Published in final edited form as:

Methods Mol Biol. 2009 ; 529: 37–49. doi:10.1007/978-1-59745-538-1_3.

Comparative Genomic Hybridization: microarray design and data interpretation

Richard Redon and Nigel P. Carter

Abstract

Microarray-based Comparative Genomic Hybridization (array-CGH) has been applied for a decade to screen for submicroscopic DNA gains and losses in tumor and constitutional DNA samples. This method has become increasingly flexible with the integration of new biological resources generated by genome sequencing projects. In this chapter, we describe alternative strategies for whole genome screening and high resolution breakpoint mapping of copy number changes by array-CGH, as well as tools available for accurate analysis of array-CGH experiments. Although most methods listed here have been designed for microarrays composed of large-insert clones, they can be adapted easily to other types of microarray platforms, such as those constructed from printed or synthesized oligonucleotides.

Keywords

probe design; clone selection; normalization; outlier detection; CNV calling; Comparative Genomic Hybridization; array-CGH

1. Introduction

Comparative Genomic Hybridization (CGH) was developed in the early nineties to screen for chromosomal deletions and duplications along whole genomes (1,2). Originally, CGH consisted of co-hybridizing one test and one reference labeled probe DNA onto metaphase chromosomes spread on glass slides, in the presence of Cot-1 DNA to suppress high repeat sequences (*see Chapter 17*). During the nineties, CGH on chromosomes was widely used by research laboratories, in particular to screen for chromosome numerical aberrations associated with the progression of solid tumors (3): chromosome analysis by G-banding was technically challenging with tumor cells, due to the frequency of highly rearranged karyotypes and to the difficulties of culturing cells *in vitro* to obtain good quality metaphase chromosomes.

However, although CGH became widely used in cancer research, it did not prove to be particularly valuable as a standard method in diagnostic laboratories for the analysis of genomic imbalance in patients with developmental disorders. This was due firstly to the poor spatial resolution of metaphase CGH which is limited to 5-15 Mb by the image acquisition of probe signals on metaphase spreads using fluorescence microscopy. Secondly, metaphase CGH is technically challenging requiring expertise for preparation of suitable metaphase chromosomes as well as image acquisition and analysis.

From the mid-nineties, the International Human Genome Sequencing Project released new information on the human genome sequence, which was derived from the construction and characterization of libraries composed of large-insert clones such as bacterial artificial chromosomes (BACs) (4). These resources allowed the CGH method to be modified such that metaphase chromosomes could be replaced by arrayed DNA fragments representing precise chromosome coordinates. This strategy was initially called matrix-CGH (6) but then

array-CGH (5), and it is this name that is now in common usage. The development of array-CGH improved significantly the potential of CGH for the analysis of small chromosomal imbalances. Initial arrays provided a more than ten-fold increase in resolution such that micro-rearrangements that were invisible previously on chromosome preparations became detectable. In addition, for the first time, deletion and duplication breakpoints could be localized directly on the human genome sequence assembly.

The large insert clones used for the first array-CGH applications - in particular BACs and fosmids - have since become widely available. This has facilitated the construction of microarrays covering the whole genome at increasingly higher resolution. However, the relatively large size of these clones (~170 kb for BACs, ~40 kb for fosmids) limits the ultimate resolution of these types of arrays. In the past couple of years, small-insert clones, PCR products and oligonucleotides have been developed for use in array-CGH (5,6) allowing a greater degree of flexibility and higher resolution (down to just a few base pairs) in the design of microarray experiments which can be tailored to the specific biological question. This chapter describes many critical factors that should be considered when designing new array-CGH experiments and discusses different possible strategies for data analysis. It focuses on microarrays composed of cloned DNA printed on slides, though some strategies and tools described here can also apply for the design of microarrays composed of printed or synthesized oligonucleotides.

2. Array-CGH design

2.1. Clone selection

The first step in array-CGH is the design or choice of the microarray to be used for interrogating test genomes. There are two common strategies: (i) the design or the selection of one microarray covering the whole genome in order to screen for every deletion or duplication in a given test genome compared to a reference DNA; (ii) the construction and use of one microarray targeted to one part of the genome only, such as one chromosome or one region.

