Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics

S Le Scouarnec and SM Gribble

Genomic rearrangements can result in losses, amplifications, translocations and inversions of DNA fragments thereby modifying genome architecture, and potentially having clinical consequences. Many genomic disorders caused by structural variation have initially been uncovered by early cytogenetic methods. The last decade has seen significant progression in molecular cytogenetic techniques, allowing rapid and precise detection of structural rearrangements on a whole-genome scale. The high resolution attainable with these recently developed techniques has also uncovered the role of structural variants in normal genetic variation alongside single-nucleotide polymorphisms (SNPs). We describe how array-based comparative genomic hybridisation, SNP arrays, array painting and next-generation sequencing analytical methods (read depth, read pair and split read) allow the extensive characterisation of chromosome rearrangements in human genomes.

Heredity (2012) 108, 75–85; doi:10.1038/hdy.2011.100; published online 16 November 2011

Keywords: array-CGH; array painting; breakpoint mapping; copy-number variant; next-generation sequencing; structural variant

INTRODUCTION

Diverse types of genomic variants have been described (Scherer et al., 2007) thanks to the development and expansion of molecular biology and cytogenetic techniques, and contribute largely to human disease, normal phenotypic variation and karyotypic evolution. Structural variants (SVs) within individual genomes result from chromosomal rearrangements affecting at least 50 bp (Alkan et al., 2011a) and include deletions and duplications known as copy-number variants (CNVs), inversions and translocations. Rearrangements are triggered by multiple events including external factors such as cellular stress and incorrect DNA repair or recombination (Mani and Chinnaiyan, 2010). Notably, segmental duplications (low-copy repeats), which are particularly frequent in subtelomeric regions (Linardopoulou et al., 2005), facilitate nonallelic homologous recombination and are considered as hotspots for recurrent rearrangements (Mefford and Eichler, 2009; Stankiewicz and Lupski, 2010; Ou et al., 2011).

The conventional cytogenetic methods, ‘chromosome banding’ and ‘karyotyping’ are very informative and are still commonly used. However, these techniques are limited to the detection of numerical chromosomal aberrations (aneuploidy, polyploidy) and microscopic SVs a few megabases in size (Table 1). Molecular cytogenetic approaches enable the detection of submicroscopic SVs and have been crucial for studying complex rearrangements, generated by more than two chromosomal breakage events, refining breakpoints and performing cross-species comparisons (Speicher and Carter, 2005). These newer approaches have mostly relied on the use of ‘fluorescence in situ hybridisation’ (FISH; Bauman et al., 1980) where fluorescence microscopy reveals the presence and localisation of defined labelled DNA probes binding to complementary sequences on targets, traditionally metaphase chromosome spreads. To facilitate detection of events such as translocations, whole chromosome-specific DNA probes or ‘paints’ have been used (‘chromosome painting’; Cremer et al., 1988; Lichter et al., 1988; Pinkel et al., 1988). To increase resolution, shorter probes have been introduced (for example, fosmids and very recently oligonucleotide libraries; Yamada et al., 2011) and/or the target has been refined by replacing condensed chromosomes with extended chromatin fibres (‘Fibre-FISH’; Heng et al., 1992; Wagiet al., 1992; Parra and Windle, 1993). Furthermore, Fibre-FISH is now facilitated by an automated procedure called ‘molecular combing’ (Michalet et al., 1997). Alternative targeted approaches have simplified CNV detection (Feuk et al., 2006). For example, ‘real-time qPCR’ (Bieche et al., 1998) and ‘MLPA’ (multiplex ligation-dependent probe amplification) are broadly used to detect recurrent events in clinical genetics (Schouten et al., 2002). While these different approaches are restricted to specific regions, some FISH-based techniques have been developed to detect genomic aberrations at the whole-genome level without prior knowledge (Table 1). For example, copy number differences between two genomes can be detected using ‘comparative genomic hybridisation’ (CGH; Kallioniemiet al., 1992); and subtle translocations and complex rearrangements can be characterised using techniques derived from chromosome painting such as ‘M-FISH’ (multiplex-FISH; Speicher et al., 1996) and ‘SKY’ (spectral karyotyping; Schrock et al., 1996) where all chromosomes are differentially coloured in a single experiment (Darai-Ramqvist et al., 2006; Stephens et al., 2011). These methods are experimentally demanding, labour-intensive and the resolution still limited by the use of chromosomes as targets (Table 1).

