Published online 24 January 2016

Nucleic Acids Research, 2016, Vol. 44, No. 9 ***e81***
doi: 10.1093/nar/gkw031

Quantification of read species behavior within whole genome sequencing of cancer genomes for the stratification and visualization of genomic variation

Dror Hibsh[1], Kenneth H. Buetow[2], Gur Yaari[3] and Sol Efroni[1,*]

[1]Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel, [2]Computational Sciences and Informatics Program, Complex Adaptive Systems Initiative, Arizona State University, Tempe AZ 85281, USA and [3]Faculty of Engineering, Bar-Ilan University, Ramat Gan 52900, Israel

Received May 21, 2015; Revised January 5, 2016; Accepted January 11, 2016

ABSTRACT

The cancer genome is abnormal genome, and the ability to monitor its sequence had undergone a technological revolution. Yet prognosis and diagnosis remain an expert-based decision, with only limited abilities to provide machine-based decisions. We introduce a heterogeneity-based method for stratifying and visualizing whole-genome sequencing (WGS) reads. This method uses the heterogeneity within WGS reads to markedly reduce the dimensionality of next-generation sequencing data; it is available through the tool HiBS (Heterogeneity-Based Subclassification) that allows cancer sample classification. We validated HiBS using >200 WGS samples from nine different cancer types from The Cancer Genome Atlas (TCGA). With HiBS, we show progress with two WGS related issues: (i) differentiation between normal (NB) and tumor (TP) samples based solely on the information structure of their WGS data, and (ii) identification of specific regions of chromosomal amplification/deletion and their association with tumor stage. By comparing results to those obtained through available WGS analyses tools, we demonstrate some of the novelties obtained by the approach implemented in HiBS and also show nearly perfect normal/tumor classification, used to identify known and unknown chromosomal aberrations. Finally, the HiBS index has been associated with breast cancer tumor stage.

INTRODUCTION

Tumorigenesis involves a series of complex cellular, genetic and epigenetic changes (1,2). Large-scale cancer genomics projects, such as TCGA (3) and the International Cancer Genome Consortium (ICGC) (4), have worked hard in characterizing these changes in the cancer genome. Diagnosis and disease stage are primarily based on the histopathology of a tissue biopsy (5). Cancer genome sequencing has recently been introduced in the clinic, and provides another approach to assist clinicians in identifying genetics and epigenetic changes in tumor cells (6). Based on this progress, personalized therapeutic strategies are made possible (7,8). With rapidly falling costs of whole-genome sequencing (WGS), this technology is becoming an accessible tool in cancer research and patient care (9–11). As the technical barriers to human WGS are overcome, high-throughput 'omics' data accumulate and bring with them a set of issues that are novel in medical care. In cancer, other medical conditions and maintaining good health, improved analytic methods (12,13) and visualization (14,15) of omics data are of utmost importance.

Over the past two decades, many studies brought forward specific algorithms for quantifying whole-genome expression behavior (16–18), and other approaches as binary classification of normal versus tumor samples (19–23). Usually these approaches put their focus on the genome coding regions. Differing from whole exome sequencing (WES) and RNA sequencing, WGS approaches can detect mutations in unexplored regions and improve our understanding of the whole landscape of cancer genome. Elucidating the functions of these unexplored human genomic regions could facilitate the discovery of genomic biomarkers and personalized cancer treatment.

Most recent approaches have analyzed WGS to explore CNV (copy number variation) and depth of coverage (24–26). We now provide a new perspective on the use of WGS (Figure 1A; Supplemental Figure S1). A species volatility model (Supplemental Figure S2) is used to introduce a novel methodology for reducing the dimensionality of WGS. This method has been developed specifically for cancer, for the rapid, model-free, analyses of high-throughput data. Our approach translates a whole genome point-of-view (~6 giga bases) to its 1×10^{-7} fraction (589 regions) (Figure 1B), thus enabling a practical view of WGS data.

*To whom correspondence should be addressed. Tel: +972 3 738 4518; Fax: +972 3 738 4518; Email: sol.efroni@biu.ac.il

Figure 1. Experimental and computational workflow of HiBS. (**A**) The general approach to obtain a HiBS index: starting with a tissue sample, and following with NGS, the sequencing reads are mapped to a reference genome using one of the conventional alignment methods. The alignment is not used in HiBS classification, only in annotation, and thus has no impact over categorical decision-making. Finally, HiBS gives a species profile for each sample. A more detailed description of the full HiBS methodology is in (**B**), where we outline the partition of the genome ($\sim 3 \times 10^9$ bases) into equal length segments of 5×10^6 bases, and exclude chromosome tails (see 'Materials and Methods' section). This procedure leads to a total of 589 regions. Unique sequence species are detected within each region (**C**). The schematic example in this figure is a zoom-in on a case with 6 of 18 reads for this region that share an alignment to the same area in region number 350. These six reads yield four unique reads that we refer to as four species (s1-purple, s2-yellow, s3-green and s4-red). s1 is constructed from two identical reads. s2 is constructed from a single read that matched the beginning of s1 (purple nucleotides), but has additional nucleotides colored in yellow. s3 overlaps with both s1 (purple nucleotides) and s2 (yellow nucleotides), but also differs in nucleotides colored in green. s4 differs from s1 in a single nucleotide only that is colored in red. By counting the number of unique species per region, the species-per-region vector (**D**) is produced. The vector contains 589 points, each representing an individual region. The value for each region represents the number of unique reads in it. In (**E**), the last step in the calculation of the index, HiBS locates high volatility regions, based on the species-per-region vector, and produces a profile for each sample. The y-axis in (E) indicates the number of unique mapped reads (species) per region. The first x-axis outlines the genomic region {1..589}, set sequentially by the 5 million bases region size and the second x-axis displays chromosome numbers {1..22}. The red horizontal line indicates the average number of species and the green horizontal dashed line indicates the 1 z-score value. The blue line represents the species per region profile for the KIRP-NB sample (Kidney renal papillary cell carcinoma normal blood test; details in Supplementary Table S1). There are only six regions in which the number of species is above a z-score threshold of 1, which have been marked in blue circles (details in 'Materials and Methods' section).

We have focused on WGS to avoid many of the specific biases associated with the choice of an enrichment method of RNA (for RNA sequencing) or exomes (for exome sequencing) (27,28). We quantify heterogeneity within a sequencing sample and translate this sequence-read heterogeneity into clinical information (Figure 1C and D), which leads to importantly biological conclusions derived from examination of the data structures and their projection over genome architecture (Figure 1E).

We also implement the suggested approach for stratification and visualization of genomic variation in cancer, termed HiBS (Heterogeneity Based Subclassification). HiBS uses the species volatility-based model to stratify and visualize the whole genome. The input of HiBS is a single 'bam' file from one sample usually obtained either from the tumor or non-tumor tissue of a patient. As explained above, the method requires only data from WGS. The output of HiBS is the species volatility index vector (one value for each region), and gives basic statistics, such as z-scores and standard deviation, over the genome. These parameters are used to provide interpretable estimates of the source of a sample—normal versus tumor tissue (diagnosis), specific tumor targeting for amplification/deletion loci and associate to tumor stage (prognosis).

