Genomics 101 (2013) 86–93

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Review

Molecular methods for genotyping complex copy number polymorphisms

Stuart Cantsilieris [a,b], Paul N. Baird [a,1], Stefan J. White [b,*,1]

[a] Centre for Eye Research Australia, University of Melbourne, Royal Victorian Eye and Ear Hospital, East Melbourne, Victoria, Australia
[b] Centre for Reproduction and Development, Monash Institute of Medical Research, Monash University, Melbourne, Victoria, Australia

ARTICLE INFO

Article history:
Received 20 July 2012
Accepted 24 October 2012
Available online 30 October 2012

Keywords:
Copy number variation
Next generation sequencing
Paralogue ratio test
Multiplex ligation-dependent probe amplification
Quantitative PCR
Southern blotting
Multiplex amplifiable probe hybridization

ABSTRACT

Genome structural variation shows remarkable complexity with respect to copy number, sequence content and distribution. While the discovery of copy number polymorphisms (CNP) has increased exponentially in recent years, the transition from discovery to genotyping has proved challenging, particularly for CNPs embedded in complex regions of the genome. CNPs that are collectively common in the population and possess a dynamic range of copy numbers have proved the most difficult to genotype in association studies. This is in some part due to technical limitations of genotyping assays and the sequence properties of the genomic region being analyzed. Here we describe in detail the basis of a number of molecular techniques used to genotype complex CNPs, compare and contrast these approaches for determination of multi-allelic copy number, and discuss the potential application of these techniques in genetic studies.

Contents

1. Introduction

The human genome contains several levels of genetic variation, from single base changes to those affecting entire chromosomes. Copy number variants (CNVs) now operationally defined as deletions

* Corresponding author at: Centre for Reproduction and Development, Monash Institute of Medical Research, 27-31 Wright Street, Clayton, 3168, Victoria, Australia. Fax: +61 3 9594 7114.
E-mail address: stefan.white@monash.edu (S.J. White).
[1] Are to be recognized as joint senior authors.

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2012.10.004

and duplications >50 bp, can be rare (<1%) or common (>5%), large (>1Mp) or small (<500 bp), and di-allelic (0–3) or multi-allelic (>3 diploid copy numbers) [1]. CNVs that segregate at appreciable frequency in the population are termed copy number polymorphisms (CNPs), and those that show a dynamic range of diploid copy number (multi-allelic) will be the focus of this review. Multi-allelic CNPs are attractive candidates for disease association studies for several reasons. Many of these CNPs contain genes, and these genes appear to be over-represented in pathways associated with immunity and interaction with the environment [2]. Furthermore, there is evidence of stratification in human populations, indicating that such regions are positively selected and are of clinical relevance [3]. Finally, gene duplication followed by adaptive evolution can facilitate new gene function, resulting in changes of phenotype [4].

Several techniques have been described to measure CNVs in the human genome, however, with such diverse genetic properties, no single existing methodology has the scope for accurately genotyping all CNV classes. The dynamic range that exists within complex CNPs poses significant challenges for accurate genotyping. In principle, this likely reflects the greater quantitative differences when detecting deletion copy numbers compared to duplications or multi-allelic loci. Distinguishing four from five diploid copy numbers reproducibly, compared to that of one and two, is difficult using standard methodologies [5]. This clearly poses problems for genotyping in association studies.

Such observations have highlighted the substantial ascertainment bias towards the detection of deletion variants from high resolution genome wide studies (77%) [6], and the limitations of certain techniques for assessing these regions [7]. This is further complicated by the fact that CNVs are enriched (10-fold) for segmental duplications (SDs, defined as sequences >95% and >1 kb in length) making characterization of these regions difficult using current methodologies [8]. Whole Genome Sequencing (WGS) in combination with sequence read depth, allows the analysis of many complex regions of the genome that were excluded from conventional genome wide association studies (GWAS) [9]. Seminal work by Sudmant et al. identified nearly 1000 genes within these regions, ranging from 0 to 48 copies at 3 kb resolution. Such a result expands our knowledge of the "assayable" portion of the genome, but indicates that many highly duplicated regions are yet to be analyzed in studies of disease. It has also been demonstrated that a significant proportion of CNPs residing in SD's are not in Linkage Disequilibrium (LD) with nearby single nucleotide polymorphisms (SNPs) [8]. A study of 192 CNPs, showed that only 40% of those located in SDs had high correlation to nearby SNPs, in comparison to 70% of 892 CNPs in unique regions of the genome [8]. These findings illustrate that for a large number of CNPs, genotypes cannot be imputed through the use of tagSNPs and must be measured directly.

There is continued interest in assessing the relevance of multi-allelic CNPs in complex disease, however, in some cases technical difficulties have impeded the reproducibility of these associations [7,10]. In addition, techniques amenable to genotyping in large scale studies can suffer from poor resolution particularly assignment of integer copy number, meaning that a compromise must be reached between accuracy and cost (Table 1). In general the most accurate techniques are the most labor intensive. Given that large numbers of individuals are required for robust associations and batch effects have the potential to create bias in genotyping, it is necessary to understand the strengths and limitations of methodologies used to analyze multi-allelic CNPs.

1.1. Hybridization-based techniques

1.1.1. Fiber FISH

Fluorescence *in-situ* hybridization (FISH) is a visual technique, typically used to identify chromosomal abnormalities from metaphase or interphase spreads using fluorescent probes. The strength of FISH lies in the direct visualization of DNA copy number at the single cell level. Multi-allelic CNPs, however, can be more difficult to analyze, especially when attempting to resolve tandem duplications. A modified approach, known as Fiber FISH, possesses sufficient resolving power to analyze complex structural rearrangements. The principle of this technique involves the release and fixation of DNA molecules from interphase nuclei onto a slide, with the DNA stretched in a linear fashion through mechanical or gravitational force [11]. The DNA fibers can be hybridized with fluorochrome-labeled DNA probes, producing a characteristic "beads on a string" pattern which is easily distinguishable from background probe signals [12]. Visualization of multiple DNA targets can be achieved using multi-colored probes, which appear as barcode-like signal patterns in the presence of tandem duplications [12]. Simple and complex genomic rearrangements, as well as repetitive sequences, can be accurately resolved using Fiber FISH. A striking example of the ability to resolve complex multi-copy gene re-arrangements was demonstrated in a study of the salivary amylase (*AMY1*) gene [13]. Individuals with tandem duplications of >10 copies could be accurately resolved using Fiber FISH. Techniques which measure changes in diploid dosage have the potential to miss assign complex copy numbers as *de-novo* events in the offspring, purely due to the number of different combinations and inheritance patterns [2]. A key advantage to using Fiber FISH, is that it allows the determination of CN per allele which is important for studies of inheritance and disease [13]. Limitations of

Table 1
Methods to measure complex copy number polymorphisms.