Precise determination of SV boundaries is crucial for accurate genotype–phenotype correlations, which are dependent on the extent of genes or regulatory regions that are disrupted or vary in copy number (Huang et al., 2010). In addition, nucleotide breakpoint resolution gives insights into the mechanisms underlying SV formation (Korbel et al., 2007; Gu et al., 2008; Kidd et al., 2008; Conrad et al., 2010a; Stankiewicz and Lupski, 2010; Mills et al., 2011).
Recent developments in molecular cytogenetics
S Le Scouarnec and SM Gribble

2011). Completion of the human genome sequence in the early 2000s (Lander et al., 2001; Venter et al., 2001) and progress in molecular biology techniques gave rise to new genome-wide screening methods, revolutionising the understanding of the genomes of healthy individuals (Iafrate et al., 2004; Sebat et al., 2004; Redon et al., 2006; Conrad et al., 2010b) as well as patients with disease. In this review, we will discuss how microarray and next-generation sequencing (NGS) technologies can be utilised to reveal and extensively characterise chromosome rearrangements. While the focus of this review is on humans, since the majority of techniques presented herein have largely been developed to study human genomes, these new advances are species-independent and hold great promise for future studies in various areas, including karyotype evolution and phylogenomics (Griffin et al., 2008; Skinner et al., 2009; Volker et al., 2010).

ARRAY-BASED TECHNIQUES

A brief introduction to arrays
DNA microarrays or ‘chips’ are currently applied to a wide range of applications in molecular biology. Originally developed for gene expression profiling, they are now commonly used to unmask copy number changes (array-based CGH), for single-nucleotide polymorphism (SNP) genotyping, as well as to study DNA methylation, alternative splicing, miRNAs and protein–DNA interactions (array-based ChIP (Chromatin Immunoprecipitation)). In short, each array consists of thousands of immobilised nucleic acid sequences (for example, oligonucleotide probes or cloned sequences). Labelled DNA or RNA fragments are applied to the array surface, allowing the hybridisation of complementary sequences between ‘probes’ and ‘targets’. The main advantages of this technology are its sensitivity, specificity and scale as it enables data for thousands of relevant genomic regions of interest to be generated rapidly in a single experiment. Lastly, but important for precious clinical samples, the amount of input sample material required is generally low, usually <1 μg.

CNV discovery using CGH and SNP arrays
While CGH arrays were fabricated specifically for the detection of CNVs in genomes, SNP arrays, initially designed for large-scale genotyping and essential for linkage and association studies, can also be used for this purpose. The genome-wide coverage of features on these arrays allows the discovery of CNVs without any prior knowledge. Some commercial arrays are designed to more easily identify recurrent rearrangements (in particular microdeletion syndromes) or to genotype CNVs present in >1% of the general population (known as copy number polymorphisms, CNP; Alkan et al., 2011a). A list of current commercial human catalogue oligonucleotide arrays is provided in Supplementary Table S1, and arrays are also available for multiple organisms. In addition, array vendors generally provide flexibility in design such that the researcher can easily adapt the content of the array in order to increase the resolution in one or more regions relevant for their study (‘custom designs’).

Table 1 Evolution of genome-wide methods for identifying different classes of chromosomal rearrangements

| Techniques | Deletions | Insertions | Unbalanced translocations | Balanced translocations | Inversions | LOH and UPD |
|------------|-----------|------------|---------------------------|------------------------|------------|-------------|
| Early 1970s Karyotyping/G-banding | Yes | Yes | Yes | Yes | No | Low (≥ several Mb) |
| Early 1990s CGH | Yes | No | Yes | No | No | Low (≥ several Mb) |
| Mid 1990s M-FISH/SKY/COBRA | Yes | Yes | Yes | Yes | No | High |
| Late 1990s RxFISH | Yes | Yes | Yes | Yes | No | Low (≥ several Mb) |

Abbreviations: BAC, bacterial artificial chromosome; CGH, comparative genomic hybridisation; COBRA, combined binary ratio labelling; FISH, fluorescence in situ hybridisation; LOH, loss of heterozygosity; M-FISH, multiplex FISH; NGS, next-generation sequencing; RxFISH, Rainbow cross-species FISH or cross-species colour banding; SNP, single-nucleotide polymorphism; SKY, spectral karyotyping; UPD, uniparental disomy.