MATERIALS AND METHODS

Case study on TCGA datasets

We tested 203 TCGA WGS datasets (Table 1), the sequencing data being derived from a total of 107 patients. Ninety-six patients had paired NB (blood-derived normal) and TP (primary solid tumor) samples. The rest of the patients provided only TP samples. All datasets are of DNA reads that have been sequenced on Illumina instrument in paired-end mode, with 101/100 bases, at the Washington University Genome Sequencing Center (WUGSC) and the Baylor College of Medicine (BCM) centers. The reference genome used for the analysis of these datasets is the GRCh37-lite (29). Diversity within the datasets was achieved through the inclusion of different cancer types: COAD (colon adenocarcinoma), READ (rectum adenocarcinoma), KIRC (kidney renal clear cell carcinoma), UCEC (Uterine corpus endometrioid carcinoma), HNSC (Head and neck squamous cell carcinoma), CESC cervical squamous cell carcinoma and endocervical adenocarcinoma), KIRP (kidney renal papillary cell carcinoma), SARC (sarcoma) and BRCA (breast invasive carcinoma) using TCGA notations. While each of the types above represented by a single patient (two samples each), the BRCA set included 99 patients (187 samples in total) (Supplementary Tables S1 and S2). Throughout the analysis, BRCA samples were separated from the other eight cancer types. We refer to the eight other cancer types as 'mixture' cancer datasets which were mainly used for learning and optimizing the algorithm, and the BRCA as a tested dataset.

Quality control

Although the files deposited in the TCGA have to pass very restricted quality control tests, to rule out the option that our analysis is biased to the quality of the samples, we used FastQC v0.11.3 (30). FastQC use the bam file as input and provided a set of analyses that gave the impression of the quality of the file. Polymerase chain reaction (PCR) duplicates analysis was also used on the mixture set, based on the samtools 'pcrdup' function (31) (Supplementary Table S3). The available Nanodrop results from the TCGA were used to assess the quality of the samples (Supplementary Table S4).

Alignment

No changes were made from the original deposited TCGA bam files because the choice of the specific alignment tool was not critical; here this step is used only to map HiBS results to meaningful chromosomal regions, and it is highly robust. This alignment step is not part of the classification, but only for annotation. Usually, the deposited data in the TCGA was aligned using BWA (32) and Bowtie (33).

Description of HiBS

HiBS provides a command-line, stand-alone tool implemented in Linux. The method has two parts, first part being the core component, the HiBS full-mode command line tool, which is an efficient bash script for processing bam files, generating summary statistics on the mapped reads and providing a suggested classification for the origin of the sample as either normal or tumor. The second part uses the output of the first part to produce profiles of chromosomes volatility.

The HiBS algorithm

HiBS uses samtools (31) 'view –h' option to measure chromosome length, and this information divides the genome into regions of equal length. Next, using the samtools 'idxstats' option, HiBS calculates the sub-sampling factor, which is in turn used to obtain a fixed number of mapped reads across differing files. HiBS can also be executed without this option when handling a single file and there is no need for a comparison between parameters. The default value is set to 900 million mapped reads, based on the tested data (Supplemental Figure S3), but can be changed from the command-line. Although HiBS proceeds by dividing the genome into equal length regions (Figure 1B; Supplemental Figure S4), the exception is around the tails of each chromosome when it is not of divisible length.

Dissimilarity between regions

The algorithm continues by counting the number of unique mapped reads, referred to as '*species*', of every region and by registering the number of species in a region into a single representing vector (Figure 1C and D). By counting the species, HiBS basically measures the dissimilarity between the regions {1..589}. HiBS starts with Linux sort function and keeps with Linux uniq function. The sort function is used to sort the DNA reads in lexicographic order for each region {1..589}. At this point before using the uniq function, HiBS counts the number of instances that belong to each species, and keeps this value for structure distribution

Table 1. TCGA dataset used in our statistics

	BRCA dataset	Mixture dataset	Together
Sum	23.6TB	2.6TB	26.2TB
Mean	126.3GB	164.3GB	129.3GB
Min	73.6GB	90.2GB	73.6GB
Max	196GB	266GB	266GB

The table provides the statistics per dataset plus the combination of the two datasets. The parameters measured are in the first column. Summation is of the sizes of all the samples in the tested data. The Mean is of the size of the samples. Min is the minimum size of a sample, and Max its maximum size.

analysis (see below). After being sorted by lexicographic order, HiBS uses the uniq function to compare each of the reads to the next read, and keeps only the DNA reads that are dissimilar from each other. This can vary between single nucleotide dissimilarity and more. This results in a set of sequences that HiBS gets as an input.

z-score

The average and standard deviation of this vector are reported, and with these values at hand, HiBS calculates a z-score for each region and collects these values into a whole genome representation vector (Figure 1E) that serves as a volatility measure for each genomic region. To calculate a z-score, HiBS has estimate a few parameters, the first being the unique reads per region {1..589}. HiBS then calculates the average and the standard deviation of the unique reads over the genome. With the average and the standard deviation, it finally calculates the z-score for each region. The z-score is defined by:

$$\frac{x - \acute{x}}{\sigma} = z$$

where x is the number of unique reads in the specific region, $\acute{x}$ is the average unique reads over the genome and σ is the standard deviation of the unique reads over the genome. While the z-score is usually used to compare a sample to the standard normal deviation, HiBS uses it for finding the datum distance from the average, without statistical significant factors and without any assumptions of normality. This distance would not benefit the score, which is a second-order calculation over these distances.

HiBS index

For the final step of HiBS, we must know the parameters that were calculated in the previous steps as the z-score for each region, and the average and standard deviation over the genome. We then examined whether the z-score for each region is more than a single standard deviation from the average. If it does, this region is referred to as a *'high volatility region'*. After finishing the whole genome regions, we sum the amount of *'high volatility regions'*, and refer to this as the *'HiBS index'*.

HiBS optimization step

Careful estimation showed that 5 million bases region is the most suitable for the algorithm (Supplemental Figures S5 and S6). An additional optimization step in HiBS before the selection of z-score threshold to be 1 is seen in Supplemental Figure S7 for z = −1.5/ −1/ −0.5/ 0.5/ 1 and 1.5. The program options allow users to specify how strictly the analysis has to be after some initial defaults. HiBS provides a classification as to whether a sample is tumor or normal based on the HiBS index. In training sets, we found HiBS index threshold to be 10 regions. A score below or equal to 10 is classified as normal, with >10 being classified as tumor (Supplemental Figures S7 and S8). HiBS also provides a chromosome profile vector that represents the volatility regions among the chromosomes {1..22}.

Structure distribution analysis

While HiBS generates the index, it also keeps the number of instances belonging to each species (as explained above). This information is used to count how many species share the same order of instances. In other words, HiBS counts how many species the specific region has that share the same number of instances. From these numbers, we can generate the species distribution structure, a more detailed example being shown in Supplemental Figure S12.