	Fiber FISH	Southern Blot	PFGE	QPCR[b]	MAPH[b]	MLPA[b]	PRT[b]	SNP array	Array CGH[a]	NGS
Detection	Absolute copy number	Inferred absolute copy number/change from diploid dosage	Inferred absolute copy number	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Absolute copy number
Sample	Cells	2–5 μg DNA	2–5 μg DNA	5–10 ng DNA	0.5–1 μg DNA	100–200 ng DNA	10–20 ng DNA	0.5–1 μg DNA	0.5–1 μg DNA	1–2 μg DNA
Loci	Single	Single	Single	Single	>40	>40	Single	>2 million	>2 million	Genome-wide
Throughput	Low	Low	Low	High	High	High	High	High	High	Low/moderate
Minimum resolution	>1 kb[12]	>1 kb[14]	0.5–1 kb[18]	100 bp	100 bp	100 bp	100 bp	5–10 kb[47]	5–10 kb[47]	>1 kb[9]
Cost per sample	Low	Low	Low	Low	Low	Low	Low	Moderate	Moderate	High
Time to result	>24 h	2–3 days	2–3 days	4 h	>24 h	>24 h	4 h	>24 h	>24 h	2–3 days
Labor requirement	High	High	High	Low	Low	Low	Low	Moderate	Moderate	High

[a] High resolution Array CGH can achieve a minimum resolution of >500 bp[6].
[b] Minimum resolution is in general the length of a single probe.

Fiber FISH include a labor intensive workflow, low throughput and a high quality sample requirement. In addition, highly variable regions are difficult to interpret if there are overlapping signals. However, this remains the most accurate methodology for typing multi-allelic CNPs.

1.1.2. Southern blotting and Pulsed Field Gel Electrophoresis

Southern blotting is a powerful method for typing structural rearrangements. The principle of the technique relies upon fragmentation of DNA with a restriction endonuclease enzyme, followed by gel electrophoresis and transfer to a nylon membrane (blotting) [14]. Hybridization using a labeled DNA probe and exposure to film allows visualization of target regions [14]. As this technique encompasses both hybridization and electrophoresis steps, detection of structural rearrangements occurs through comparison of hybridization intensities between a normal control and unknown samples [15] and/or the creation of altered fragment sizes (Fig. 1).

Complex CNPs can be resolved using Southern Blot Hybridization. For example, it has been shown that Southern blot analysis can resolve copy number at the *FCGR3B* locus (commonly exists between 1 and 4 copies per diploid genome), through differential band intensities after normalization on a reference locus, with decreased or increased ratio indicative of CNV [16]. Disadvantages to this technique include labor intensive workflow and the requirement for large amounts of high quality genomic DNA. A particularly important caveat is that uneven transfer of DNA to the nylon membrane or incomplete washing of probes can result in misinterpretation of band intensities. Careful consideration of probe design and choosing restriction enzymes that create distinguishable length differences between fragments, can minimize the impact of this limitation.

Hybridization intensity alone is not ideal for determination of complex CNPs, as intensity differences can be subtle and copy numbers can potentially be mis-scored. Pulsed field gel electrophoresis (PFGE) in combination with Southern Blotting, is a powerful technique with respect to resolution (ability to resolve DNA sequences >12 Mb) and accuracy for measuring copy number (analysis of altered junction fragments) [17]. PFGE is useful for both restriction mapping and characterizing large genomic re-arrangements [18]. A limitation of conventional agarose gel techniques is that size resolution is dependent on migration of DNA molecules through a relatively small gel pore. Large DNA molecules which exist as random coils, must unravel to enable electrophoresis through a much smaller gel matrix [17]. This leads to size independent mobility and loss of resolution. To circumvent this limitation, PFGE includes periodic alteration of the electric field which produces continuous re-orientation of DNA molecules. This allows the resolution of large DNA fragments. The utility of this technique in the detection of chromosomal re-arrangements extends not only to deletions and duplications but also translocations, insertions and inversions which are difficult to detect by other methods beyond PFGE and FISH [15].

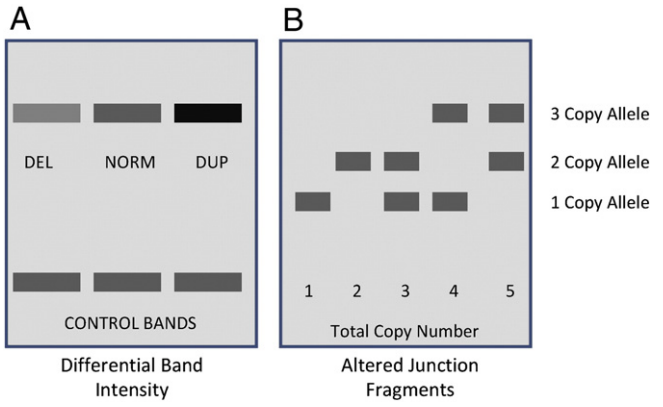

Fig. 1. Schematic showing the detection of copy number changes through the use of differential band intensities (Southern Blot) and altered junction fragments (Pulse Field Gel Electrophoresis). A) Gene dosage effects are clearly visible as differential hybridization intensities. Deletions bands are approximately half as intense as control bands, whilst duplications bands are 50% more intense than control bands. B) Copy number changes clearly visible as altered fragment sizes differentiated by length. Combining allele information allows inference of total gene copy number.

PFGE is useful for inferring absolute copy number of multi-allelic CNPs, as the presence of additional fragments corresponding to length differences of the repeat structure is a more accurate measure of the number of gene copies in comparison to intensity differences. For example, a study of the Fcγ receptors by Hollox et al. demonstrated that estimated copy numbers of 3, 4 and 5 by paralogue ratio test corresponded to additional fragments sizing at increasing 84 kb intervals, matching the length of the duplicon containing the Fcγ region [19]. Further work by Aldred et al., determined absolute copy number using PFGE on two α defensin genes, *DEFA1* and *DEFA3* [20]. *DEFA1* and *DEFA3* range between four and eleven copies in the general population, and *Hpa* I restriction fragments resulted in different length alleles corresponding to three, four and five copies of both genes [20]. A key advantage of PFGE is that all types of rearrangements are detectable as altered junction fragment sizes, thus comparison of hybridization intensities is not required. Limitations include the labor intensive workflow, the requirement of high molecular weight DNA that is not always available from archived samples, and the characterization of rearrangements being restriction enzyme site dependent.