Methods in the grey-shaded area are discussed in this review.

Array-CGH. The first array-based CGH experiments (Solinas-Toldo et al., 1997; Pinkel et al., 1998) were designed to improve the resolution obtained with conventional CGH (Kallioniemi et al., 1992). Normal metaphase chromosomes were replaced with arrays containing thousands of DNA sequences. Initially, these sequences were large genomic clones of typically 80–200 kb in length, namely BAC or PAC (bacterial/P1-derived artificial chromosome) clones selected throughout the genome at 1-Mb intervals (~3000 BAC clones per array) (Snijders et al., 2001; Fiegler et al., 2003a; Chung et al., 2004). In 2004, the first whole-genome tiling path array was created (Ishkanian et al., 2004). This array comprised >30 000 overlapping BAC clones covering the entire genome, increasing the array resolution and the potential to detect copy number changes. Array resolution has further improved since technology has allowed an increase in the number of features present on an array and shorter sequences have been used as targets: cDNA (Pollack et al., 1999; Heiskanen et al., 2000), PCR amplicons (Mantripragada et al., 2004; Dhami et al., 2005) and above all, oligonucleotide probes that are now...
widely used (Brennan et al., 2004; Carvalho et al., 2004). This recent significant increase in array resolution has allowed the detection of genetic imbalances as small as just a few kilobases in size and also enables the boundaries of an imbalance to be better defined.

In array-CGH, test and reference DNAs are labelled with different fluorophores (for example, Cy5 and Cy3), and then simultaneously hybridised onto arrays in the presence of Cot-1 DNA to reduce the binding of repetitive sequences (Figure 1). If only low amounts of DNA are available (for example, in prenatal diagnosis or tumour analysis), amplification methods can be applied before labelling (Guillaud-Bataille et al., 2004; Le Caignec et al., 2006; Fiegler et al., 2007) although data quality is in general substantially reduced (Talseth-Palmer et al., 2008; Przybytkowski et al., 2011). After hybridisation, washing and scanning, Cy5 and Cy3 fluorescence intensities are measured for each feature on the array, normalised, and log2 ratios of the test DNA (for example, Cy5) divided by the reference DNA (for example, Cy3) are then plotted against chromosome position. Theoretically, for each position, a value of 0 indicates a normal copy number \( \log_2 (2/2) = 0 \) result, while a log2 ratio of 0.58 \( \log_2 (3/2) = 0.58 \) indicates a one copy gain in test compared with reference, and a log2 ratio of -1 \( \log_2 (1/2) = -1 \) indicates a one copy loss in test compared with reference. To minimise the influence of CNVs in the reference DNA for the identification of CNVs in the test DNA, a pool of ‘normal’ DNA samples, ideally \( >100 \), can be used as a reference. A large variety of algorithms designed to detect CNVs from array-CGH data (‘calling’ algorithms) have been published, for example, ‘DNA-copy’ (Oshen et al., 2004), ‘SW-ARRAY’ (Price et al., 2005), ‘SMAP’ (Andersson et al., 2008), ‘GADA’ (Pique-Regi et al., 2008) and ‘ADM3’ (R package available at http://cran.r-project.org). These algorithms search for intervals where the average log2 ratio exceeds specified thresholds. If probe response is good and background noise is low, a few probes can be sufficient to detect imbalanced regions with confidence (generally a minimum of 3–10 probes are used, depending on platforms; Alkan et al., 2011a). Algorithms can more accurately detect CNVs and will produce less false positive calls if data are normalised to correct for artefacts such as GC-bias, waves (Marioni et al., 2007) or dye-bias (Fitzgerald et al., 2011).