Benchmarking

Benchmarking involved three tools other than HiBS, from which we could compare the profiles they generated to explore the differences between HiBS and these others. The first tool used for the benchmarking was Bedtools (v2.17.0–137-g83ce948) (34), which tests the genome coverage/depth-of-coverage (a schematic example is given in Supplemental Figure S11A). Bedtools was executed with the function 'genomecov –d', which reports the depth at each genome position with 1-based coordinates. As HiBS works on a 5M base window size, we calculated the average for each of the genomic regions in such a window to produce a genome coverage representing a vector in size of 589 as for Bedtools. Bedtools used a BAM format as its input. The second tool was Control-FREEC (v7.2) (25) for examining copy number changes and allelic imbalances using deep-sequencing data (a schematic example is shown in Supplemental Figure S11B). To achieve the best detection of copy number changes for Control-FREEC, we used it with two genomes at a time (normal versus tumor). Control-FREEC uses the normal genome as a reference. The parameters that Control-FREEC executed with were: window size of 5 Million bases, ploidy equal to two and the input was BAM format with paired-end samples. Control-FREEC uses samtools v1.2. The third tool used was CNV-seq (v2014.08.12) (35). This was used as a second method to detect CNV using high-throughput sequencing. To achieve the best detec-

tion of CNV for CNV-seq, we used it also with two genomes at a time (normal versus tumor), the normal genome being the reference. The parameters that CNV-seq was used with were: window size of 5Million bases and human genome reference, with the input being BAM format with paired-end samples. CNV-seq uses samtools v1.0. For the benchmark, we focused on HNSC (a tumor sample from the previously described mixture dataset), in which we studied chromosomes {1..22}. Since chromosome sizes differ and do not divide by 5M, we omitted the tail of the chromosomes from the benchmark, and analyzed all other regions.

RESULTS

Quality control

All samples passed FastQC basic statistics for quality estimation, per base sequence quality, per tile sequence quality, per sequence quality score, per base N content, length distribution, over-represented sequences and adapter content. PCR duplicates analysis showed that, even after the removal of the PCR duplicates, each of the samples has >750M reads (Supplementary Table S3). Available information on the quality of the samples from Nanodrop analysis indicated that all the samples have >1.89 (a260_a280_ratio) (Supplementary Table S4).

Inter-tumor WGS total mapped reads analysis

Analysis of the total mapped reads for the 'mixture' cancer datasets showed that each of the 16 samples has variable amounts of total mapped reads (Supplemental Figure S3). The smallest number related to the rectum adenocarcinoma normal sample (READ-NB) with 0.927 billion total mapped reads. The largest number of mapped reads belonged to the cervical squamous cell carcinoma tumor sample (CESC-TP) with 1.943 billion. There were samples in which the total mapped amount was larger in the tumor sample than the non-tumor sample, whereas in other samples the opposite was true. In five of the eight cancer types (CESC, HNSC, KIRP, SARC and UCEC), the tumor samples had nearly double number of total mapped reads than the normal samples. In two cases (COAD and READ), the gap in mapped reads was very small, whereas for the kidney renal clear cell carcinoma the normal sample (KIRC-NB) was with more total mapped reads compared to the tumor sample.

Region size impacts on the relative average and standard deviation of the species profile

An initial and necessary step in translating the genome to a reduced number of regions for HiBS to analyze is the choice of the number of regions to cover the genome. The total number of bases in chromosomes {1..22}, according to GRCh37/hg19 is 2 881 933 286 (29). We studied a range of sizes for the examined regions of 5k, 50k, 500k, 5m and 50m. The genome was divided into different total number of regions: 576 216, 57 633, 5776, 589 and 67, respectively (Supplemental Figure S4). For the 'mixture' cancer datasets, the kidney renal papillary cell carcinoma normal sample (KIRP-NB) had the largest average of species across the genome, whereas the smallest average belonged to the sarcoma tumor sample (SARC-TP) (Supplemental Figure S5). We also established that the choice of region size did not affect the relative average of species. This absence of effect applies to the additional samples (Supplemental Figure S5). The sample with the largest average remains as such for all choices of region sizes. Region size also did not influence relative standard deviation (Supplemental Figure S6). The only difference was with the KIRP samples, which had higher standard deviations in the tumor sample for 5k, 50k and 500k compared to the normal sample, but with the 5M and 50M region sizes the normal sample had a lower standard deviation (Supplemental Figure S6D and E).

Region size impact over HiBS classification within the mixture cancer datasets

An increase of region size (5k–50m) was associated with a unique behavior of the HiBS index. For region sizes 5k and 50k, the index gave seemingly random results, without any association to the source (normal/tumor) (Supplemental Figure S7A and B). Yet the transition to a region size of 500k gave an association with tumor status, such that tumor sample constantly had a higher HiBS index than its paired normal sample (Supplemental Figure S7C). However, this pattern was not consistent between cancer types. For example, SARC-TP's HiBS index is 52, but the HiBS index is 55 for UCEC-NB. For region size 50M we could not identify regions above the z-score at all (Supplemental Figure S7E). Finally, we show that the choice of a 5m region size provides a perfect HiBS classification. With this choice for the rest of the analyses, we found a clear separation between the indices of normal samples and tumor samples. In the former, the HiBS index did not exceed 10 across all the sample types (Supplemental Figure S7D), but in index for tumor samples was always equal to or larger than 11 (Supplemental Figure S7D).

Z-score threshold impact on HiBS classification of the mixture cancer datasets

Following the identification of the optimal region size, optimization was done by controlling for a z-score threshold within a range of values of −0.5, −1, −0.5, 0.5 to 1.5 to test if a choice of threshold would provide better stratification over a default threshold of 1 (although the value demonstrated a perfect classification). For the negative values (−1.5, −1 and −0.5), it would only be possible to classify parts of the data (Supplemental Figure S8). For a z-score threshold of −1.5, there was a close to perfect classification, with two misclassifications for the SARC-TP and the KIRP-TP. Classification was significantly diminished for thresholds of −1 and −0.5,. With positive z-scores both normal and tumor samples shared very similar indices for 1.5, but there were two misclassifications for UCEC-NB and SARC-NB as tumor for a value of 0.5 (Supplemental Figure S8). For a plot of the number of species per region as presented in Figure 2, we also found that genomic profiles of normal samples were very different from those of tumor samples. Normal samples show similar profiles across genomic regions, across patients, and even along