1.3. PCR techniques

1.3.1. Quantitative PCR

Quantitative PCR is a high throughput technique for determining gene copy number. The basic principle behind qPCR is the measurement of PCR amplicon accumulation in real time. The fractional cycle number (Ct) is proportional to the amount of starting template, when the amplification during the exponential phase of the reaction reaches a defined threshold [21]. Amplicon accumulation is measured by fluorescent based chemistry, which primarily consists of either DNA intercalating dyes such as SYBER green or probe based methodologies such as TaqMan®, Scorpion and molecular beacons. The most popular approach is the TaqMan® primer and probe chemistry supplied by Applied Biosystems, which incorporates a duplex real time reaction for the gene of interest (GOI) and a reference gene (RG) (generally RNase P) for normalization. Data analysis is conducted using the comparative Ct method (ΔΔCt) where the Ct values of the GOI and RG are compared between a control sample and an unknown, with the equation presented as $2^{-\Delta\Delta Ct}$ [22]. CopyCaller™ software automatically calculates the equation presenting the predicted copy number and confidence in the call. An underlying requirement for using the ΔΔCt method is that the PCR amplification efficiencies for both the GOI and RG, should be similar as there is no correction for the difference in efficiencies in the equation [22]. Another commonly used approach to convert raw Ct values into normalized relative quantities is the standard curve method, which is used to calculate both PCR efficiency and interpret samples of unknown quantity [23]. Absolute or relative quantitation of an unknown sample can be achieved using a standard curve, which is constructed by amplifying known amounts of target DNA (usually a serial dilution), plotting the resultant Ct values as log concentrations and fitting a linear trend line to the data [23]. Either methodology is in principle suitable for analyzing CNP data. Studies of the *CCL3L1* gene, known to exist from 0 to 10 copies per diploid genome have been estimated using standard curve methodology [24]. Copy number was calculated by converting the Ct value into template quantity using the standard curve and determining the ratio of *CCL3L1* template quantity to reference gene β globin multiplied by 2. The copy number value was rounded to the nearest copy number integer [24].

QPCR has been the method of choice for large scale association studies, with the advantages of simple workflow procedure (3–4 h from sample preparation to result), high throughput capabilities and cost effectiveness in typing large sample numbers, and requiring relatively small amounts of DNA in comparison to many other methodologies for detecting CNVs (Table 1). Despite these advantages, there are significant limitations regarding the use of qPCR in genotyping multi-allelic CNPs. Several studies comparing Paralogue Ratio Test (PRT) and qPCR have found that, results gained by qPCR can be influenced by differential sample preparation, storage conditions and DNA degradation [7,25]. In addition, it has been shown that qPCR can generate results where there is extensive overlap between copy number integer classes [7,25]. Such a result would have dramatic affects on association studies. It is also worth noting that a study directly comparing qPCR and Multiplex Ligation-dependent Probe Amplification (MLPA) for analysis of beta-defensin yielded markedly different results using qPCR, often mis-scoring CN by several copies [26]. The effects observed by direct comparison between qPCR and other well established techniques have important implications for large scale association studies. It is clear that "Binned" or rounded data to the nearest integer, can result in spurious associations. The inability of qPCR to multiplex more than one locus for CNV analysis means that multiple measurements cannot be performed simultaneously, therefore many replicates maybe required to reach a consensus result.

1.3.2. Multiplex amplifiable probe hybridization (MAPH)

MAPH, first described by Armour and colleagues in 2000, is a technique based on quantitative recovery of probes after hybridization to immobilized DNA [27]. The amount of probe hybridization is proportional to the relative copy number of the binding site in the genomic DNA. MAPH probes were initially created by cloning small DNA fragments into vectors [27]. A less laborious approach involves the use of primer pairs containing identical 5′ sequences (MAPH sequences), to amplify target sequences from genomic DNA followed by pooling into a probe mix [28]. A limitation of this approach is that more than one PCR product per target locus could potentially be amplified during the creation of the probe mix, and as the probes have common ends, these extra products will be carried over in the amplification step [28].

The MAPH procedure utilizes genomic DNA immobilized to a nylon filter, overnight probe hybridization followed by a washing step to remove unbound probes. Bound probes are then released into solution by heating, and an aliquot is used for PCR. The original MAPH technique utilized radioactive labeling followed by polyacrylamide gel electrophoresis to visualize the PCR amplicons [27]. This was subsequently replaced by fluorescent labeling and separation by capillary electrophoresis [28]. Probes can be normalized by dividing the peak height (or area) of the probe by the heights of control peaks (loci known to be present in two copies), dividing the height of each peak by the sum of all peaks in a reaction, known as "global normalization", or by dividing the height of each peak by the sum of the four nearest probes in a mix known as the "nearest neighbor approach" [28]. MAPH has several advantages compared to other techniques. As the probe sequences are generally long (100–400 bp) and PCR is conducted directly from hybridized probes, MAPH is not affected by unsuspected polymorphisms under the probe binding sites. MAPH can generally multiplex up to 40 loci making it a cost effective and accurate method for typing many CNV loci simultaneously. Limitations include labor intensive probe preparation, DNA needing to be fixed to a membrane for stringent washing, and the fact that without a ligation step the specificity of the probes is dependent on the sequence properties of the genomic region.

MAPH has been shown to be effective for resolving several complex CNVs such as the beta-defensin region, which commonly exists between 2 and 9 copies per diploid genome (although as many as 12 has been observed) [29]. The beta-defensin region has a median copy number of four in European populations [25], and poses a considerable challenge for accurate genotyping. Studies of European control populations have shown that more than 80% of samples have 3–5 copies. The relative difference between the median copy number of 4 normalized to 1.0 is 0.75 for 3 copies and 1.25 for 5 copies. Distinguishing this difference reproducibly requires measurement of a relative increase of 0.25, which is not easily achieved using standard methodologies (Fig. 2) [5]. Multiple measurements indeed increase the precision of copy number calls and previous studies of beta-defensin have taken six independent MAPH measurements, calculated the average of each probe and rounded the mean to the nearest integer copy number [29,30]. When compared to other techniques such as FISH and PFGE [29] this appears to yield accurate copy number estimations.

1.3.3. Multiplex ligation-dependent probe amplification (MLPA)

MLPA is a versatile method for analyzing copy number variation at multiple loci (>40) from relatively low amounts of genomic DNA [31]. The technique relies upon hybridization and ligation of two adjacently situated oligonucleotides to a specific genomic DNA sequence. As with MAPH, all probes have identical 5′ sequences, allowing amplification with a single primer pair. Typically products are separated on a capillary sequencer based on size, and the resultant fluorescence intensities are exported for further analysis [28]. The sensitivity of the ligation step makes this method extremely powerful for analyzing sequences of high identity, by designing probes with mismatches at the ligation site. It is effective for measuring SDs, as notionally a probe can differentiate two sequences differing only by a single mismatch at the

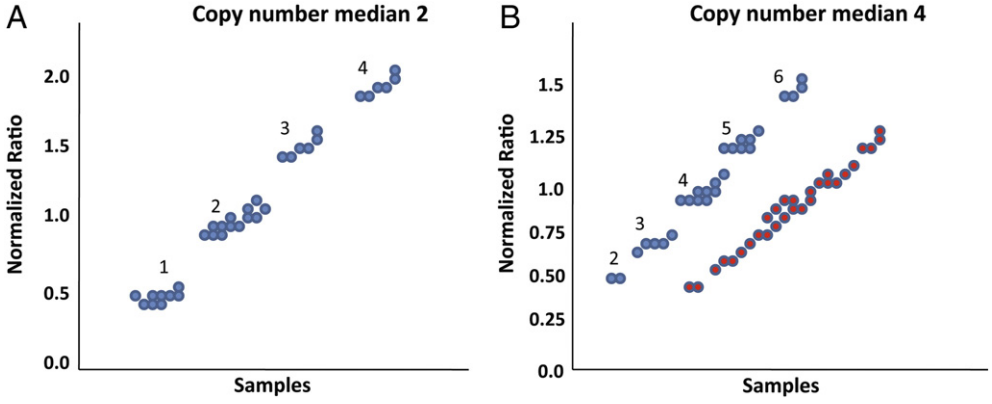

Fig. 2. Relative difference between integer copy number. The quantitative differences between 1 and 2 copies is greater than that of 4 and 5 copies (A) and (B). A) Stepwise increase in integer copy number from 0–4 copies. The relative difference between each of the integers is 0.5, consistent with an increase or decrease of 1 copy. B) Stepwise increase in integer copy number from 2–6 copies. A median copy number of 4 results in a relative difference of 0.25 between each of the integer copies (blue clusters). In practice this is not always achievable as experimental variability can result in extensive overlap of clusters obscuring accurate integer assignment (red clusters).

