) publications is different than the method described here. There are two reasons to prefer the new method. First, unlike the earlier method, it is applicable to cancer-free cells, with copy number of 2 nearly everywhere in the genome. Secondly, the minimal value of the mean squared rounding error indicates the quality of the estimate and can be used for quality control.

2.7 Center the profiles and determine segments with altered copy number

Tumor cells commonly undergo whole-genome copy number gains and losses, resulting in changed ploidy. It is important to be aware of these changes in ploidy in order to correctly infer individual copy number alterations in single-cell profiles. For example, copy number of 2 would be a loss in a triploid but not in a diploid genome. To estimate the ploidy of a cell, we first find, for each chromosome other than X or Y, the integer copy number where most of the chromosome's length resides, then take the most frequent of these integers among the chromosomes considered as the ploidy of the cell. Copy number changes in individual segments are computed relative to ploidy.

In the analysis of bulk tumor DNA cells comprising the sample need not all have the same ploidy or harbor the same copy number alterations. As a result, segments comprising the profile need not differ by integer copy number, and it is not obvious what value of the copy number estimate corresponds to the normal state of the cells, relative to which somatic alterations occur. We therefore adopted the following multi-step process to infer the likely normal state and identify segments with significant copy number deviation from that state. First, a bootstrap sample is generated from the log normalized bin count data for the median value of each segment. The bootstrap sample size is proportional to the number of bins in the segment (see Note 7). These samples are joined, and the result is modeled as a normal mixture using expectation maximization (EM) algorithm (15). We usually allow the mixture components to both have differing means and standard deviations. The optimal number of components for the mixture is chosen based on the Bayesian information criterion (BIC). The algorithm is re-initialized at random 10 times, and the result with the highest BIC is chosen as final. This result contains, in addition to the component means and standard deviations, the component identities for all values in the sample. The sample is therefore divided into clusters depending on the component label. Next, we seek the maximal overlap of any two components. If that overlap exceeds a predetermined value (typically 0.25 to 0.4), the two clusters are joined. This search and joining of the most overlapping pair of clusters is repeated until no two of the remaining clusters overlap by more than the preset value. At this point in the process we seek the cluster whose median value is nearest 0. If that cluster represents more than a certain predetermined portion of the genome, it is declared a normal-state cluster (see Note 8). Otherwise, the two clusters with medians nearest 0 are joined, a process repeated until the result represents at least the required portion of the genome. Once the normal-state cluster is determined, its median is taken as the estimate of the normal-state log segmented value. We then estimate and tabulate, for each segment, its marginal probability to belong to that cluster. To that end, the segment median is bootstrap-sampled a fixed number of times (typically 20 to 50), and the marginal probability to belong to the normal-state cluster is averaged over the sample values. Segments for which this probability is low are flagged as likely to be amplified or deleted, depending on the sign of their deviation from the normal state. This criterion may be used alone or in combination with a

⁷This proportionality coefficient typically is in the 0.02–0.05 range.

⁸The minimal portion of the genome to be represented by the normal-state cluster is tuned depending on the cancer type. In breast cancer the expectation generally is that no less than 40–50% of the genome retains its original copy-number; for ovarian cancer that expectation may be as low as 15–20%.

criterion based on comparison of the tumor-derived data to those derived from cancer-free tissues.

2.8 If analyzing bulk DNA, determine segments with altered copy number by comparison with profiles of cancer-free tissues

If a representative set of copy number profiles is available for cancer-free tissues, it can be used to identify segments with significant copy number deviation from the norm in tumor-related profiles. For this identification to be meaningful it is essential for the tumor-related and cancer-free DNA to have undergone the same preservation, extraction and sequencing procedures. If this is the case, we remove from the cancer-free data all segments whose copy number is likely to deviate from 2. These are short segments, usually defined as shorter than 5Mb, and segments in the chromosomes X and Y. For the remaining cancer-free data we determine, separately for each profile, the genome-wide median for the log of segmented values and tabulate, for each segment, its deviation from the median. Next, tumor-derived segments are compared one by one to the joint set of long, autosomal cancer-free segments. This comparison yields, for each tumor-derived segment, the fraction of the total length of the cancer-free segment set for which the log segmented value deviation from the norm exceeds that of the tumor-derived segment. This fraction is small for tumor-derived segments with altered copy number and may therefore be used to identify amplified or deleted segments, alone or in combination with the clustering-based criterion described in the previous sub-section.

2.9 Mask inherited copy number alterations

Some of the amplified or deleted segments identified in tumor-derived copy number profiles are not somatic and specific to the tumor tissue or cell but inherited and common to all cells of the organism. These inherited alterations may be masked or retained in the profile, depending on the final goal of the analysis. In particular, if the goal is to determine somatic alterations across multiple tumors, it is advisable to mask the inherited alterations. If, on the other hand, the goal is to reconstruct the genealogy or the population structure of cells within a tumor, this step may not be necessary, as such inherited alterations are shared by all the cells under investigation and will have no impact on the analysis downstream.