The design of a whole genome microarray is dependent on the resources available to construct the array. Construction of arrays from large insert clones requires physical spotting of the clone DNA onto microscope slides which typically limits the number of elements on the array to less than 50,000. For this reason, many laboratories used BAC clones for whole genome coverage, because with an average length of 170kb coverage of the whole genome with overlapping clones requires approximately 30,000 BACs while it would require more than 120,000 fosmids (40kb in length). Covering the whole genome at tiling path resolution is an important investment in time and resources, which may not be suitable for many laboratories. For this reason, most BAC microarrays used for whole genome screening were composed of only approximately 3,000 clones. They cover the whole genome with clones regularly interspaced, each single clone positioned at an interval of approximately one megabase apart. Although this strategy is not efficient for the detection of copy number changes below 1-2 Mb in size, it has proved to be valuable for the screening of most large-scale deletions or duplications, such as those responsible for severe congenital anomalies.

Several sets of clones designed specifically for the construction of CGH microarrays are publicly available. The Wellcome Trust Sanger Institute has developed two sets of large-insert clones for the construction of microarrays covering the whole genome at 1-Mb and tiling path resolutions (1-Mb and 30k TPA sets, respectively). The coverage of the human genome by these two sets of clones can be visualized on the Ensembl browser (www.ensembl.org, see Fig. 1A) and clones are available through GeneService (www.geneservice.co.uk). Another selection of 32,000 overlapping BAC clones covering

the whole genome can be obtained from the BACPAC Resources Center at CHORI (bacpac.chori.org).

To design a microarray targeted to specific loci, there is a larger choice of clones which could be used, depending on the size of the genomic segments to cover and on the resolution which is required. While BAC clones are usually selected for the construction of whole-genome microarray, fosmid clones represent a good alternative for custom arrays. Overlapping fosmids provide better resolution than overlapping BACs (down to 10 kb in case of high redundancy in coverage versus approximately 50 kb) but can be prepared for spotting using the same protocols (see Chapter 16). The fosmid library WIBR-2 is particularly useful as it has been extensively characterized by end-sequencing: most clones from this library are precisely mapped on the human genome assembly and all read-pair positions can be visualized on the UCSC genome browser (genome.ucsc.edu, see Fig. 1B). Read-pair coordinates can be downloaded from the UCSC browser for further selection of the clones required to cover the regions of interest. All fosmids can be purchased at the BACPAC Resources Center (bacpac.chori.org).

For example, after selecting fosmids for the construction of a small custom microarray, we applied array-CGH for high-resolution breakpoint mapping of two deletions at 9q22.3, responsible for a syndrome involving mental retardation and overgrowth in two unrelated children (7). The result obtained for one child is shown in Fig. 2A. Further increase in array-CGH resolution can be achieved by selecting small-insert clones (1.5 to 4 kb, see Fig. 2B) or PCR products (less than 1 kb), which can be used to cover all exons of any gene of interest (5). Today, synthetic oligonucleotides have largely replaced these approaches to custom array construction. Several companies - such as Agilent Technologies, Inc. and NimbleGen Systems, Inc. - are now commercializing microarray platforms with custom oligonucleotide synthesis, which provides virtually unrestricted flexibility in the design of CGH.

2.2. Controls

The microarray design should always include a selection of control target sequences, which will be used to estimate the performance of the microarray as well as the quality of array-CGH hybridizations.

Some negative controls should be included to estimate the intensity of fluorescence resulting from the non-specific hybridization of genomic probes on the target DNA. For printed arrays, negative control spot positions commonly contain bacterial genomic DNA or DNA sequences from other species, such as *Drosophila*. After image acquisition and spot intensity quantification, the intensity of fluorescence on these negative controls should always be monitored and be extremely low when compared to the test intensities along the microarray.

It is also valuable if possible within the array design to include controls for the estimation of the dosage response on the array. For example, adding clones representing sequences on chromosome X can be used to estimate the ratio deviation due to the presence of one copy in a male test DNA compared to 2 copies in a reference female DNA. This strategy has been widely used to validate the performance of new microarray platforms (8,9).

In addition, it may be useful to include some normalization probes, in particular for microarrays covering only small region. Selecting a number of clones that are located in one or several regions of the genome unlikely to be variable in copy number in test and reference DNA samples can be critical for normalization steps (see Fig. 2). The control clones can either be located on a chromosome which is known to contain no gross anomaly or can cover genes which are known to be present in normal copy number in the test and the reference DNA. When working on copy number variations in humans, one common strategy

consists in selecting only clones located at chromosomal loci not reported to show variation in the literature (data available in the Database of Genomic Variants, projects.tcag.ca/variation).