ligation site. One potential limitation is that sequence polymorphisms at and around the ligation site can disturb ligation sufficiently to give the appearance of an apparent mosaic deletion event (a limitation not shared by MAPH) [32].

Essentially, MLPA is analyzed in the same manner as MAPH (described above). Multi-allelic CNPs can be calculated by ordering the normalized ratios from low to high and assigning copy number integers defined by eye in a cluster analysis. By plotting these values in a scatter plot and calculating the relative differences between the subgroups (0.5 = 1 copy, 1.0 = 2 copies, and 1.5 = 3 copies) an estimation of DNA copy numbers can be performed (Fig. 2). For example, a study by White et al. performed MLPA analysis on the *NSF* gene, and was able to identify discrete copy number classes ranging from 2 to 7 copies with distinct differences between each of the subgroups [33]. This methodology has also been shown to be effective for the *FCGR3B* region. In addition to analyzing the clusters by eye, however, Marques et al. confirmed the cluster analysis using a statistical approach that validated the intensity ratios and cut-offs for assigning discrete copy number integers [34] Previous studies have also utilized the "nearest neighbor approach" for analyzing beta-defensin, where each test probe was normalized against the sum of five nearest neighbor reference peaks (CN of 2) [35]. An average of 10 beta-defensin test probes plotted on a scatter plot demonstrated discrete steps corresponding to a stepwise increase in copy number integers from 2 to 9 copies [35].

The strengths of MLPA lie in the number of loci that can be analyzed in a single reaction, the specificity of the ligation step, the reliability and accuracy of CNV measurement, and the relative low cost for conducting large scale association studies. Multiple steps are required to complete the MLPA procedure thus at several points there is the possibility for errors to be introduced.

1.3.4. Paralogue ratio test: PRT

One copy deletions and duplications can be detected using Quantitative Multiplex PCR methods, performed under semi quantitative conditions. Application to more complex CNPs has not yet been assessed, but presumably, the differential PCR amplification properties of the test and reference loci will impact the reliability of accurate copy number estimation. To mitigate the impact of experimental variability, Armour et al. reported the use of paralogue ratio test (PRT), a comparative PCR method based on amplification of dispersed repeat sequences [36]. The principle of this technique relies on careful primer design, in which two nearly identical sequences (differing in size) can be amplified using a single set of primers (a reference and test locus) [36]. The PCR is performed under quantitative conditions (30 cycles), and size differences are measured by capillary electrophoresis. Utilizing a two color dye system, multiple peak area measurements from the same sample can be compared to generate concordant data [37]. The ratio of test and reference loci allows an inference of gene copy number. This methodology is particularly suited for analysis of complex regions of the genome such as SDs. Inter-chromosomal repeat sequences are preferable as they are unlikely to be linked to the CNV locus. In recent times this technique has been widely adopted for analyzing multi-allelic CNPs including Beta-defensin, *CCL3L1*, *FCGR3B* and Complement Component 4 (*C4*) [7,19,37,38]. This has proven to be an accurate method of CNV detection [7,37]. One particular strength of PRT is that the first pass assignment of integer copy number is quite high at 85% for *CCL3L1*, 93% for *FCGR3B*, 93% for the beta-defensin locus, and 91% for *C4* [19,36–38]. This is particularly useful for association studies as repeat testing of 10% of samples is certainly within acceptable limits with respect to cost and labor. PRT has been shown to be superior to qPCR in terms of determining integer copy number and accuracy in copy number measurement >4 [7,25]. In addition, the workflow procedure is quite rapid as it does not require the lengthy overnight hybridization steps, that are components of both MLPA and MAPH. With respect to cost, it is a relatively inexpensive means to determine copy number in large scale association studies, and requires low quantities of genomic DNA (10–20 ng). Studies have shown comparable accuracy to other well validated methods such as MLPA and MAPH [36]. Limitations to this methodology relate to the fact that it is assumed that the paralogous reference loci do not itself vary in copy number, that sites of primer binding are free from polymorphisms that can affect primer binding efficiency, and the dependence of identifying paralogous sequences outside the target region. One technical limitation is that regions of high sequence identity, such as *FCGR3A*/*FCGR3B* and *C4A*/*C4B* do not contain the sequence differences required for primer specificity. To circumvent this limitation both genes are amplified and cut using restriction enzymes, with a digestion time >4 h [19,38] Another drawback is a lower multiplexing capacity compared to techniques such as MLPA and MAPH.

2. Array based approaches

2.1. Array CGH

Array CGH is a technique based on dual hybridization of test and reference DNA to either short or long oligonucleotides (and historically bacterial artificial chromosomes (BACs)) immobilized on a glass slide [39]. The signal ratio between test and reference sample is normalized and used to infer copy number. Initial approaches using BAC clones with a resolution between 100 and 200 kb provided important insights into the landscape of structure variation in the human genome [40], however, poor breakpoint resolution typically led to overestimation of CNV size. Subsequent studies implementing long oligonucleotide arrays have provided a more accurate picture of the CNV landscape, with improved resolution between 0.5 and 2 kb [6,41].

2.2. SNP microarrays

SNP microarrays are also hybridization based and have the advantage of analyzing both single nucleotide differences and in some cases non-polymorphic copy number probes that are not restricted by sequence properties of SNPs [41]. In comparison to Array CGH, SNP microarrays analyze a single sample per microarray and compare signal intensities from a sample with clustered intensities from a set of reference samples, or the whole sample population to generate a log ratio [41,42]. SNP microarrays generate two types of fluorescent information; the total fluorescence gained from intensity of both alleles, and the allelic ratio gained from the relative intensity of each allele. By plotting the normalized intensity of each allele against one another, three types of clusters should emerge corresponding to the AA, AB and BB genotypes. Results indicative of a deletion can result in three additional clusters corresponding to reduced intensity of homozygous alleles indicative of a heterozygous deletion, or no signal (failure to cluster) at all indicative of a homozygous event. The intensity of one allele as a proportion of the total allele signal, can also be used as an additional measure to confirm the presence of CNPs (Fig. 3) [42,43]. For example the proportion of the B allele should correspond to 0, 0.5 and 1.0 for genotypes AA, AB and BB respectively [42,43]. In this case heterozygous deletions are indicated by the loss of heterozygosity such that the BAF is 0 or 1 for AA and BB genotypes. Homozygous deletions result in no signal. More complex CNVs can also be indicated by the loss of the one to one ratio of each allele, such that there is a skew in BAF at heterozygous sites indicative of an AAAB, BBBA or AABB genotype (reflective of 4 copies). This is useful for accurately predicting CNPs of 0–4 diploid copy numbers.