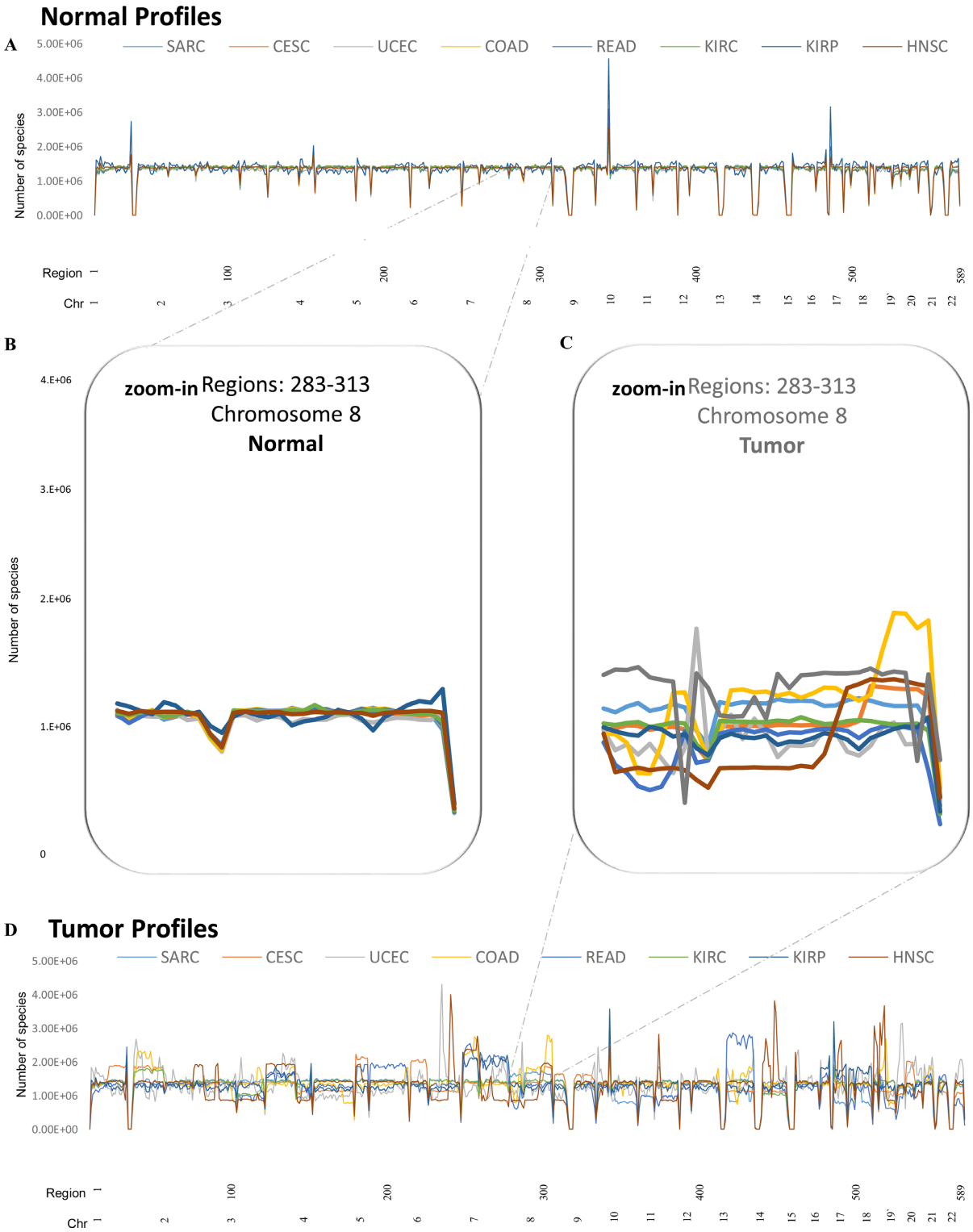

Figure 2. Major differences in volatility measurements between normal and tumor samples within the 'mixture' cancer datasets. The profiles come from eight blood-derived normal (NB) samples (**A** and **B**) and eight primary solid tumor (TP) samples (**C** and **D**), i.e. from eight patients with paired samples in total. The samples were obtained from TCGA ('Materials and Methods' section). (A and D) give the full genomic species profiles of samples from the same eight patients. The color of each line represents a different cancer type (indicated in the figure legend) consistently throughout panels (further details can be found in Supplementary Table S1). The axes are those above in Figure 1E. (B) This is a zoom-in over specific regions, 283–313, that cover chromosome 8 within the eight normal samples, whereas the same regions are shown in (C) for the set of tumor samples. In the normal sample set (B), all eight samples show similar species profiles across genomic regions. They are consistent across patients and even change simultaneously with acute curve changes. The stretches of zero-reads are heterochromatin regions seen in Figures 1E, 2A , D and 3B, C. The set of tumor samples in (D) clearly demonstrates that these samples do not share any common behavior, in contrast to the normal samples.

acute changes in function, such as heterochromatin areas (Figure 2A). In contrast, tumor samples do not share any common behavior (Figure 2D), and a closer look at specific regions, for example in chromosome 8 (Figure 2B and C), shows even larger differences.

Depth-of-coverage and the influence on HiBS index

To test the stability of HiBS index and rule out depth-of-coverage as an underlying parameter influencing the HiBS index, correlations between the total number of mapped reads (without subsampling), HiBS index and depth-of-coverage were measured. There was a perfect correlation between total mapped reads and depth-of-coverage, but a lack of any correlations between HiBS index and the others (Supplemental Figure S3). We also ran HiBS on the 'mixture' cancer dataset without the limitation of 0.9 billion mapped reads per samples, for which it also achieved perfect classification (Supplemental Figure S9).

Sample classification for breast cancer

HiBS also gives nearly perfect classification for the 187 breast cancer samples (Figure 3A; Supplementary Table S2), with a 98.4% succession rate. The species-per-region profile is given in Supplemental Figure S10. Only three samples were misclassified by HiBS (Supplemental Figure S10C and D). By examining the variability of species per region and per chromosome, these measures of variability provide a clear view of the acute differences between normal genomes and tumor genomes (Figure 3B and C; Supplemental Figure S10). For chromosomes 1, 3, 5, 6, 8, 10 and 16, the variety of species per chromosome proved much larger in the tumor samples than normal samples (Figure 3B and C).

Specific chromosomal arm amplification

In chromosomes 5, 10, 12 and 16, the species distribution was wider compared to other chromosomes, especially for their shorter arms (Figure 4A). In contrast to this set of chromosomes, in chromosomes 1, 3, 8, 13, 14 and 15 it was the longer arm that had a wider species distribution (Figure 4B). For chromosomes 2, 4, 6, 7, 9, 11 and 17–22, the difference in the species distributions was as significant.

High volatility regions

HiBS identified 499 out of the 589 possible genomic regions as high volatility regions in all 99 BRCA tumor samples, with an average of 53 regions per sample and 275 regions in the tumor samples of the 'mixture' cancer datasets with an average of 60 (Supplementary Tables S5 and S6). However, only 57 regions were tagged as high volatility regions across all 88 BRCA normal samples, with an average of 5 regions per sample and 7 regions in normal samples of the 'mixture' cancer set, with an average of 5 regions (Supplementary Tables S7 and S8). Tumor samples in the BRCA dataset and the 'mixture' dataset had 269 overlapping regions, and the BRCA normal dataset and the 'mixture' cancer datasets normal samples overlapped in 7 regions. Chromosomes 1 and 8 had the largest effect on high volatility regions in the BRCA tumor datasets (Supplementary Table S6). More specifically, regions chr8:125 000 000–130 000 000 (8q24.13–q24.21) and chr1:150 000 000–155 000 000 (q21.2-q21.3) were 'hotspots'. These were the genomic hotspots identified among the samples with values 66 and 62, respectively, for their HiBS indices for the 99 patients (Figure 4C; Supplementary Table S6). Regions chr10:40 000 000–45 000 000 and chr16:30 000 000–35 000 000 have higher volatility regions, but are also present in normal samples. In addition to an intersection between these findings and data from the genome browser refFlat (35), we identified that 8q24.13–q24.21 contained 38 unique genes, including the oncogene, *MYC* (36).

Tumor stage

Two groups are readily observed from the chromosomes profiles in Figure 4C. The first comprises chromosomes 4, 13, 14, 15, 18, 21 and 22, in which the number of volatility regions among the samples is low (0–2). The second group comprises all other chromosomes in which the number of volatility regions among samples is >2. When associating tumor stages with volatility quantification, stage I samples were found to present very little over-the-threshold volatility measurements in chromosomes 7, 11 and 20 (Figure 4C and D). In contrast, summing up samples of stages II and III gave a significantly larger number of over-the-threshold volatility measurements (Figure 4D). For stage II, there were almost no over-the-threshold volatility-measurements for chromosome 9, and for stage III the same applied to chromosome 2.