Commercial arrays (Supplementary Table S1) provided by companies such as Agilent Technologies (Santa Clara, CA, USA) BlueGnome (Cambridge, UK) Oxford Gene Technology (Oxford, UK) and Roche NimbleGen (Madison, WI, USA) in the UK, offer robustness, sensitivity and flexibility compared with early BAC arrays. As previously stated, the researcher can order a custom design including dense coverage focusing on single or multiple chromosomal regions where higher resolution is required. Conrad et al. (2010b) describe the use of a set of 20 ultra-high resolution oligonucleotide arrays comprising 42 million probes in total, with a median probe spacing of just 56 bp across the entire genome. Such high resolution enabled the identification of 11,700 CNVs \( >443 \) bp in the genomes of 40 normal individuals. The fabrication processes of the arrays vary between manufacturers. For example, Agilent Technologies utilises in situ inkjet technology (‘SurePrint Technology’, Agilent Technologies) to synthesise 60-mer oligonucleotide array features (Barrett et al., 2004). This technology produces highly reproducible features and excellent signal-to-noise ratios, assuring maximum sensitivity and specificity. Custom arrays can be designed and ordered using the online eArray application (https://earray.chem.agilent.com/earray), which contains at present over 28 million in silico-validated human
oligonucleotide sequences. These 60-mers span exonic, intronic, intergenic, pseudoautosomal, segmented duplication DNA regions and copy number variable regions. In addition to sequences contained in the database, any custom oligonucleotide sequence with a size ranging from 25 to 60 bp can be printed. For every oligonucleotide on the array, scores can be provided by array manufacturers, which can predict their performance on a genomic array and help to interpret derivative log2 ratio values in breakpoint regions (Sharp et al., 2007). Scores are based on various parameters such as melting temperature (Tm), SNP content, sequence complexity and uniqueness of the oligonucleotide sequence. For cost-effectiveness, the user can choose between different formats, from 8 x 60 K up to 1 x 1 M for SurePrint G3 arrays (Supplementary Table S1). Furthermore, designs can be shared with collaborators through the online application. Roche NimbleGen high-density array manufacturing is based on photo-mediated synthesis process using the Maskless Array Synthesizer technology (Nuwaysir et al., 2002). In comparison to other in situ synthesis technologies such as inkjet deposition, this method enables the production of more features on the glass slide, and oligonucleotide lengths are usually ranging between 50 and 75 bp. They have recently introduced very high-resolution arrays composed of 4.2 million array features (284 bp median feature spacing), and different array formats are also available (Supplementary Table S1). The array design is made on-demand by Roche NimbleGen from a list of regions of interest supplied by the customer. Similar to Agilent Technologies, Roche NimbleGen offers whole-genome catalogue arrays and custom solutions designed to study a range of various organisms.

SNP arrays. As with array-CGH technology, SNP arrays have undergone huge developments over the last few years (Kennedy et al., 2003; Gunderson et al., 2005; LaFramboise, 2009), with the ability to genotype a few thousands SNPs at first, rising to millions of SNPs today in the latest arrays. In addition to the advances in resolution, the design of the arrays is continually incorporating more informative SNPs, as a result of large-scale studies such as the HapMap Project (The International HapMap Consortium, 2003) and the 1000 genomes project (Durbin et al., 2010). Although SNPs account for a substantial part of genetic variation, chromosomal rearrangements have a tremendous role in disease, evolution and tumorigenesis, and SNP arrays have progressively started to be used to simultaneously genotype SNPs and detect rare and common genomic rearrangements (Bignell et al., 2004; Huang et al., 2004; Peiffer et al., 2006). Besides amplifications and deletions detected by both CGH and SNP arrays, SNP arrays can reveal mosaicism, extended regions of loss of heterozygosity, uniparental disomy (Conlin et al., 2007). Accurate calculation of copy numbers (Greenman et al., 2010) and appropriate labelling as they are present on different derivative chromosomes have been replaced by arrays (Fiegler et al., 2003b).