2.3. Application of arrays to CNP genotyping

The success for resolving complex CNPs into discrete copy number classes is typically dependent on the genetic properties of the CNP, the probe density/performance and the normalization parameters [6]. A particularly important point is that the number of probes on an array does not necessarily translate into improved coverage or

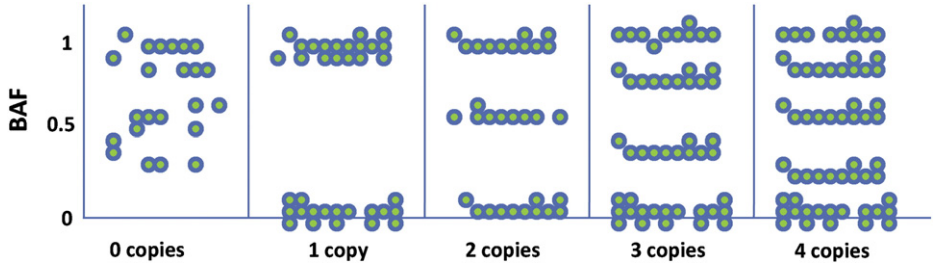

Fig. 3. The Allelic Ratio or B allele Frequency (BAF) to infer multi-allelic copy number. The BAF calculates the proportion of the B allele from the total allele signal, thereby using allelic ratio to infer gene copy number. For example the BAF for AA, AB and BB genotypes should be 0, 0.5 and 1.0 respectively. No signal is indicative of a homozygous deletion. One copy genotypes resulting in complete loss of heterozygosity signal indicated by BAF clustering around BAF = 0 (AA) and BAF = 1.0 (BB). More complex copy numbers such as three and four copy number genotypes can result in quite different BAF values, e.g. allelic ratios of 0.33/ 0.67 (ABB) and 0.25/ 0.75 (ABBB) respectively.

resolution. It is recognized that commercial SNP arrays such as Affymetrix and Illumina SNP arrays do not sufficiently cover regions of genome complexity such as SDs [8,44]. In addition, a study by Conrad et al. using customized tiling array CGH could only reliably genotype 61% of CNPs that map to SDs [6]. In general, SDs show higher levels of false positive and false negative call rates for both SNP and Array CGH platforms in comparison to unique regions of the genome [42,44]. However, both the Nimblegen and Agilent array CGH platforms contain substantially more probes in SD regions than do Affymetrix and Illumina SNP arrays [44]. Not surprisingly, large studies using customized array CGH platforms have found that duplications and multi-allelic loci are more difficult to detect than deletion variants [6,45]. This may be at least partially explained by an ascertainment bias for some commercial arrays towards deletion variants [44]. In particular, it is important to note that for multi-allelic CNPs, the choice of normalization algorithm impacts the resolution of the data [45] and that each individual locus must be treated separately to achieve the best results.

Numerous analytical algorithms have been described to analyze data from microarray platforms. However, different algorithms often provide marked differences in the both the number and reliability of CNV calls [44,46]. Studies have recommended that merging calls from multiple algorithms may improve sensitivity, in comparison to using multiple algorithms to increase confidence in CNV calling which can result in substantial increases in false negative call rates [44]. In addition, it has been shown that software developed specifically for the particular array performs better than software independent algorithms. For example, algorithms such as Birdsuite developed specifically for the Affymetrix 6.0 platform performed better in a direct comparison with platform-independent algorithms such as Nexus Copy Number [44].

The main strength of microarray platforms is the ability to screen CNVs on a genome-wide level at relatively low cost in large data sets, an advantage yet to be equaled by next generation sequencing platforms. In general, parameters for reliable genotyping are set at five consecutive probes and a minimum size of 1 kb [44]. Typically, most single channel array platforms lose sensitivity to detect variants below 10 kb [47]. There are, however, several limitations. The ability of commercial microarray platforms to resolve breakpoints and detect smaller rearrangements is generally poor, and is dependent on probe location and spacing (minimum 10 kb). No positional information is obtained, meaning that structural rearrangements that do not affect copy number (such as inversions and translocations) will not be detected. Estimation of copy number is relative, and it is assumed that the diploid copy number for a given region is two, which is not always the case (particularly in regions of segmental duplication). The reliability of Array CGH CNV calls is also heavily dependent on how well characterized the reference sample is. For example, a loss in the reference sample can be interpreted as a gain in the test sample even though the test sample may have a diploid copy number of two.

3. Sequencing based approaches

3.1. Next generation sequencing

The abundance of data generated from next generation sequencing (NGS) platforms has provided a wealth of information relating to the extent of structural variation in the human genome [48]. The application of a single comprehensive bioinformatic approach has remained elusive, due to the short read limitations of NGS technology. Several methodologies have recently emerged that are amenable to detecting specific types of structural variation, including split read methods, paired end mapping, read depth methods and whole genome assembly [49]. Of these computational approaches only read depth methods and whole genome assembly are suitable for genotyping complex CNPs and are discussed below.

3.2. Read depth methods

Read depth methodology is a useful measure for determining absolute copy number as essentially, the number of sequencing reads that map to a specific region is proportional to the number of copies that the region is present in the genome [49]. This basis for read depth methods assumes a Poisson distribution of sequencing reads, thus a region can be assumed to be deleted or duplicated if the region has fewer or more mapped reads than expected. This technique has been successfully applied to complex genomic regions containing multi-allelic CNPs [9]. Typically, smaller CNV events and those that contain high genomic copy number require deeper sequence read depth (SRD) to achieve accurate CNV measurement. Recent data by Hollox et al. found a reciprocal relationship between the size of the CNV, and the required SRD to accurately distinguish four from five copies. Interestingly, while increases in SRD initially improve the accuracy of CNV measurement (up to 50x coverage), much deeper SRD is required to achieve the same gains in resolution beyond this point [2]. One limitation of using massively parallel short-read sequencing is the inability to uniquely map short reads to regions such as SDs. Singly unique nucleotides (SUNs) provide a unique identifier within duplicated sequences and have been successfully applied to complex regions of the genome [9]. Using a combination of SUNs and SRD, copy number can be measured in most segmental duplications with short reads of >30 bp and $<5\times$ read depth. The accuracy of read depth approaches for determination of multi-allelic CNPs was supported by showing high correlation with qPCR and Array CGH experiments [9].