Benchmarking

Comparing the alterations of chromosomes {1..22} for HNSC on which HiBS had reported, to the other tools validated results for these regions. (Figure 5 and Supplemental Figures S13 and 14). For example, all tools showed loss of p13.3-p12 for chromosome 1 (regions 23–24; Figure 5A), and all tools showed a gain event in the longest arm of chromosome 8 (Figure 5C). We also found that the four profiles were very similar to one another because the interpretation of the results are similar across chromosomes (Figure 5 and Supplemental Figure S13 and 14). By focusing on the differences between the four profiles for chromosome 8, two of the tools, Bedtools and Control-FREEC, show almost no differences. These tools were identical not only in their plots, but also in the raw numbers that produced the plots, and differed only in the third place after the floating point (Supplemental Figure S13), while HiBS and CNV-seq shared some differences from the others. By measuring a Euclidian distance between the tool profiles, we found that for chromosomes 1–4, 6–9, 11–13, 15–16, 18 and 20–22 HiBS provides the largest distance from other tools (Supplemental Figure S14). For chromosomes 5, 10, 14, 17 and 19 CNV-seq stands at a greater distance (Supplemental Figure S14). When analyzing the actual distances between the tools we see that there are chromosomes for which the tools share between $\sim$12% (chromosome 10) and $\sim$25% (chromosome 14) of differences (Supplemental Figure S14). In contrast to the similarities between the tools, and specific for chromosome

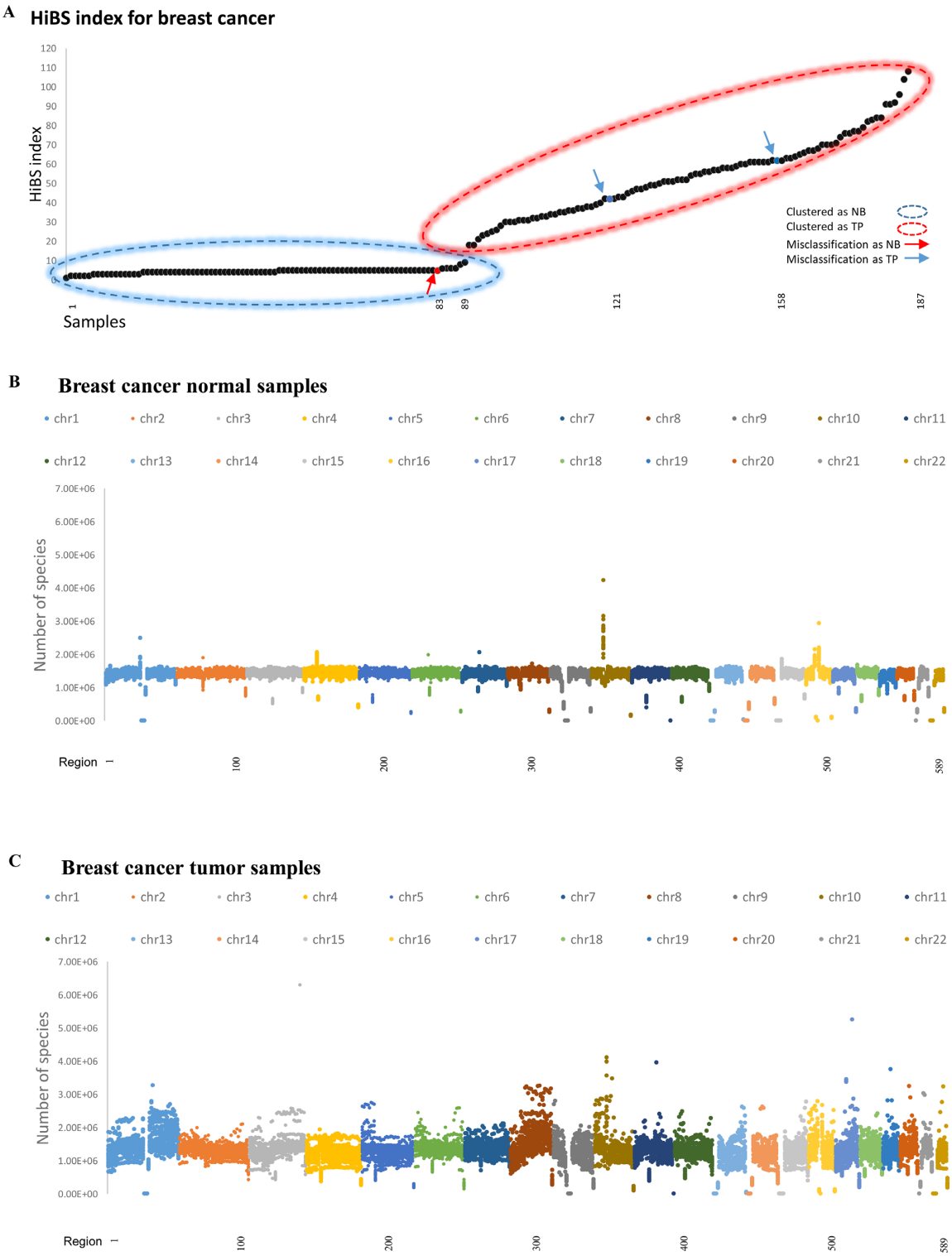

Figure 3. A whole genome point-of-view of breast cancer using HiBS classification. (**A**) Provides HiBS indices for a collection of 187 samples (breast cancer normal and tumor samples). The x-axis gives a sample's serial number, ordered by the HiBS classification. The y-axis represents the HiBS index. The first 89 points were classified as normal samples by HiBS and are followed by 98 samples classified as tumor. Samples defined as NB by TCGA that are also classified by HiBS as normal ('true negatives') are surrounded with a blue dashed circle; samples defined by TCGA as TP and also classified by HiBS as tumor ('true positives') are surrounded with a red dashed circle. The figure represents three misclassifications, points 121 and 158, where HiBS classified the sample as tumor whereas TCGA defined it as normal ('false positives', blue points), and one sample, point 83, in which HiBS classified the sample as normal whereas TCGA defined it as tumor ('false negative', red point). The profiles presented in (**B** and **C**) come from the 187 breast cancer samples that were analyzed from TCGA. The x-axes in (B and C) provide region numbers, whereas the y-axes provide the number of species per region. Panels (B and C) show the full range of variability between species profiles of normal tissue (B), and tumor samples (C). Each dot of color belongs to a different chromosome in correspondence to the figure legend. The full profiles of these datasets can also be seen in Supplement Figure S10.

Figure 4. Chromosome arm-specific amplification and tumor stages as portrayed by the HiBS index. The profiles come from the same set of 20 breast cancer patients as in Figure 3. Panel (**A**) gives chromosome profiles for chromosomes 5, 10, 12 and 16. Panel (**B**) gives chromosome profiles for chromosomes 1, 3, 8, 13, 14 and 15. Each of these chromosome profiles has a schematic chromosome under its profile, which were taken from the HapMap project (51) (http://hapmap.ncbi.nlm.nih.gov/karyogram/gwas.html). The axes in (A and B) are the same as in Figure 1E. Line colors correspond to different patients. Colors are consistent between panels (A and B). The proportion of each chromosome is kept and the y-axis is limited to 3.5×10^6 species per region. The matrix in (**C**) shows a connection between specific genomic regions and breast cancer tumor stage (generated by R gplots package 34). Each column stands for one chromosome {1..22}, and each row stands for the samples' cancer stage. The stages have been divided into categories I, II and III, separated by thick green lines. Each cell is color-coded according to the color key (middle right) that reflects the number of regions HiBS detected as high volatility in a specific chromosome. Panel (**D**) provides a sum-up of these measurements for chromosomes 7, 11, 20, exposing differences between the stages. Panel D provides the binary mode. We chose the highest value in the linear combination for stage I to be the cutoff for the binary classification.