First, the two aberrant or ‘derivative’ chromosomes involved in the reciprocal translocation are isolated. This can be achieved by flow-sorting (Gribble et al., 2009) or by microdissection (Backx et al., 2007). Subsequently, each derivative chromosome, represented by one (Gribble et al., 2004) or generally multiple copies, is amplified using DOP-PCR or commercially available whole-genome amplification kits to provide sufficient DNA. The amplified products are then differentially labelled with fluorescent dyes (Cy5 and Cy3), and co-hybridised onto an array, which is then scanned after excess labelled probe is washed off (Figure 2). As for array-CGH, log2 ratios for Cy5/Cy3 intensities are plotted against chromosome position for each feature. Because the chromosomal regions flanking each side of the breakpoint are differentially labelled as they are present on different derivative...
chromosomes, the position where the log2 ratios changes from high to low ratios (or vice versa) defines the breakpoint, and breakpoint spanning clones usually show intermediate ratios (Fiegler et al., 2003b; Backx et al., 2007) (Figure 2). Fine-mapping of breakpoints is only dependent on the resolution of the array. In the initial reports of the array painting method, 1-Mb whole-genome or custom tiling BAC arrays were used (Fiegler et al., 2003b). Array painting benefited from the evolution of array-CGH technology and BAC arrays have been replaced by whole-genome or region-specific high-resolution oligonucleotide arrays, allowing higher resolution and better accuracy of breakpoint determination (Gribble et al., 2007). Precise breakpoint mapping of balanced translocations can give insights into associated phenotypes in patients. For example, array painting performed with a 244 K CGH array for a t(10;13)(q22;p13) balanced translocation suggested that C10orf11, which was disrupted by the translocation, could contribute to the mental retardation phenotype in 10q22 deletion patients (Tzschach et al., 2010). Breakpoints identified by array technologies can be independently validated by FISH assays to visually demonstrate the rearrangements in individual cells.

This robust procedure can be used to determine the composition of any isolated chromosome and has applications other than mapping balanced translocation breakpoints. Thus, complex chromosome rearrangements, involving more than two chromosomes, can be deciphered (Fauth et al., 2006), and in some instances other inter-chromosomal aberrations may be identified. Furthermore, array painting can replace conventional chromosome painting to determine cross-species homology, which can give insights into karyotype evolution. For example, white-cheeked gibbon chromosome 14 was hybridised onto a human 1-Mb array, which identified syntenic blocks on human chromosomes 2 and 17 (Gribble et al., 2004).

An alternative technique to array painting for fine-mapping of translocation and complex rearrangements breakpoints, based on ‘Chromatin Conformation Capture on Chip’ or 4C (Simonis et al., 2009) has been described. Briefly, many fragments across the breakpoints are captured by cross-linking physically close parts of the genome, followed by restriction enzyme digestion, locus-specific inverse PCR and templates hybridised to 4C-tailored microarrays. Clustering of positive signals displaying increased intensities predicts breakpoints at the resolution of the array. It claims to be particularly valuable if isolation of derivative chromosomes is not achievable, and to characterise inversions.

**NGS-BASED TECHNIQUES**  
**A brief introduction to NGS**

Using conventional Sanger sequencing, it has taken more than a decade of international effort to sequence the human genome (Lander et al., 2001; Venter et al., 2001). Since the development of NGS (or ‘second-generation’ sequencing) technologies in 2005, sequencing of a whole human genome can now be achieved in a few days and at much lower cost. Also known as ‘massively parallel sequencing’, these technologies allow the sequencing of millions of DNA molecules simultaneously after library preparation of fragments, to produce sequence reads. Sequence reads are generally aligned to the reference genome and base variants, small insertions/deletions (indels) and SVs (>80 bp) can be detected. The most commonly used platforms at present have been developed by Illumina (Genome Analyzer/HiSeq,
Deciphering chromosomal rearrangements with NGS technology

Information provided by read mapping and sequence coverage enables the detection of SVs and NGS is becoming an attractive alternative to array-based assays in the field of molecular cytogenetics. Among the many advantages of high-throughput NGS, SVs of all types and sizes can theoretically be detected, breakpoints can be mapped with high resolution, down to the basepair level in some instances, and complex rearrangements can be characterised with the possibility to study multiple breakpoints in a single experiment. Four different approaches have been described to characterise SVs: (i) read-depth analysis, which can only detect gains and losses; (ii) pair-read analysis (paired-end mapping); (iii) split-read analysis; and (iv) assembly methods, all of which can detect in theory all types of rearrangements including copy-neutral rearrangements (inversions and translocations) (Figure 3). A variety of tools based on one or more of these methods have been developed to analyse chromosomal rearrangements according to the genomic regions affected, the size-range and breakpoint precision (Medvedev et al., 2009; Alkan et al., 2011a; Mills et al., 2011). We will discuss how each method can be used to characterise genomic rearrangements, with the exception of local assembly approaches that are still limited by read length and cost (Alkan et al., 2011b).