If masking is necessary, its implementation and results depend on the availability of the copy number profile of a matching normal tissue. If available, a matching profile makes it possible to mask all inherited alterations, independent of their frequency in the population. We then proceed as follows. First, we identify amplified (deleted) segments in the tumor profile that are matched closely by amplified (deleted) segments in the normal profile. The matching is quantified as the fraction of the tumor-derived segment covered by the segment from the normal profile. We usually require this measure to exceed 0.6 to trigger masking. The segment slated for masking is then removed as described in Figure 7: a point is chosen at random within the segment and neighboring segments are extended to the chosen point.

If a matching profile is not available, we must rely on population-wide inherited copy number variation data to generate a mask. The mask is defined by two frequency thresholds, the high threshold T , $0 < T < 1$, and the low threshold t , $0 < t < T$ (see Note 10). For a given type

of alteration (amplification or deletion) we then identify all contiguous regions of the genome where the alteration frequency exceeds T . Each such region is embedded in a (usually broader) region where the frequency exceeds t . The latter region is added to the mask. The mask is thus a set of genomic intervals. If an altered segment of the same type in the tumor is matched by one of these intervals, it is masked as described above. By definition, there will be no masking of copy number alterations whose frequency in the population does not exceed t .

2.10 Identify copy number altering events

Segments observed in a copy-number profile do not necessarily coincide with the copy-number altering events that gave rise to the observation. However, these events may be reconstructed from the observed profile under certain assumptions. Our reconstruction procedure, called slicing, is especially straightforward for single-cell profiles, where all copy-number values are integer. The assumptions underlying slicing are that (a) all copy-number events that have occurred in the ancestral history of the cell are observable in its copy-number profile, (b) there is no event overlap involving both amplification and deletion events and (c) the only allowed form of event overlap is nesting. The procedure consists of first identifying event blocks in the profile. An event block is a succession of segments, all of the same type (amplification or deletion), either ending at a segment that is not of that type or at a chromosome boundary. Consistent with the assumptions (a–b), the events are the found by maximally extending each segment within the profile (Figure 8).

For the slicing procedure as described to work events in the event block must all correspond to integer copy-number alterations. This property makes it easy to determine which segments form a copy-number event. However, this property does not hold for copy-number profiles obtained from bulk tumor DNA. Application of slicing to such profiles requires the following preliminary step we call tiering (16). The tiers are formed by first selecting a group of seed segments in the block with a maximal total length subject to a condition that any two segment medians in the group are different with 95% confidence. The tiers in a block are numbered in the order of increasing seed median. The remaining segments are each assigned to the tier with the nearest seed median. Slicing is then carried out in the block, with the tier being the analog of the integer copy number in the single-cell case.

References

1. Beroukhi R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M. The landscape of somatic copy-number alteration across human cancers. *Nature*. 2010; 463:899–905. [PubMed: 20164920]
2. Hicks J, Krasnitz A, Lakshmi B, Navin NE, Riggs M, Leib E, Esposito D, Alexander J, Troge J, Grubor V, Yoon S, Wigler M, Ye K, Borresen-Dale AL, Naume B, Schlicting E, Norton L, Hagerstrom T, Skoog L, Auer G, Maner S, Lundin P, Zetterberg A. Novel patterns of genome

- rearrangement and their association with survival in breast cancer. *Genome Res.* 2006; 16:1465–1479. [PubMed: 17142309]
3. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M. Tumour evolution inferred by single-cell sequencing. *Nature.* 2011; 472:90–U119. [PubMed: 21399628]
 4. Illumina sequencing technology. [Accessed May 20, 2013] Available from: http://www.illumina.com/technology/sequencing_technology.ilmn.
 5. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B, Wigler M, Navin N, Hicks J. Genome-wide copy number analysis of single cells. *Nat Protoc.* 2012; 7:1024–1041. [PubMed: 22555242]
 6. Bric A, Miething C, Bialucha CU, Scuoppo C, Zender L, Krasnitz A, Xuan ZY, Zuber J, Wigler M, Hicks J, McCombie RW, Hemann MT, Hannon GJ, Powers S, Lowe SW. Functional Identification of Tumor-Suppressor Genes through an In Vivo RNA Interference Screen in a Mouse Lymphoma Model. *Cancer Cell.* 2009; 16:324–335. [PubMed: 19800577]
 7. Krasnitz A, Sun G, Andrews P, Wigler M. Target inference from collections of genomic intervals. *Proc Natl Acad Sci.* 2013 in press.
 8. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 2010; 38:1767–1771. [PubMed: 20015970]
 9. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 1998; 8:186–194. [PubMed: 9521922]
 10. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10
 11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25:2078–2079. [PubMed: 19505943]
 12. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 2002; 30
 13. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics.* 2004; 5:557–572. [PubMed: 15475419]
 14. Wigler M. Private communication. 2013
 15. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc.* 2002; 97:611–631.
 16. Xue W, Krasnitz A, Lucito R, Sordella R, VanAelst L, Cordon-Cardo C, Singer S, Kuehnel F, Wigler M, Powers S, Zender L, Lowe SW. DLC1 is a chromosome 8p tumor suppressor whose loss promotes hepatocellular carcinoma. *Gene Dev.* 2008; 22:1439–1444. [PubMed: 18519636]

Figure 1.