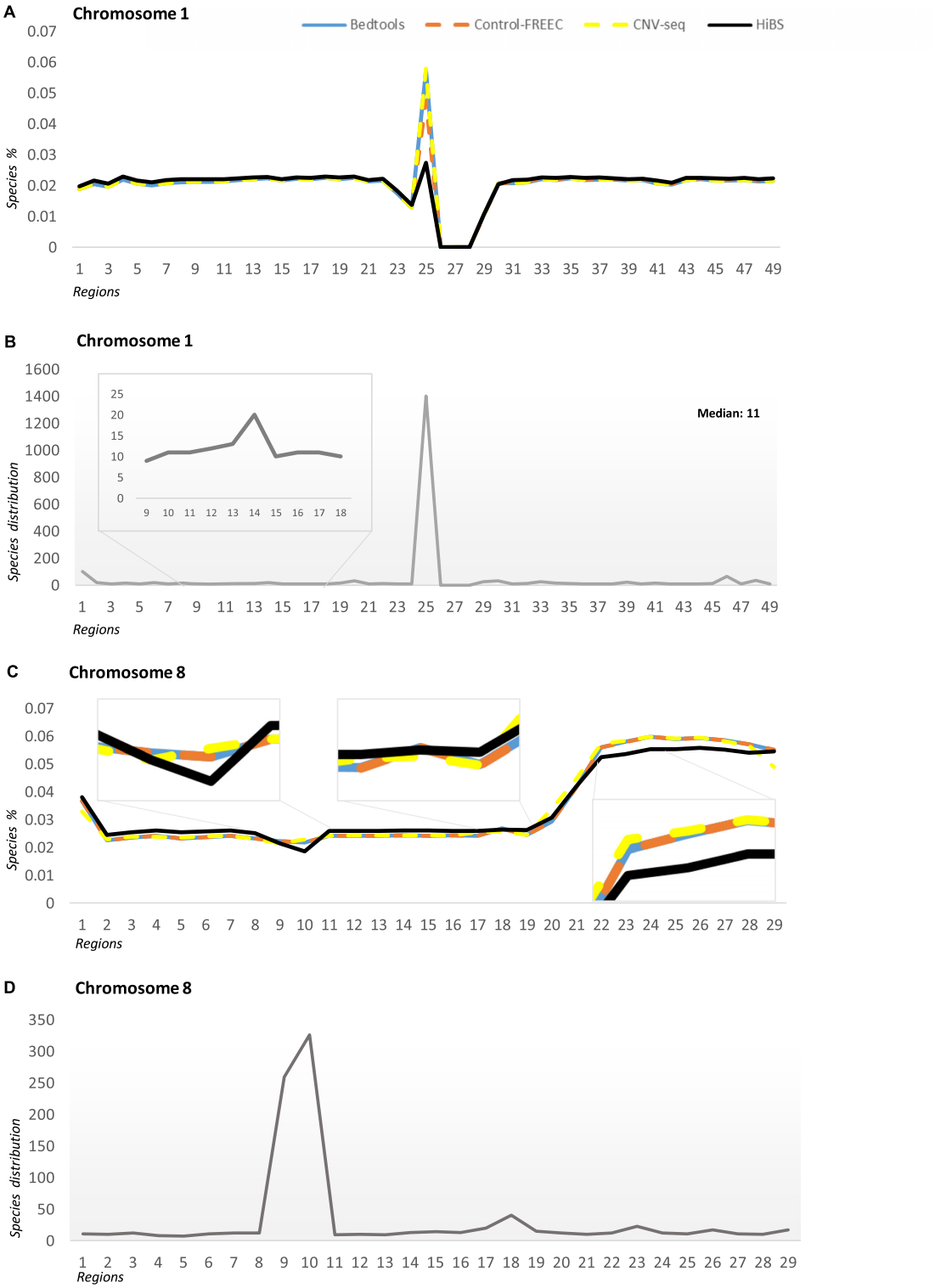

Figure 5. Benchmarking and species distribution structure. Specific cases for chromosomes 1 (**A** and **B**) and 8 (**C** and **D**) in HNSC (TCGA-BA-4076–01A-01D-2266–10_Illumina.bam) are in the panels. The x-axis is the region number across all panels. Since chromosomes length does not divide by 5M and since HiBS uses a window of 5 Mb, we neglected the tail (last region) and explored the other regions. The y-axis in panels (A and C) is the species fraction. The y-axis in panels (B and D) is the structure distribution number. Each tool has a different color and it is consistent along the figure. The gray rectangles in panels (B and D) are a zoom-in. The species analysis and the species distribution structure analysis, both generated by HiBS, and both for the same regions.

8, there were three specific areas in which the discrepancies between the profiles were higher (Figure 5C, gray rectangles); more data is given below regarding species distribution structure analysis.

Species distribution structure

One example for novel findings is chromosome 1 region number 14 (1p31.3-p31.1) (Figure 5A and B). This region has been picked up by HiBS structure analysis due to the fact that its structure distribution value is higher than the median for chromosome 1, while the other tools were unable to expose this region. We can see how this specific region differs from other regions in Figure 5B (gray rectangle). One of the genes included in the region is GADD45A. GADD45A is a member of the GADD45 family of genes, which are known stress sensors (37). The regulation of the expression of this gene is extensively studied in many types of cancer, such as breast cancer (37), head and neck cancer (38), human cancer cells (39) and many other cases. Further, associating the benchmarking profiles for chromosomes 8 in the HNSC to species distribution structure analysis by HiBS showed three areas with major differences between the tools, the biggest variation being along the species distribution structure (Figure 5C and D). The first area was region number 10, which is the centromere area of chromosome 8. The second was region number 18 (8q21.2–8q21.3), known for its abnormality in HNSC (40) and other cancers, including cervical carcinoma (41) and breast cancer (42). The third was area regions 22–23 (8q23.2–8q23.3), those of CSMD3 noted for its abnormality in HNSC and several other cancers (43).

DISCUSSION

We have shown a method that relaxes the complexity of WGS raw reads to an index that allows sample classification, detects specific amplification/deletion loci and indicated the stage of a tumor sample. HiBS was developed as a novel algorithm that stratifies and visualizes WGS data from cancer samples by deriving information about intra-tumor WGS heterogeneity within a sample. The three major advantages of HiBS are: first, the read species based method helps in decisions regarding the tumor/normal source of the sample using only the structure of the data (Figures 1–3A). Second, the algorithm is an extremely sensitive method of identifying specific amplification/deletion loci, making it suitable for loci targeting approaches (Figures 3B and C, 4B and C, 5B). Third, the algorithm can be easily integrated within a tumor staging system to assist in diagnosis (Figure 4C and D). We also describe the use of HiBS to successfully confirm breast cancer hotspots within the genome. For example, chromosome 8, is known to contain breast cancer affiliated loci (44–46), with the more specific *MYC* hotspot region being in chromosome 8—*MYC* is amplified in breast cancer (36,47). Furthermore, comparing the sensitivity of the algorithm to other tools on a specific cases, such as chromosomes 1 and 8 in HNSC, we found specific regions (1p31.3–p31.1 and 8q21.2–8q21.3) (Figure 5) that were uniquely ascertained by HiBS species structure distribution analysis, known in the literature to be in HNSC (38,40) and other cancers (37,39,41–42). In summary, results from experiments over a broad range of WGS data, through appropriate statistical modeling using the sequence-read species behavior of the data themselves, provide by HiBS processing tumor classification at an extremely high success rate, specific chromosomal fluctuation regions and an association with tumor stage classification.