Read-depth method. Read-depth NGS data (Campbell et al., 2008; Chiang et al., 2009; Yoon et al., 2009) are essentially providing similar information to that obtained from array-CGH, by indicating copy-number gains (≥2 copies for a diploid genome) or losses (<2 copies). Sequence read depth, that is, the number of reads mapping at each chromosomal position, is in theory randomly dispersed and significant divergence from the normal Poisson distribution indicates copy-number variation (Figure 3). Duplications and amplifications are indicated by the presence of regions showing excessive read depth, whereas low read depth indicates heterozygous deletion and absence of coverage is suggestive of homozygous deletion. Statistical power is limited for smaller CNVs but increasing sequence coverage can in some instances improve sensitivity (Chiang et al., 2009). Factors such as GC content, homopolymeric stretches of DNA or preferential PCR amplification at the library preparation stage can introduce biases. Repetitive DNA regions are also problematic as reads are aligned with low confidence (low ‘mapping quality’; Li et al., 2008), providing poor information on copy-number status, but longer reads will increase mapping specificity in the future. Applying read-depth analysis to cancer cell lines has shown that the dynamic range for absolute copy-number evaluation is greater than that detected by SNP arrays (Campbell et al., 2008), which tend to saturate for high intensity values. For example, Chiang et al. (2009) found a 55.6-fold increase by NGS compared with only a 16-fold increase by SNP array for the ERBB2 locus in a breast carcinoma cell line. This increased dynamic range of NGS may lead to new insights into segmental duplications (Alkan et al., 2009) and multicopy gene families (Sudmant et al., 2010).

Read-pair method. Currently, the most powerful method to study chromosome rearrangements is the paired-end read mapping technique (Tuzun et al., 2005; Korbel et al., 2007) (Figure 3). Sequence read pairs are short sequences from both ends of each of the millions of DNA fragments (‘inserts’) generated by library preparation. Clustering of at least two discordant pairs of reads, either by size or by orientation, is suggestive of a chromosome rearrangement. When aligned to the reference genome, read pairs (> <) are expected to map at a certain distance (> ---- <) corresponding to the average library insert size (typically 200–500 bp and up to 5 kb for large-insert libraries); a spanning distance significantly different from the average insert size indicates putative SVs. Deletions are identified by read pairs spanning a larger genomic region when mapped to the reference not carrying the deletion (>-(del)< → >----<). By contrast, insertions or tandem duplications in the sequenced sample will cause the reads to map closer as they are absent from the reference genome (>-(ins/dup)< → <---->). In addition to the expected span distance of a sequence read pair, aberrant mapping orientation can identify inversions (---- ----) and tandem duplications (<----->) (Korbel et al., 2007; Kidd et al., 2008) (Figure 3). Novel insertions, as compared with published reference genomes, are identifiable when only one read of the pair is mapping (<---->). Data from short-insert libraries often need to be supplemented by data from large-insert libraries generated by large circular fragments of DNA typically of 2–5 kb, providing higher physical coverage at breakpoints thereby facilitating SV detection (Shendure et al., 2005; Bentley et al., 2008) (Figure 4). Short-insert libraries (200–500 bp) have a limited capacity to detect SVs mediated by segmental duplications (or low-copy repeats) that harbour a substantial part of SVs (Sharp et al., 2005; Cooper et al., 2007; Kidd et al., 2008; Conrad et al., 2010a), because reads map to multiple similar genomic loci (Li et al., 2008). Another example is the limit to detect insertions larger than the library insert size. Conversely, small events (<400 bp) can be missed with large-insert libraries because the expected size variance between the mate pairs will not be significantly altered. The lower resolution associated with large-insert libraries can also mistake complex events, where several breakpoints are in close proximity such as small insertions flanked by deletions, for simple deletions (Bentley et al., 2008).