At last, using one or a small group of clones that will be printed in replicate distributed regularly on the surface of the microarray can help in detecting problems of signal heterogeneity after hybridization and imaging. Furthermore, a control DNA sequence spotted in replicate along the array can be used to estimate and correct the spatial heterogeneity of log₂ratio values (see §3.1.3).

3. Array-CGH data analysis

This section describes different strategies for the analysis of CGH profiles on large-insert clone microarrays that have been constructed and hybridized as described in **Chapters 16 and 17**.

3.1. Post-processing

After hybridization and washes, microarray images are acquired on an array scanner. Test and reference fluorescence intensities are measured for each spot position, and test *versus* reference intensity ratios are calculated, usually after subtraction of local background fluorescence in each channel (see **Chapter 17, Note 6**).

3.1.1. Exclusion of poor quality data points—After ratio calculation, any data point which does not fulfil a series of quality criteria should be excluded from further analysis. The usual criteria are:

- a. The signal fluorescence intensity should be significantly higher than the background intensity, at least in one channel. The exclusion threshold varies depending on image acquisition and quantification systems. Commonly, every data point with signal intensity lower than twice the background intensity is rejected. The background intensity can be either the local background intensity on the slide or the median signal intensity calculated from all negative controls (bacterial DNA or cloned DNA from an unrelated species).
- b. If each target DNA sequence is printed in several copies (several spots), any clone with discordant replicate ratio values should be excluded from analysis.
- c. Some microarray analysis programs, such as BlueFuse (BlueGnome, Ltd), calculate scores estimating quality criteria for each spot on the array (shape, regularity, concordance between the 2 colour channels and/or signal vs. background ratio). These quality scores can be used to exclude poor quality spots.

3.1.2. Global normalization methods—After the exclusion step, intensity ratios are normalized to generate the final log₂ratio profiles. One usual way to normalise whole genome log₂ratio profiles is to subtract the median log₂ratio value for all clones (or all clones located on autosomes) from each individual log₂ratio (*as shown in Chapter 17 - Fig. 2B*). This method - called global median normalization - is suitable for experiments comparing the genomes of healthy individuals or in constitutional genetics (where there are only limited numbers of rearrangements along the genome). However, it may not perform as well with test samples showing many gross chromosomal rearrangements, such as DNA from cancer cells. In this particular situation, it may be preferable to normalize the log₂ratio profiles using the modal rather than the median value: the modal value can be estimated for example using the Kernel method, which is available through the R software environment (www.r-project.org). Finally, array-CGH profiles that cover only one or several

chromosomal loci of interest should be normalised by the median log₂ratio value from control clones previously selected for their location in other non-variable regions of the genome (*see §2.2 and Fig. 2*).

3.1.3. Other normalization methods—In addition to the global methods, additional normalisation steps can be implemented to correct for some technical biases that can occur during array-CGH.

A spatial normalisation can be required when hybridising large DNA microarrays using manual procedures: the hybridisation process can result in uneven hybridisation across the slide and lead to a gradient of log₂ratio values along and/or across the array. A simple method for spatial normalization involves dividing the microarray into sub-arrays (or blocks) - each of them containing a sufficient number of spots - and normalising each block individually. Contact printers using pins for spotting arrays often generate arrays already segmented into a number of blocks that can be convenient for this approach. Alternatively, control spots replicated across the slide that have been included during the array design (*as described in §2.2*) can be used for local normalisation of ratios. The log₂ratio gradient is then corrected by normalizing separately each block by the median value of all spots (or by the median value of all control replicates) from this same block. Note that the block median normalisation using all spots is valid only if all the clones have been randomly distributed along the array, with no use of their genome or chromosomal position to order them on the surface of the slide.

When analysing DNA samples from various sources by array-CGH, we have observed that log₂ratio values show sometimes some strong variations, resulting in a “wavy” profile (or “auto-correlation”). Although the experimental origin of this phenomenon is still unclear, the observed variations correlate with the GC content of the corresponding clone sequences. In consequence, to overcome this problem, we have introduced a GC correction, which consists of normalizing the log₂ratios of each clone using the content in GC percent of that clone (*see Fig. 3*). This last step has enabled us to generate useful microarray data from some DNAs with poor quality, by eliminating the wavy patterns often visible in array-CGH with these types of samples..