HiBS can initially configure runtime parameters over a set of 16 WGS samples obtained from eight patients representing a 'mixture' of cancer types from TCGA (Supplementary Table S1). These datasets represent a non-uniform technical population due to their inter-tumor WGS heterogeneity, in contrast to intra-tumor WGS heterogeneity (48) Figure 2). This phenomenon calls for a recalibration of the amount of mapped reads for quantitative benchmarking (Supplemental Figure S3). Selecting different region sizes and using a variety of z-score thresholds (Supplemental Figures S4–S7), we saw convergence into a set of optimal parameters for the data. Empirical comparisons to ground truth showed that HiBS, with a region size of 5 million bases and with a z-score threshold of 1, outperformed all other parameter choices on the basis of accuracy of classification determined by the 'validity' of volatility measurement (Supplemental Figures S7 and S8). With these parameters, HiBS achieved a perfect classification between normal and tumor samples. Performance remained consistent whether we analyzed each dataset using a fixed or the original size (Supplemental Figures S7D and S9). Stability of HiBS index was checked by ruling out any associating to depth-of-coverage (Supplemental Figure S3C and D).

To achieve and demonstrate a robust evaluation of HiBS, we used the complete set of available whole genome sequenced breast cancer samples from TCGA (Supplementary Table S2). This shows that HiBS achieved near-perfect classification. Out of the 187 BRCA samples, 184 were successfully classified (98.4%), only 3 misclassifications being found. This high success rate calls for a detailed study of these specific samples; unfortunately, more informative data, both from the clinic or on the misclassifications samples is unavailable. Using these data, we identified BRCA specific targets of genomic amplification, especially in chromosomes 8 and 1, particularly for their longer arms (Figure 3B and C). Both chromosomes seem to be involved in quantifiable genome anomalies in many breast tumors (44–45,49). Moreover, by intersecting our identified high volatility regions with external information about breast cancer from refFlat (version 3/8/2014) (35), the highest observed HiBS index was in 8q24.13–q24.21. This section contains 38 unique genes, including *MYC*, which is amplified in many cases of breast cancer (36,47). Out of 589 possible regions, 499 had high volatility regions at least in one region for the tumor BRCA samples. These findings are supported by spatial heterogeneity phenomena, in which sub-clonal populations of cancer cells exist across different topological regions of the tumor (50).

To understand more fully the use of the volatility model, we correlated chromosome profiles with tumor stage. For a compatible comparison between different stages of the BRCA samples, seven samples were selected from stages I, II and III (Supplementary Table S9). There were only seven stage I (whole genome sequenced) samples in TCGA, which

limited the analyses to a smaller number. Nevertheless there were significant differences between the regions prone to high volatility or no volatility (Figure 4C). For example, chromosomes 7, 11 and 20 have high volatility regions in samples from stages II and III, whereas samples from stage I mostly lacked volatility (Figure 4D). Although the index displayed a 'decrease' gradient from stage I to stage III when presented in Figure 4C, it still hard to claim that a single chromosome can differentiate the stages.

For sensitivity exploration, we compared the information derived by HiBS to other methodologies, such as CNV and genome coverage (Figure 5). The results from these methodologies are very similar, but we show how we can have a uniquely point of view on the species distribution structure by using the species model to shed more light on the WGS as an innovative field (Figure 5B and D). But we remain unsure as to why the species distribution structure in those regions varied from the other regions, but we do know that the points it raises share abnormal behavior.

Perhaps the strongest feature of our method is that the algorithm findings are only based on the distribution and similarity of reads within the (BAM/FASTQ) file. Interpretation of the findings calls for a biological model, for example, that associate specific regions with specific genes. Normalization or a reference (normal genome) is not needed for analyses, making for greater tool robustness. While the biological, technological and genomics reasoning behind this phenomenon remains to be elucidated, the findings call for special attention to the interplay between cancer genome structure and NGS. HiBS provides a means to monitor this interplay, which in and of itself leads to immediate implementation.

In conclusion, the HiBS method offers a novel layer of interpretation to genome sequencing data. We have addressed the urgent need for tools to interpret WGS data in the diagnosis/prognosis of cancer genomes. The tool is freely available, and its further development should make it apply to other forms of genomic and epigenomic data, such as detection of unstable genomic loci in various disease and non-disease situations.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors would like to thank Prof. Edith Heard for her useful comments and pointers. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at the project website [http://cancergenome.nih.gov/].

FUNDING

The research leading to these results has received funding from the European Union FP7-People Programme under the project MODICELL, grant agreement n° 285875.

Conflict of interest statement. None declared.

REFERENCES

1. Ryan,R.J.H. and Bernstein,B.E. (2012) Genetic events that shape the cancer epigenome. *Science*, **336**, 1513–1514.
2. Jones,P.A. and Baylin,S.B. (2007) The epigenomics of cancer. *Cell*, **128**, 683–692.
3. Collins,F.S. (2007) The Cancer Genome Atlas (TCGA). *Nature* , **458**, 719–724.
4. Cancer,T. (2011) International cancer genome consortium. *Cancer* , **2011**, 1–20.
5. Zerbino,D.D. (1994) Biopsy: its history, current and future outlook. *Lik. Sprava*, **3–4**, 1–9.
6. Ding,L., Ley,T.J., Larson,D.E., Miller,C.A., Koboldt,D.C., Welch,J.S., Ritchey,J.K., Young,M.A., Lamprecht,T., McLellan,M.D. *et al.* , (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**, 506–510.
7. Cordero,P. and Ashley,E.A. (2012) Whole-genome sequencing in personalized therapeutics. *Clin. Pharmacol. Ther.*, **91**, 1001–1009.
8. Roychowdhury,S., Iyer,M.K., Robinson,D.R., Lonigro,R.J., Wu,Y.-M., Cao,X., Kalyana-Sundaram,S., Sam,L., Balbin,O.A., Quist,M.J. *et al.* (2011) Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.*, **3**, 111ra121.
9. Wang,K., Yuen,S.T., Xu,J., Lee,S.P., Yan,H.H.N., Shi,S.T., Siu,H.C., Deng,S., Chu,K.M., Law,S. *et al.* (2014) Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.*, **46**, 573–582.
10. Foley,S.B., Rios,J.J., Mgbemena,V.E., Robinson,L.S., Hampel,H.L., Toland,A.E., Durham,L. and Ross,T.S. (2015) Use of whole genome sequencing for diagnosis and discovery in the cancer genetics clinic. *EBioMedicine*, **2**, 74–81.
11. Bainbridge,M.N., Wiszniewski,W., Murdock,D.R., Friedman,J., Gonzaga-Jauregui,C., Newsham,I., Reid,J.G., Fink,J.K., Morgan,M.B., Gingras,M.-C. *et al.* (2011) Whole-genome sequencing for optimized patient management. *Sci. Transl. Med.*, **3**, 87–102.
12. Haibe-Kains,B., El-Hachem,N., Birkbak,N.J., Jin,A.C., Beck,A.H., Aerts,H.J.W.L. and Quackenbush,J. (2013) Inconsistency in large pharmacogenomic studies. *Nature*, **504**, 389–393.
13. Dewey,F.E., Grove,M.E., Pan,C., Goldstein,B.A., Bernstein,J.A., Chaib,H., Merker,J.D., Goldfeder,R.L., Enns,G.M., David,S.P. *et al.* (2014) Clinical interpretation and implications of whole-genome sequencing. *JAMA*, **311**, 1035–1045.
14. Gehlenborg,N., O'Donoghue,S.I., Baliga,N.S., Goesmann,A., Hibbs,M.A., Kitano,H., Kohlbacher,O., Neuweger,H., Schneider,R., Tenenbaum,D. *et al.* (2010) Visualization of omics data for systems biology. *Nat. Methods*, **7**, S56–S68.
15. Pabinger,S., Dander,A., Fischer,M., Snajder,R., Sperk,M., Efremova,M., Krabichler,B., Speicher,M.R., Zschocke,J. and Trajanoski,Z. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.*, **15**, 256–278.
16. van 't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A. a M., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
17. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
18. Rhee,J.-K., Kim,K., Chae,H., Evans,J., Yan,P., Zhang,B.-T., Gray,J., Spellman,P., Huang,T.H.-M., Nephew,K.P. *et al.* (2013) Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic Acids Res.*, **41**, 8464–8474.
19. Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.
20. Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