Split-read method. The third approach commonly applied to NGS data is the split-read method (Figure 3). Although this method was originally developed for Sanger sequencing (Mills et al., 2006) and will be much more efficient with longer read length, it is already capable of precisely mapping breakpoints for small deletions (1 bp–10 kb) in unique regions of the genome using the algorithm Pindel (Ye et al., 2009) and read lengths as low as 36 bp. The first stage is to map all reads to the reference genome, and select read pairs responding to the following criteria: one read maps perfectly (no mismatches) and uniquely (no other genomic location), and the other read of the
pair cannot be mapped (that is, it is across the rearrangement breakpoint). For each of these pairs, using the location and orientation of the mapped read, Pindel searches for the paired unmapped read ('split' read) by performing multiple local alignments. In the case of deletions, candidate unmapped reads are split into two fragments that map separately, and analysis of the alignment deciphers the breakpoint at the basepair level. The AGE algorithm described recently has been designed to identify exact breakpoints for tandem duplications, inversions and complex events (Abyzov and Gerstein, 2011). Thus, this method has a significant advantage on others applied to array or NGS data, which can identify breakpoints with high resolution (Gribble et al., 2007; Mills et al., 2011) but require an additional PCR or high-throughput capture step (Conrad et al., 2010a; Mills et al., 2011) followed by conventional or NGS to reach basepair resolution.

**Fine-mapping of translocation breakpoints using NGS.** In one of the first studies applying NGS to fine map a reciprocal translocation breakpoint, derivative chromosomes were isolated by flow-sorting, sequenced and single-end reads aligned to the two corresponding chromosomes (Figure 4a). Read-depth analysis identified breakpoints within 1 kb, which were subsequently confirmed at the basepair level by PCR amplification and sequencing (Chen et al., 2008). With whole-genome sequencing becoming more affordable and paired-end

---

**Figure 3** Four methods to identify SVs from NGS data. These methods are often used in combination to detect chromosomal rearrangements and characterise breakpoints (red arrows) with precision. *De novo* assembly methods are still challenging but have the potential to accurately and rapidly characterise all classes of rearrangements. MEI, mobile-element insertion; RP, read pair. For further details and full figure legend, see Alkan et al., 2011a. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (Alkan et al., 2011a), copyright 2011.
technology now available, flow-sorting of derivative chromosomes becomes less critical, as essentially, pairs of reads mapping to different chromosomes will identify translocations (Figures 4b and c; Slade et al., 2010). Large-insert paired-end libraries (Figure 4c) of ~3 kb are generally preferred to short-insert libraries to increase physical coverage, and maximise chances of observing read pairs consistently spanning the breakpoint (Chen et al., 2010; Slade et al., 2010). If high sequence coverage is reached and reads span the breakpoint (‘split’ reads) (Figure 4), it should be straightforward to directly identify the exact breakpoint without the need for an extra PCR/sequencing step. For example, a method called SLOPE can rapidly identify sequence breakpoints for translocations using read-depth and split-read data (Abel et al., 2010).

**Insights from cancer genomes.** NGS has also revolutionised the understanding of cancer genomes by identifying not only the full spectrum of somatic point mutations (Mardis et al., 2009; Pleasance et al., 2010) but also giving more insights into complex whole-genome acquired rearrangements (Campbell et al., 2008; Stephens et al., 2009) (Figure 5). These studies showed that intra- and inter-chromosomal somatic rearrangements can be detected and are more frequent than envisaged, partly because they involve small aberrations beyond the resolution of previous molecular cytogenetics methods, emphasising the utility of NGS to study rearrangements (Meyerson et al., 2010). Discovery of fusion genes resulting from these rearrangements and having potential functional consequences is greatly facilitated. Furthermore, transcriptome sequencing using next-generation technologies can identify or validate putative fusion transcripts in a high-throughput manner (Maher et al., 2009).

**IMPACT ON PRESENT AND FUTURE STUDIES**

Recently developed molecular cytogenetic methods have provided new tools to accurately characterise chromosomal rearrangements and have uncovered the great complexity of human genome architecture (Pang et al., 2010). We have shown that each strategy has limitations, emphasising that approaches often need to be combined to capture the entire range of genetic variation (Alkan et al., 2011a; Mills et al., 2011).