3.2. Automatic detection of DNA copy number changes

The last step of data analysis is the application of statistical methods for the automatic detection of significant copy number changes on log₂ratio profiles. Many strategies can be followed to detect copy number changes objectively. The choice of the best method depends on the type of DNA samples that needs to be analyzed.

The simplest method is the arbitrary definition of one fixed threshold to determine which log₂ratio values correspond to DNA gains or losses. Fixed thresholds were initially applied for the analysis of DNA samples from patients with constitutional anomalies, because only few data points were expected to be variable by array-CGH (10). Fixed thresholds can also be applied on profiles focused on particular loci and particularly for breakpoint mapping of previously identified chromosomal imbalances (7). However, because the same fixed threshold is arbitrarily defined for every array-CGH profile, independently from differences in experimental log₂ratio variability (“noise”), this simple method results in higher false discovery rates in “noisier” profiles.

One other simple strategy to avoid this problem consists in applying a threshold of significance which is proportional to the experimental variability. The variability can be roughly estimated with the standard deviation (SD) of all normalized log₂ratios (considering that the experimental variability should be distributed normally). This method has been

commonly applied by array-CGH studies in constitutional genetics, with thresholds equal to three or four times the SD of all (autosomal) clones (11).

With the development of larger microarrays covering the whole genome with up to 32,000 clones, the use of single thresholds has become inappropriate. In a normal distribution of 32,000 log₂ratio values, 96 will be expected to be above a 3x SD threshold just by chance (and 3 or 4 above a 4x SD threshold). To address this problem, we have developed a more elaborate algorithm, CNVfinder, which enables the automatic detection and delineation of copy number changes using whole genome BAC/PAC array-CGH profiles (12). CNVfinder is based on a robust estimation of the inherent log₂ratio variance and applies different thresholds for single-clone and multi-clone copy number changes. It also incorporates a post-processing step to obtain the most likely bounds of each copy number change (12). This algorithm, written in the Perl language, is freely available and can be downloaded at www.sanger.ac.uk/humgen/cnv/software.

CNVfinder is an efficient method to detect copy number variants on array-CGH profiles in constitutional genetics (Fig. 4A-B). However, this algorithm is not suitable for detecting gross chromosomal anomalies or to detect copy number changes in highly rearranged DNA samples, such as tumor genomes. Many other statistical methods have been developed recently and can be used for the detection of copy number alterations in tumor cells. Three of them, all freely available, are listed below.

- SW-array (13) applies the Smith-Waterman algorithm (14) to identify segments with deviating values within a log₂ratio profile. SW-array is available as a package in the R Project (cran.r-project.org/doc/packages/cgh.pdf).
- aCGH-smooth (15) is a heuristic method, which identifies potential breakpoints and smoothes the observed array CGH values between consecutive breakpoints to a suitable common value. aCGH-smooth can be downloaded at www.few.vu.nl/~vumarray.
- DNACopy is a package in Bioconductor (www.bioconductor.org), which detects segments with abnormal copy number on array-CGH profiles using a method called circular binary segmentation (16).

3.3. Classification of copy number variation (CNV) by cross-sample analysis

Automatic detection methods have been designed to report, from each array-CGH profile, a list of chromosomal segments showing copy number variation (CNV) in the test DNA compared to the reference, using the rest of the genome (or some control chromosomal regions) as the baseline. Projects involving array-CGH analysis for larger cohorts of individuals may require additional information, such as the exact frequency of each CNV in the test population or the presence of any differences between samples showing the same gain or the same loss compared to the single reference sample. Cross-sample comparison can provide such additional information by examining for each clone - or each group of consecutive clones - the distribution of log₂ratio values from all individuals (Fig. 4C). Cross-sample analysis has been successfully applied for a global description of copy number variation in the human genome (17) and will be more widely used for CNV association studies in a near future (18).

4. Conclusion

Since its development in the late nineties, array-CGH has been widely used by research laboratories for the detection of copy number changes in DNA samples from individuals with cancer and constitutional disease. It has quickly become a reference method for the

diagnosis of patients with severe developmental defects. This method has also been instrumental for the discovery of an unexpected level of DNA copy number variation in the human genome, between individuals from the general population. New array-CGH platforms dedicated to variable regions of the genome are now in development for CNV association studies on common diseases. Array-CGH will certainly become a standard method in human genetics in years to come: new improved strategies will be required for better statistical interpretation of array-CGH results.