A flowchart for the computational procedure described in the Methods section.

A**Read data from sequencing machine in fastq format.**

```
@GA200001:79:FC64N77AAXX:7:69:10027:15968 1:N:0:
GACAAAATTTTCATTCTTGTGACCAGGTTGGAGTGCAA
+
DDDDDBDD>DDDDDDDDDDDDDD<DDBDDDDDDDD>9,B>B
```

B**Read data after mapping in SAM format.**

```
GA200001:79:FC64N77AAXX:7:69:10027:15968 16 chr10 77560 255
38M * 0 0 TTGCACTCCAACCTGGTCAACAAGAATGAAATTTTGTC
B>B,9>DDDDDDDBDD<DDDDDDDDDDDDDD>DDBDDDD XA:i:0 MD:Z:38 NM:i:0
```

Figure 2.

Illumina sequencing data format. A - For each spot on the flowcell surface the Illumina sequencing machines output 4 lines of data in FASTQ format. Line 1 is a read ID, line 2 lists the bases called, line 3 is a plus sign and line 4 is the phred quality score for each base, ASCII-encoded. B - A SAM file format of an alignment of a sequence read to the human genome. There is one tab-delimited line for each read aligned, showing the read ID, read sequence and quality scores and the alignment position in the reference genome.

Figure 3. Identification and removal of anomalous bins. The left panel shows a segmented copy-number profile of a diploid single-cell genome with all 20,000 bins included in the data analysis. The right panel shows a segmented profile of the same cell after removing from analysis bins at centromeres with consistently high read counts in many diploid cells. In this case 38 bins are removed leaving 19,962 bins. The normalized bin counts prior to segmentation are shown in gray in both panels.

Figure 4.

Sample-dependent GC bias and its removal. The two plots on the left show read count versus GC content for the genome bins in two different diploid cells from a breast cancer patient. The bins with very low read count are from the Y chromosome. These illustrate that the GC bias differs between samples and must be adjusted for separately for each sample. The right two panels show the same data after GC bias removal.

Figure 5.

The effect of GC bias removal on segmentation. A - A segmented profile for a diploid cell prior to GC bias removal. Multiple spurious deletions are seen in regions of high GC content, frequently located near telomeres. B - A segmented profile for the same cell after GC normalization. The copy number zero region on the far right of both panels is the Y chromosome. The normalized bin counts prior to segmentation are shown in gray in both panels.

Figure 6.

Integer copy number estimation. A - A segmented profile for a hypo-diploid tumor cell with DNA content of approximately 0.85 that of a diploid genome. The segmented values are so that their genome-wide mean is 1. The horizontal lines are drawn at the values of 0.5, 1.0, 1.5 and 2.0 of the vertical coordinate. B – The integer-valued copy-number profile of the same cell, determined by minimization of the mean squared rounding error. The horizontal lines are drawn at copy number values of 1, 2, 3 and 4.

Figure 7.

Masking of inherited copy-number variation. A – A portion of a segmented profile. An amplified segment (shown in red) is overlapped by an inherited amplified region (a region between the green dashed lines). B – A position, shown by a green dashed line, is chosen at random within the amplified segment. The neighboring segments are extended to that position.

Figure 8. Slicing. A portion of the integer copy number profile of chromosome 1 in a cell is sliced into four unique interval events (in red). The ploidy value for this cell is 4 (shown by a dashed line).

Table 1

A bin count data table for the first 5 bins of chromosome 1 in a genome partitioned into 20,000 bins. Each row corresponds to a bin. The columns, from left to right, tabulate the ordinal bin number (“bin”), the chromosome number (“chrom”), the bin start position in the chromosome (“chrompos”), the read count in the bin (“bincount”), the read count normalized by the mean read count per bin (“ratio”), the proportion of G or C in the bin (“gc.content”), the normalized read count with the GC bias removed (“lowratio”), the segmented normalized read count (“seg.mean”), the “lowratio” column multiplied the optimal M (cf Methods, 2.6, “ratio.q”), the optimal MS (cf Methods, 2.6, “seg.q”)

bin	chrom	chrompos	bincount	ratio	gc.content	lowratio	seg.mean	ratio.q	seg.q
1	1	0	53	1.09	0.45	1.03	1.00	3.15	3.06
2	1	892036	31	0.65	0.63	0.82	1.00	2.50	3.06
3	1	1033528	49	1.01	0.60	1.18	1.00	3.60	3.06
4	1	1172212	59	1.21	0.62	1.53	1.00	4.67	3.06
5	1	1311902	56	1.15	0.57	1.23	1.00	3.74	3.06