SRD approaches have recently been applied to exome sequencing data. [50]. A challenging aspect of this analysis is that, in comparison to whole genome sequence, exome sequence coverage is non-uniform and can be biased by factors such as sequence capture design. Encouragingly, analysis of three or more exons could estimate absolute copy number of multi-allelic genes with an accuracy of 78% with an enhanced ability to detected smaller CNVs >14 kb [50]. However, there

were several limitations. The study was based on exome capture kits which essentially target unique regions of the genome, pseudogenes can obscure accurate copy number prediction due to extensive sequence homology, normalization of the X chromosome means the algorithm is unsuitable for chromosome aneuploidy and discovery of novel polymorphic CNPs is less sensitive in comparison to genotyping known CNP regions.

An advantage in performing SRD analysis is being able to transform relative copy number obtained from array CGH data into absolute copy number [8,9,48]. Comparing absolute copy number predictions based on SRD, with data generated by SNP or Array CGH allows integer copy number to be calibrated and adjusted. A limitation of using intensity data is that inferred integer copy number is prone to inaccuracy, especially at high copy number counts. Campbell et al. was able to correct a number of regions that were discordant by one integer copy using SRD which gave an accurate assessment of baseline copy number [8]. This will have important implications for future disease association studies as we move from associating traits with a deviation from the population median copy number (low or high) to assessing individual integer copy number in a population.

3.3. Whole genome sequence assembly

Complete genome assembly compared to a high quality reference genome is currently the most accurate way for discovering structural variation. *De novo* assembly using NGS technology from a small number of genomes has already yielded between 5 and 50 Mb of novel sequence not identified in the NCBI reference assembly build 36 [51,52]. Analysis of 185 human genomes (as part of the 1000 genomes project) compared three sequence-based computational approaches (split read, paired end mapping and read depth) and found differences between the methods in terms of genomic regions ascertained, accessible structural variation size range and breakpoint precision [53]. Of the methodologies, read depth analysis and paired end mapping proved to be the most sensitive, and a newly developed method ((Genome STRiP) which integrates both approaches) proved to be the most accurate [53].

The strength of NGS technology lies in the ability to identify many different levels of genetic variation, from SNPs to small indels and large CNVs. Recent advances in computational approaches allow accurate estimation of multi-allelic CNPs that cannot be determined using conventional capillary sequencing. In addition, the ability to identify and characterize CNV break points will ultimately lead to a comprehensive map of CNVs in the human genome. Despite the obvious advantages in using NGS technology, there are several limitations. The major issue is that no currently available computational approaches are comprehensive. Both split read and paired end mapping approaches are unreliable in duplicated regions of the genome, and read depth method is poor at characterizing CNV breakpoints. In addition, algorithms to assist in *de novo* assembly will require further improvement before they can accurately map complex regions of the genome. The quest for the perfect reference genome, (which may never be possible) is further complicated by genes embedded in repeat rich regions such as SDs, and such regions are incomplete even in the latest genome reference build [2]. For instance, a recent study by Alkan et al. found that two recent *de novo* assemblies were 16.2% shorter than the reference genome, with several Mb of missing sequences, including a large proportion of repeat and duplicated sequence [54]. An integrated approach using longer read technologies and new assembly algorithms will substantially improve the standard of *de novo* assembly, especially in complex regions of the genome.

4. Future directions

Multi-allelic CNPs will contribute to a proportion of relative risk in a variety of human complex diseases, although the significance is likely to be dependent on ethnicity, clinical phenotype and interaction with other genomic loci. Accurate detection and measurement of multi-allelic CNPs including direct assessment of integer copy number, in large ethnically matched cohorts is required to fully understand their contribution to complex disease. We have described the advantages and disadvantages of a number of commonly used methodologies to assess complex CNPs. It is clear from this that no single method is yet, suitable for all types of analysis. It is noteworthy that in combination, these techniques can be much more powerful with respect to resolution and accuracy. For example, intensity data from single channel array CGH combined with SRD analysis can be used to calibrate multi-integer copy estimates [9]. Although Fiber FISH, Southern blotting and PFGE are low throughput methodologies, they are useful for accurately characterizing copy number reference samples to be used for downstream genotyping applications [38]. Confirming copy number estimates using two of the four PCR based methodologies (MLPA, MAPH, QPCR and PRT) can allow an estimate to be made of the reliability of the copy number typing [55]. The introduction of NGS technology, and the development of improved computational approaches such as read depth mapping using SUNs, have vastly improved the accuracy of determining absolute copy number of multi-allelic CNPs. Finally, the implementation of single cell sequencing will provide allele and inheritance information, which is currently not achieved for any CNV methodology, with the exception of Fiber FISH. Recent work in breast cancer tumors has demonstrated that high resolution copy number profiles can be achieved by flow sorting, whole genome amplification (WGA) and massively parallel sequencing [56]. The cost of performing WGS in large disease cohorts is not yet achievable for the average researcher. Targeted, cost effective and high throughput techniques such as PRT, MLPA, MAPH and qPCR will continue to be used for CNV genotyping in the near future.

Acknowledgments

This work is supported by funding from Monash University, Translational Clinical Research in Major Eye Diseases, and the National Health and Medical Research Council (Centre for Clinical Research Excellence #529923, NHMRC Senior Research Fellowship 1028444 (PNB)). Both MIMR and The Centre for Eye Research Australia are supported by the Victorian Government's Operational Infrastructure Support Program.

References

[1] S. Girirajan, C.D. Campbell, E.E. Eichler, Human copy number variation and complex genetic disease, Annu. Rev. Genet. 45 (2011) 203–226.
[2] E.J. Hollox, The challenges of studying complex and dynamic regions of the human genome, Methods Mol. Biol. 838 (2012) 187–207.
[3] K.M. Steinberg, F. Antonacci, P.H. Sudmant, J.M. Kidd, C.D. Campbell, L. Vives, M. Malig, L. Scheinfeldt, W. Beggs, M. Ibrahim, G. Lema, T.B. Nyambo, S.A. Omar, J.-M. Bodo, A. Froment, M.P. Donnelly, K.K. Kidd, S.A. Tishkoff, E.E. Eichler, Structural diversity and African origin of the 17q21.31 inversion polymorphism, Nat. Genet. 44 (2012) 872–880.
[4] M.Y. Dennis, X. Nuttle, P.H. Sudmant, F. Antonacci, T.A. Graves, M. Nefedov, J.A. Rosenfeld, S. Sajjadian, M. Malig, H. Kotkiewicz, C.J. Curry, S. Shafer, L.G. Shaffer, P.J. de Jong, R.K. Wilson, E.E. Eichler, Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication, Cell 149 (2012) 912–922.
[5] E.J. Hollox, Beta-defensins and Crohn's disease: confusion from counting copies, Am. J. Gastroenterol. 105 (2010) 360–362.
[6] D.F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y.J. Zhang, J. Aerts, T.D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C.H. Ihm, K. Kristiansson, D.G. MacArthur, J.R. MacDonald, I. Onyiah, A.W.C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N.P. Carter, C. Lee, S.W. Scherer, M.E. Hurles, Origins and functional impact of copy number variation in the human genome, Nature 464 (2010) 704–712.
[7] M.C. Aldhous, S. Abu Bakar, N.J. Prescott, R. Palla, K. Soo, J.C. Mansfield, C.G. Mathew, J. Satsangi, J.A.L. Armour, Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease, Hum. Mol. Genet. 19 (2010) 4930–4938.
[8] C.D. Campbell, N. Sampas, A. Tsalenko, P.H. Sudmant, J.M. Kidd, M. Malig, T.H. Vu, L. Vives, P. Tsang, L. Bruhn, E.E. Eichler, Population-genetic properties of differentiated human copy-number polymorphisms, Am. J. Hum. Genet. 88 (2011) 317–332.
[9] P.H. Sudmant, J.O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure, P. Genomes, E.E. Eichler, Diversity of human copy number variation and multicopy genes, Science 330 (2010) 641–646.