Acknowledgments

This work was supported by the Wellcome Trust.

References

1. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*. 1992; 258:818–821. [PubMed: 1359641]
2. du Manoir S, Speicher MR, Joos S, Schrock E, Popp S, Dohner H, Kovacs G, Robert-Nicoud M, Lichter P, Cremer T. Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. *Human genetics*. 1993; 90:590–610. [PubMed: 8444465]
3. Gebhart E, Liehr T. Patterns of genomic imbalances in human solid tumors (Review). *International journal of oncology*. 2000; 16:383–399. [PubMed: 10639584]
4. Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, Kim UJ, Kuo WL, Olivier M, et al. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature*. 2001; 409:953–958. [PubMed: 11237021]
5. Dharni P, Coffey AJ, Abbs S, Vermeesch JR, Dumanski JP, Woodward KJ, Andrews RM, Langford C, Vetrie D. Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *American journal of human genetics*. 2005; 76:750–762. [PubMed: 15756638]
6. Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, Nair P, Brothman AR, Stallings RL. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer*. 2005; 44:305–319. [PubMed: 16075461]
7. Redon R, Baujat G, Sanlaville D, Le Merrer M, Vekemans M, Munnich A, Carter NP, Cormier-Daire V, Colleaux L. Interstitial 9q22.3 microdeletion: clinical and molecular characterisation of a newly recognised overgrowth syndrome. *Eur J Hum Genet*. 2006; 14:759–767. [PubMed: 16570072]
8. Pinkel D, Seagraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*. 1998; 20:207–211. [PubMed: 9771718]
9. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*. 1997; 20:399–407. [PubMed: 9408757]
10. Veltman JA, Schoenmakers EF, Eussen BH, Janssen I, Merckx G, van Cleef B, van Ravenswaaij CM, Brunner HG, Smeets D, van Kessel AG. High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am J Hum Genet*. 2002; 70:1269–1276. [PubMed: 11951177]
11. Shaw-Smith C, Redon R, Rickman L, Rio M, Willatt L, Fiegler H, Firth H, Sanlaville D, Winter R, Colleaux L, et al. Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J Med Genet*. 2004; 41:241–248. [PubMed: 15060094]
12. Fiegler H, Redon R, Andrews D, Scott C, Andrews R, Carder C, Clark R, Dovey O, Ellis P, Feuk L, et al. Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome research*. 2006; 16:1566–1574. [PubMed: 17122085]

13. Price TS, Regan R, Mott R, Hedman A, Honey B, Daniels RJ, Smith L, Greenfield A, Tiganescu A, Buckle V, et al. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic acids research*. 2005; 33:3455–3464. [PubMed: 15961730]
14. Smith TF, Waterman MS. Identification of common molecular subsequences. *Journal of molecular biology*. 1981; 147:195–197. [PubMed: 7265238]
15. Jong K, Marchiori E, Meijer G, Vaart AV, Ylstra B. Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics (Oxford, England)*. 2004; 20:3636–3637.
16. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics (Oxford, England)*. 2004; 5:557–572.
17. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. Global variation in copy number in the human genome. *Nature*. 2006; 444:444–454. [PubMed: 17122850]
18. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet*. 2007; 39:S37–42. [PubMed: 17597780]
19. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. Diet and the evolution of human amylase gene copy number variation. *Nature genetics*. 2007; 39:1256–1260. [PubMed: 17828263]

Fig. 1. Selection of large-insert clones for array-CGH using Genome Browsers

(A) The Ensembl browser (www.ensembl.org) enables the user to visualize many physical or biological annotations in the context of the genome sequence. The box displays the respective positions of genes (Ensembl annotation, top panel), clones from the Sanger 1 Mb set (middle panel) and clones from the 30k TPA set (bottom panel) between coordinates 95-100 Mb on human chromosome 9. Lists of clones from these two sets can be downloaded as delimited tables from the same website (select option “Graphical overview”).