Despite the enormous potential of high-throughput sequencing, array technology has progressed in the past few years and is still appropriate for a broad range of research projects. In addition to robustness, flexibility, and low input material required, array technologies do not demand as many resources as NGS technologies in terms of equipment and computational power. Arrays also give the possibility to study a large number of samples in a cost-effective manner. For example, CNVs identified in discovery phases can be subsequently genotyped by arrays in large population samples and used in disease association studies (Craddock et al., 2010), however data can be less accurate than sequencing at high-copy number states (Chiang et al., 2009). Array-based assays have replaced karyotyping for the diagnosis

**Figure 4** Mapping translocation breakpoints by NGS. Bars depict sequencing reads mapping to distinct chromosomes (chromosome 1 and chromosome 2) each side of the translocation breakpoint. Sequence coverage (number of times the breakpoint is covered by sequencing reads) vs physical coverage (number of times the breakpoint is covered by library fragments) are indicated. (a) Single-end sequencing. (b) Paired-end sequencing from a short-insert library (<500 bp). (c) Paired-end sequencing from a large-insert library (>1 kb), increasing physical coverage at the breakpoint site and likelihood of characterising the translocation. Reads spanning the translocation breakpoint are called ‘split reads’ and can identify breakpoints at basepair resolution. Higher depth of sequence coverage (using short-insert libraries) and longer read lengths theoretically generates more informative split reads. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (Meyerson et al., 2010), copyright 2010.
of developmental disabilities or congenital anomalies (Miller et al., 2010), and will remain the gold standard method until sequencing costs drop dramatically and downstream analyses are facilitated.

Array-based methods have revealed an unexpected level of rearrangement complexity such as imbalances in apparent balanced translocations (Gribble et al., 2005; Howarth et al., 2011), but they are mostly restricted to the detection of CNVs, and FISH is still required to distinguish tandem from dispersed duplications and decode complex rearrangements. Moreover, resolution achieved using arrays can be limited by the density of features printed on the glass slide and there has clearly been a bias towards detecting larger events thus far, even if sets of custom arrays can be employed to increase resolution (Conrad et al., 2010b; Park et al., 2010). The emergence of techniques based on high-throughput sequencing is opening new perspectives for chromosome rearrangement analyses. Whole-genome sequencing is comprehensive and reveals point mutations, indels, as well as all types of chromosome rearrangements including balanced events, and can be used to reconstruct genome architecture. Success of sequencing approaches is often dependent on obtaining sufficient coverage because of the relatively high level of sequencing error in NGS. Current analytical methods mostly rely on sequence alignment against a unique reference genome, and unspecific mapping of short reads to repetitive regions is problematic with many events mediated by repetitive elements potentially being missed (Conrad et al., 2010a).

However, third-generation sequencing technologies (Metzker, 2010) will provide longer reads more cheaply, enabling accurate de novo assembly and will help to overcome these issues.

With the increase in resolution and the larger number of SVs detected in each genome with current methods, the challenge is now to infer their phenotypic impact on normal variation and health (Huang et al., 2010). More resources will be needed to guide the interpretation, especially with the growing interest for personalised medicine. Up until now, NGS technologies have largely been applied to study the human genome, but complete sequencing of more than a thousand organisms (997 prokaryotes and 39 eukaryotes, May 2011; http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html) has now been completed and hundreds more are in progress. Methods described in this review can be utilised to detect and comprehend SV between species or strains and give new insights into recent evolution.
New developments in molecular cytogenetics
S Le Locquire and SM Grabb
Pinkel D, Landegent J, Collins C, Fuscoe J, Segraves R, Lucas J et al. (2011). The use of ultra-dense array CGH analysis for \textit{BMC Med Genomics} 4: 400–405. 

Parr P, Windle B (1993). High resolution visual mapping of stretched DNA by fluorescent hybridization. \textit{Nature Genet} 5: 17–21. 

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F et al. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. \textit{Nature Genet} 38: 1136–1148. 

Pinkel D, Landegent J, Collins C, Fuscoe J, Segraves R, Lucas J et al. (1998). Fluorescence \textit{in situ} hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. \textit{Proc Natl Acad Sci USA} 85: 9138–9142. 

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D et al. (2002). Comparative assessment of array-based platforms and calling algorithms for detection of copy number variants. \textit{Biotechnol Bioeng} 79: 512–520. 

Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S (2008). Fine-scale map resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. \textit{Nature Genet} 38: 1136–1148. 

Pikkarainen M, Muttinenka U, Kalliokoski K, Kallioniemi A, Pyrhönen S, Seppänen-To明确了人全身基因组的变异的综合研究。\textit{Nature Genet} 38: 1136–1148。