21. Tian,R., Basu,M.K. and Capriotti,E. (2014) ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples. *Bioinformatics*, **30**, i572–i578.
22. Peng,S., Zeng,X., Li,X., Peng,X. and Chen,L. (2009) Multi-class cancer classification through gene expression profiles: microRNA versus mRNA. *J. Genet. Genomics*, **36**, 409–416.
23. Victo Sudha George,G. and Cyril Raj,V. (2015) Accurate and stable feature selection powered by iterative backward selection and cumulative ranking score of features. *Indian J. Sci. Technol.*, **8**, 11–17.
24. Xie,C. and Tammi,M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80–89.
25. Boeva,V., Popova,T., Bleakley,K., Chiche,P., Cappo,J., Schleiermacher,G., Janoueix-Lerosey,I., Delattre,O. and Barillot,E. (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**, 423–425.
26. Miller,C.A., Hampton,O., Coarfa,C. and Milosavljevic,A. (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*, **6**, e16327.
27. Lohr,J.G., Adalsteinsson,V. a, Cibulskis,K., Choudhury,A.D., Rosenberg,M., Cruz-Gordillo,P., Francis,J.M., Zhang,C.-Z., Shalek,A.K., Satija,R. *et al.* (2014) Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat. Biotechnol.*, **32**, 479–484.
28. Van Dijk,E.L., Jaszczyszyn,Y. and Thermes,C. (2014) Library preparation methods for next-generation sequencing: Tone down the bias. *Exp. Cell Res.*, **322**, 12–20.
29. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**.D48–D55.
30. Andrews,S. (2010) FastQC: a quality control tool for high throughput sequence data. babraham Bioinforma.
31. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
32. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
33. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, 1–10.
34. Warnes,G.R., Bolker,B., Bonebakker,L., Gentleman,R., Liaw,W.H.A., Lumley,T., Maechler,M., Magnusson,A., Moeller,S., Schwartz,M. *et al.* (2015) gplots: various R programming tools for plotting data. R Package version 2.17.0.
35. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,a. D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
36. Xu,J., Chen,Y. and Olopade,O.I. (2010) MYC and breast cancer. *Genes Cancer*, **1**, 629–640.
37. Tront,J.S., Willis,A., Huang,Y., Hoffman,B. and Liebermann,D.A. (2013) Gadd45a levels in human breast cancer are hormone receptor dependent. *J. Transl. Med.*, **11**, 131–137.
38. Zhang,X., Qu,X., Wang,C., Zhou,C., Liu,G., Wei,F. and Sun,S. (2011) Over-expression of Gadd45a enhances radiotherapy efficacy in human Tca8113 cell line. *Acta Pharmacol. Sin.*, **32**, 253–258.
39. Shan,Z., Li,G., Zhan,Q. and Li,D. (2012) Gadd45a inhibits cell migration and invasion by altering the global RNA expression. *Cancer Biol. Ther.*, **13**, 1112–1122.
40. Manvelyan,M., Cremer,F.W., Lancé,J., Kläs,R., Kelbova,C., Ramel,C., Reichenbach,H., Schmidt,C., Ewers,E., Kreskowski,K. *et al.* (2011) New cytogenetically visible copy number variant in region 8q21.2. *Mol. Cytogenet.*, **4**, 1–3.
41. Rao,P.H., Arias-Pulido,H., Lu,X.-Y., Harris,C.P., Vargas,H., Zhang,F.F., Narayan,G., Schneider,A., Terry,M.B. and Murty,V.V.V.S. (2004) Chromosomal amplifications, 3q gain and deletions of 2q33-q37 are the frequent genetic changes in cervical carcinoma. *BMC Cancer*, **4**, 5–14.
42. Shadeo,A. and Lam,W.L. (2006) Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res.*, **8**, R9–R23.
43. Ma,C., Quesnelle,K.M., Sparano,A., Rao,S., Park,M.S., Cohen,M.A., Wang,Y., Samanta,M., Kumar,M.S., Aziz,M.U. *et al.* (2009) Characterization CSMD1 in a large set of primary lung, head and neck, breast and skin cancer tissues. *Cancer Biol. Ther.*, **8**, 907–916.
44. Orsetti,B., Nugoli,M., Cervera,N., Lasorsa,L., Chuchana,P., Rougé,C., Ursule,L., Nguyen,C., Bibeau,F., Rodriguez,C. *et al.* (2006) Genetic profiling of chromosome 1 in breast cancer: mapping of regions of gains and losses and identification of candidate genes on 1q. *Br. J. Cancer*, **95**, 1439–1447.
45. Ghoussaini,M., Song,H., Koessler,T., Al Olama,A.A., Kote-Jarai,Z., Driver,K.E., Pooley,K.A., Ramus,S.J., Kjaer,S.K., Hogdall,E. *et al.* (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl. Cancer Inst.*, **100**, 962–966.
46. Pole,J.C.M., McCaughan,F., Newman,S., Howarth,K.D., Dear,P.H. and Edwards,P.A.W. (2011) Single-molecule analysis of genome rearrangements in cancer. *Nucleic Acids Res.*, **39**, e85.
47. Singhi,A.D., Cimino-Mathews,A., Jenkins,R.B., Lan,F., Fink,S.R., Nassar,H., Vang,R., Fetting,J.H., Hicks,J., Sukumar,S. *et al.* (2012) MYC gene amplification is often acquired in lethal distant breast cancer metastases of unamplified primary tumors. *Mod. Pathol.*, **25**, 378–387.
48. Marusyk,A. and Polyak,K. (2010) Tumor heterogeneity: Causes and consequences. *Biochim. Biophys. Acta Rev. Cancer*, **1805**, 105–117.
49. Lingle,W.L., Barrett,S.L., Negron,V.C., D'Assoro,A.B., Boeneman,K., Liu,W., Whitehead,C.M., Reynolds,C. and Salisbury,J.L. (2002) Centrosome amplification drives chromosomal instability in breast tumor development. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 1978–1983.
50. Martelotto,L.G., Ng,C.K., Piscuoglio,S., Weigelt,B. and Reis-Filho,J.S. (2014) Breast cancer intra-tumor heterogeneity. *Breast Cancer Res.*, **16**, R48–R59.
51. Gibbs,R.A., Belmont,J.W., Hardenbol,P., Willis,T.D. and Yu,F. (2003) The International HapMap Project. *Nature*, **426** 789–796.