(B) Part of the same interval (99-100 Mb), displayed on the UCSC Genome Browser (genome.ucsc.edu), one alternative to Ensembl. The bottom panel shows positions of clones from the 30k TPA set. The top panel displays the positions of many fosmids mapped by end-pair sequencing. Some of the fosmid clones can be selected by their chromosomal locations for high-resolution coverage of the locus by array-CGH.

Figure 2. High resolution breakpoint mapping by array-CGH

(A) Array CGH profiles at the proximal (left) and distal (right) breakpoints of a 9q22.3 deletion detected in a patient with overgrowth syndrome (7). The deletion was first detected with a microarray covering the whole genome at 1Mb resolution (positions of 1Mb clones are represented as large grey bars). One custom microarray composed of fosmids (represented as short black bars) was then constructed to cover the two breakpoints regions at tiling path resolution. CGH with the custom array refined the deletions breakpoints to intervals of less than 50kb. Note that the 1Mb array profile was normalized by a block median method, while the custom array was normalized by the median of log2ratios from 26 fosmids located on chromosome 18 and used as controls (9).

(B) Detailed views of the same deletion breakpoint intervals. Using a small custom microarray composed of small-insert clones (1.5 to 4 kb in length, represented as small grey bars), it was possible to map each deletion breakpoint at a resolution of less than 5kb. Long-range PCR amplification and sequencing confirmed that array-CGH applied with increasing resolution enables accurate mapping of deletion breakpoints. The actual breakpoints are shown below the profiles on the UCSC browser: the proximal breakpoint disrupts the first intron of the PHF2 gene while the distal breakpoint is distal to the NR4A3 gene.

Figure 3. Normalization of array-CGH log2ratio profile using clone GC content

(A) After image quantification, log2ratio calculation and block median normalization, log2 ratios are plotted against the GC contents for all clones from autosomes: there is an apparent linear correlation between GC content and log2ratio (left panel: the fitted linear model is represented by the grey line). The influence of clone GC content on log2ratios results in local variations - called “waves” or “auto-correlation” - on the genome profile (right panel: only chromosomes 1 to 22 are displayed): waves are obvious for example on chromosomal arms 1p, 4p and 9q (grey arrows).

(B) By normalizing the log2ratios using clone GC content, the linear correlation is eliminated (left panel: the fitted linear model, in grey, gives a perfectly horizontal line at $y=0$). As a result, no wave is visible anymore on the log2ratio profile (right panel). Note that the GC correction in this example was performed by applying the linear model function ‘lm(y~x)’ in the R language environment (the output corrected values are the ‘residuals’).

Figure 4. Alternatives strategies to investigate copy number variation by array-CGH

(A) Superposition of 30 normalized array-CGH profiles on chromosome 1, resulting from the comparisons of 30 human individuals from the general population with one single reference individual.

The profile superposition enables to visualize regions showing copy number variation (CNV) in one or many individuals.

(B) UCSC Browser display for the entire chromosome 1, showing all copy number changes detected automatically on each individual profile with the CNVfinder algorithm. The length of each bar represents the frequency of gains (top panel) and losses (bottom panel) at the corresponding clone positions. CNVfinder reports that the region 3 is lost in one individual, the region 1 is gained in 15 individuals and the region 2 is gained or lost in 15 and 5 samples respectively.

(C) Analysis the distributions of log2ratio values in all 30 individuals for regions number 1 (clone Chr1tp-7D2), 2 (clones Chr1tp-30C7 and Chr1tp-6D2) and 3 (clones Chr1tp-19D11 and Chr1tp-21G1). Applying an univariate model-based clustering method for Chr1tp-7D2 shows that log2ratio values can be classified in 3 different states: 6 samples are reported as deleted (black squares) and 9 as carrying duplicated (black triangles) in comparison to the 15 samples falling in the intermediary state (black circles). Bivariate model-based clustering using log2ratios for clones Chr1tp-30C7 and Chr1tp-6D2 shows high levels of variation between individuals for the region 2: Five distinct clusters are detected using results from both clones. They correspond to five distinct copy numbers of region 2. This locus, which contains the AMY1 gene family, shows a highly unusual extent of copy number differentiation between human populations (19). Bivariate model-based clustering for clones Chr1tp-19D11 and Chr1tp-21G1 confirms the deletion of region 3 for one single individual, as detected by CNVfinder. Normal mixture modeling and model-based clustering were performed using the MCLUST package in the R language environment (www.stat.washington.edu/mclust/).