[10] S. Cantsilieris, S.J. White, Correlating multiallelic copy number polymorphisms with disease susceptibility, Hum. Mutat. (Jul 26 2012), http://dx.doi.org/10.1002/humu.22172.
[11] I. Parra, B. Windle, High resolution visual mapping of stretched DNA by fluorescent hybridization, Nat. Genet. 5 (1993) 17–21.
[12] J. Kraan, A.R.M. Bergh, K. Kleiverda, J.-W. Vaandrager, E.S. Jordanova, A.K. Raap, P.M. Kluin, E. Schuuring, Y.-S. Fan, Multicolor fiber FISH, Methods Mol. Biol. 204 (2003) 143–153.
[13] G.H. Perry, N.J. Dominy, K.G. Claw, A.S. Lee, H. Fiegler, R. Redon, J. Werner, F.A. Villanea, J.L. Mountain, R. Misra, N.P. Carter, C. Lee, A.C. Stone, Diet and the evolution of human amylase gene copy number variation, Nat. Genet. 39 (2007) 1256–1260.
[14] G. Mellars, K. Gomez, Mutation detection by Southern Blotting, Methods Mol. Biol. 688 (2011) 281–291.
[15] J.T. den Dunnen, G.-J.B. Ommen, C.G. Mathew, Application of pulsed-field gel electrophoresis to genetic diagnosis, Methods Mol. Biol. 9 (1992) 313–325.
[16] T.J. Aitman, R. Dong, T.J. Vyse, P.J. Norsworthy, M.D. Johnson, J. Smith, J. Mangion, C. Roberton-Lowe, A.J. Marshall, E. Petretto, M.D. Hodges, G. Bhangal, S.G. Patel, K. Sheehan-Rooney, M. Duda, P.R. Cook, D.J. Evans, J. Domin, J. Flint, J.J. Boyle, C.D. Pusey, H.T. Cook, Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans, Nature 439 (2006) 851–855.
[17] J. Herschleb, G. Ananiev, D.C. Schwartz, Pulsed-field gel electrophoresis, Nat. Protoc. 2 (2007) 677–684.
[18] J. den Dunnen, G. Van Ommen, Methods for pulsed-field gel electrophoresis, Appl. Biochem. Biotechnol. 38 (1993) 161–177.
[19] E.J. Hollox, J.-C. Detering, T. Dehnugara, An integrated approach for measuring copy number variation at the FCGR3 (CD16) locus, Hum. Mutat. 30 (2009) 477–484.
[20] P.M.R. Aldred, E.J. Hollox, J.A.L. Armour, Copy number polymorphism and expression level variation of the human beta-defensin genes DEFA1 and DEFA3, Hum. Mol. Genet. 14 (2005) 2045–2052.
[21] R. Hiquchi, C. Fockler, G. Dollinger, R. Watson, Kinetic PCR analysis: real time monitoring of DNA amplification reactions, Biotechnology 11 (1993) 1026–1030.
[22] T. Schmittgen, K. Livak, Analyzing real-time PCR data by the comparative C_t method, Nat. Protoc. 3 (2008) 1101.
[23] J. Hoebeeck, F. Speleman, J. Vandesompele, E. Hilario, J. Mackay, Real-time quantitative PCR as an alternative to Southern Blot or fluorescence insitu hybridization for detection of gene copy number changes, Methods Mol. Biol. 353 (2007) 205–226.
[24] E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R.J. Nibbs, B.I. Freedman, M.P. Quinones, M.J. Bamshad, The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility, Science 307 (2005) 1434–1440.
[25] P. Fode, C. Jespersgaard, R.J. Hardwick, H. Bogle, M. Theisen, D. Dodoo, M. Lenicek, L. Vitek, A. Vieira, J. Freitas, P.S. Andersen, E.J. Hollox, Determination of beta-defensin genomic copy number in different populations: a comparison of three methods, PLoS One 6 (2011) e16768.
[26] A. Perne, X. Zhang, L. Lehmann, M. Groth, F. Stuber, M. Book, Comparison of multiplex ligation-dependant probe amplification and real time PCR accuracy for gene copy number quantification using the beta defensin locus, Biotechniques 47 (2009) 1023–1028.
[27] J.A.L. Armour, C. Sismani, P.C. Patsalis, G. Cross, Measurement of locus copy number by hybridisation with amplifiable probes, Nucleic Acids Res. 28 (2000) 605–609.
[28] J.T. den Dunnen, S.J. White, MLPA and MAPH: sensitive detection of deletions and duplications, in: Curr Protoc Hum GenetJohn Wiley & Sons, Inc., 2006
[29] E.J. Hollox, J.A.L. Armour, J.C.K. Barber, Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster, Am. J. Hum. Genet. 73 (2003) 591–600.
[30] E. Hollox, J. Davies, U. Griesenbach, J. Burgess, E. Alton, J. Armour, Beta-defensin genomic copy number is not a modifier locus for cystic fibrosis, J. Negat. Results Biomed. 4 (2005) 9.
[31] J.P. Schouten, C.J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens, G. Pals, Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification, Nucleic Acids Res. 30 (2002) 1–13.
[32] A.J. Notini, J.M. Craig, S.J. White, Copy number variation and mosaicism, Cytogenet. Genome Res. 123 (2008) 270–277.
[33] S. White, L. Vissers, A. Geurts Van Kessel, R. De Menezes, E. Kalay, A. Lehesjoki, P. Giordano, E. Van De Vosse, M. Breuning, H. Brunner, J. Den Dunnen, J. Veltman, Variation of CNV distribution in five different ethnic populations, Cytogenet. Genome Res. 118 (2007) 19.
[34] R.B. Marques, M.M. Thabet, S.J. White, J.J. Houwing-Duistermaat, A.M. Bakker, G.J. Hendriks, A. Zhernakova, T.W. Huizinga, A.H. van der Helm-van Mil, R.E. Toes, Genetic variation of the Fc gamma receptor 3B gene and association with rheumatoid arthritis, PLoS One 5 (2010) e13173.
[35] M. Groth, K. Szafranski, S. Taudien, K. Huse, O. Mueller, P. Rosenstiel, A.O.H. Nygren, S. Schreiber, G. Birkenmeier, M. Platzer, High-resolution mapping of the 8p23.1 beta-defensin cluster reveals strictly concordant copy number variation of all genes, Hum. Mutat. 29 (2008) 1247–1254.
[36] J.A. Armour, R. Palla, P. Zeeuwen, M. den Heijer, J. Schalkwijk, E.J. Hollox, Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats, Nucleic Acids Res. 35 (2007) e19.
[37] D. Carpenter, S. Walker, N. Prescott, J. Schalkwijk, J. Armour, Accuracy and differential bias in copy number measurement of CCL3L1 in association studies with three auto-immune disorders, BMC Genomics 12 (2011) 418.
[38] M.M.A. Fernando, L. Boteva, D.L. Morris, B. Zhou, Y.L. Wu, M.-L. Lokki, C.Y. Yu, J.D. Rioux, E.J. Hollox, T.J. Vyse, Assessment of complement C4 gene copy number using the paralog ratio test, Hum. Mutat. 31 (2010) 866–874.
[39] N.P. Carter, Methods and strategies for analyzing copy number variation using DNA microarrays, Nat. Genet. 39 (7 Suppl) (2007) S16–S21.
[40] R. Redon, S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, Global variation in copy number in the human genome, Nature 444 (2006) 444–454.
[41] S.A. McCarroll, F.G. Kuruvilla, J.M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M.H. Shapero, P.I.W. de Bakker, J.B. Maller, A. Kirby, A.L. Elliott, M. Parkin, E. Hubbell, T. Webster, R. Mei, J. Veitch, P.J. Collins, R. Handsaker, S. Lincoln, M. Nizzari, J. Blume, K.W. Jones, R. Rava, M.J. Daly, S.B. Gabriel, D. Altshuler, Integrated detection and population-genetic analysis of SNPs and copy number variation, Nat. Genet. 40 (2008) 1166–1174.
[42] G.M. Cooper, T. Zerr, J.M. Kidd, E.E. Eichler, D.A. Nickerson, Systematic assessment of copy number variant detection via genome-wide SNP genotyping, Nat. Genet. 40 (2008) 1199–1203.
[43] G.M. Cooper, H. Mefford, Detection of copy number variation using SNP genotyping, Methods Mol. Biol. 767 (2011) 243–252.
[44] D. Pinto, K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A.C. Lionel, B. Thiruvahindrapuram, J.R. MacDonald, R. Mills, A. Prasad, K. Noonan, S. Gribble, E. Prigmore, P.K. Donahoe, R.S. Smith, J.H. Park, M.E. Hurles, N.P. Carter, C. Lee, S.W. Scherer, L. Feuk, Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants, Nat. Biotechnol. 29 (2011) 512–520.
[45] "The Wellcome Trust Case Control Consortium", genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls, Nature 464 (2010) 713–720.
[46] A.E. Dellinger, S.-M. Saw, L.K. Goh, M. Seielstad, T.L. Young, Y.-J. Li, Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays, Nucleic Acids Res. 38 (2010) e105.
[47] C. Alkan, B.P. Coe, E.E. Eichler, Genome structural variation discovery and genotyping, Nat. Rev. Genet. 12 (2011) 363–376.
[48] J.M. Kidd, G.M. Cooper, W.F. Donahue, H.S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, E. Haugen, T. Zerr, N.A. Yamada, P. Tsang, T.L. Newman, E. Tuzun, Z. Cheng, H.M. Ebling, N. Tusneem, R. David, W. Gillett, K.A. Phelps, M. Weaver, D. Saranga, A. Brand, W. Tao, E. Gustafson, K. McKernan, L. Chen, M. Malig, J.D. Smith, J.M. Korn, S.A. McCarroll, D.A. Altshuler, D.A. Peiffer, M. Dorschner, J. Stamatoyannopoulos, D. Schwartz, D.A. Nickerson, J.C. Mullikin, R.K. Wilson, L. Bruhn, M.V. Olson, R. Kaul, D.R. Smith, E.E. Eichler, Mapping and sequencing of structural variation from eight human genomes, Nature 453 (2008) 56–64.
[49] P. Medvedev, M. Stanciu, M. Brudno, Computational methods for discovering structural variation with next-generation sequencing, Nat. Methods 6 (2009) S13–S20.
[50] N. Krumm, P.H. Sudmant, A. Ko, B.J. O'Roak, M. Malig, B.P. Coe, N. NHLBI Exome Sequencing Project, A.R. Quinlan, D.A. Nickerson, E.E. Eichler, Copy number variation detection and genotyping from exome sequence data, Genome Res. 22 (2012) 1525–1532.
[51] R. Li, Y. Li, H. Zheng, R. Luo, H. Zhu, Q. Li, W. Qian, Y. Ren, G. Tian, J. Li, G. Zhou, X. Zhu, H. Wu, J. Qin, X. Jin, D. Li, H. Cao, X. Hu, H. Blanche, H. Cann, X. Zhang, S. Li, L. Bolund, K. Kristiansen, H. Yang, J. Wang, J. Wang, Building the sequence map of the human pan-genome, Nat. Biotechnol. 28 (2012) 57–63.
[52] D.A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G.T. Roth, X. Gomes, K. Tartaro, F. Niazi, C.L. Turcotte, G.P. Irzyk, J.R. Lupski, C. Chinault, X.-z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D.M. Muzny, M. Margulies, G.M. Weinstock, R.A. Gibbs, J.M. Rothberg, The complete genome of an individual by massively parallel DNA sequencing, Nature 452 (2008) 872–876.
[53] R.E. Mills, K. Walter, C. Stewart, R.E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S.C. Yoon, K. Ye, R.K. Cheetham, A. Chinwalla, D.F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L.M. Iakoucheva, Z. Iqbal, S. Kang, J.M. Kidd, M.K. Konkel, J. Korn, E. Khurana, D. Kural, H.Y.K. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X.J. Mu, J. Nemesh, H.E. Peckham, T. Rausch, A. Scally, X. Shi, M.P. Stromberg, A.M. Stutz, A.E. Urban, J.A. Walker, J. Wu, Y. Zhang, Z.D. Zhang, M.A. Batzer, L. Ding, G.T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E.E. Eichler, M.B. Gerstein, M.E. Hurles, C. Lee, S.A. McCarroll, J.O. Korbel, Mapping copy number variation by population-scale genome sequencing, Nature 470 (2011) 59–65.
[54] C. Alkan, S. Sajjadian, E.E. Eichler, Limitations of next-generation genome sequence assembly, Nat. Methods 8 (2011) 61–65.
[55] E.J. Hollox, U. Huffmeier, P. Zeeuwen, R. Palla, J. Lascorz, D. Rodijk-Olthuis, P.C.M. van de Kerkhof, H. Traupe, G. de Jongh, M. den Heijer, A. Reis, J.A.L. Armour, J. Schalkwijk, Psoriasis is associated with increased beta-defensin genomic copy number, Nat. Genet. 40 (2008) 23–25.
[56] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W.R. McCombie, J. Hicks, M. Wigler, Tumour evolution inferred by single-cell sequencing, Nature 472 (2012) 90–94.