GeNeViSTA

Low-Pass Genome Sequencing: A Good Option for Detecting Copy Number Variations

Somya Srivastava, Shubha R Phadke

Department of Medical Genetics, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, India

Correspondence to: Dr Shubha R Phadke *Email:* shubharaophadke@gmail.com

Abstract

Low-pass genome sequencing (LPGS) is a technique to detect copy number variants and map their breakpoints by using the technology of next-generation sequencing (NGS). Cytogenetic microarray (CMA) has a high resolution but is restricted to only those areas of the genome for which it has probes, thereby missing many duplications and deletions. Next-generation sequencing can identify single nucleotide changes. LPGS utilizes the strengths of NGS to empower the field of cytogenetics by helping in identifying accurate breakpoints of genes disrupted by chromosomal aberrations. Since it uses the existing infrastructure of NGS, it is cheaper, has high throughput, requires low input DNA, and has a quick turnaround time consequently making it an ideal technique for prenatal samples where time and amount of sample are crucial. Various studies have found good concordance between the results of CMA and LPGS. The yield of testing does not increase, rather the ability to identify copy number variants in areas without probe, better delineation of breakpoint, technical ease and low cost per sample is where LPGS proves to be useful. LPGS, however, is afflicted by the bane of short read length which plagues next generation sequencing. In this article we discuss the various methods of LPGS and its advantages and disadvantages, and its applicability in the clinical setting.

Keywords: Low-pass genome sequencing, copy number variants, chromosomal microarray

Introduction

Traditional karyotyping has long been the gold standard for the diagnosis of various cytogenetic abnormalities in global developmental delay/ intellectual disability with or without congenital malformations. For antenatal cases, karyotyping was reserved for the diagnosis of aneuploidies in women with abnormal screening test for aneuploidies or those with ultrasonographically detected anomalies or aneuploidy markers. Traditional karyotyping can detect aneuploidy, chromosomal rearrangements (both balanced and unbalanced), triploidy and mosaicism. Though karyotyping allows the view of the whole genome in one go, the level of resolution is a major limitation of karyotyping and only large deletions and duplications of sizes larger than 10 Mb anywhere in the genome can be detected. Also, the exact breakpoints of translocated segments cannot be delineated by karyotyping. Hence, the yield of karyotype is limited to 7.4% in children with non-syndromic global developmental delay (Sadek et al., 2018) and 18.2% in fetuses with structural malformations (Fu et al., 2018). The other disadvantages of karyotyping are requirement of live tissue cells along with expertise and substantial time needed to analyze the results. The era of molecular cytogenetics has revolutionized chromosomal analysis by tremendous increase in resolution and eliminating the need of live cells and of a fastidious, painstaking culture process. Chromosomal microarray (CMA) is a high-resolution genomic technique to interrogate the whole genome in one go. Its higher resolution can provide a magnification of up to 100 times over traditional G-banding karyotyping (Friedman et al., 2009), and identify deletions and duplications of size 0.05Mb-0.1Mb anywhere in the genome (Martin et al., 2015). Such deletions or duplications involving larger segments of the genome typically more than 1kb and ranging up to several Mb are known as copy number variants (CNV) (Valsesia et al., 2013) and are responsible for various neuro-developmental disorders, autism, and congenital birth defects. They are too small to be detected on a karyotype and not routinely analysed in next-generation sequencing. Hence, chromosomal microarray has become the first line test for prenatal samples as well as for evaluation of neurodevelopmental disorders with or without malformations.

An intrinsic disadvantage with CMA is that it cannot study those areas for which it does not have probes and hence will miss small CNVs in the range of few hundred bases. Other limitations include inability to identify point mutations, balanced translocations and cryptic breakpoints. Moreover, analysis of areas of low-level mosaicism and interpretation of CNVs of uncertain significance is challenging (Martin et al., 2015). With the availability of NGS, the resolution of genomic analysis has increased up to the nucleotide level. Over the last 15 years of CMA and NGS, there has been a growing demand for having one assay which could detect both CNV and SNV (single nucleotide variation) in a single pipeline. This led to the advent of low pass NGS for detection of CNVs as a replacement for microarray CGH and as a complement to exome sequencing.

Definitions of the terms used in this article are mentioned in **Table 1**.

Next-Generation Sequencing: Principle

Low-pass genome sequencing is largely based on the principle of massively parallel DNA sequencing also known as next generation sequencing. NGS technique has 3 basic processes, namely:

1. Library preparation which involves fragmentation of the DNA and adapter ligation
2. Amplification (emulsion/bridge)-depending upon the platform used
3. Sequencing either by synthesis or ligation

Despite identifying variations at nucleotide level, NGS comes with its own pitfalls largely due to the techniques involved. As mentioned above, it involves fragmentation of DNA. These fragments of the DNA ultimately need to be re-assembled either against a reference genome or de novo. We know that it is easier to put together a photograph torn into 4 pieces rather than 40 pieces. The same analogy can be applied to the reassembly of DNA fragments; the shorter the read length, the higher the chances of errors in aligning the fragments. In the huge domain of genetic diseases, it will be difficult to distinguish one such error from the real single nucleotide variant in the sample DNA. Barter for this situation is to sequence the same fragment multiple times. Reading the same length multiple times will strengthen the call whether to consider a single base change as a systematic error or a true variant. This process of representation of a single nucleotide for a fixed number of times in a particular sequencing platform is defined as the depth of sequencing. It depends upon the read length, number of reads and entire length of the haploid genome sequenced. All areas of the genome are not equally covered (due to GC rich regions, bias in sampling, repetitive regions, poor DNA quality, pseudogenes) during massively parallel sequencing. Naturally, more the depth of sequencing, better would be its ability to detect a variant (**Figure 1**). Although currently associated with higher cost and time, this type of deep sequencing has found its use to study variants in cancer samples, viral infections, and drug resistance.

As per Moore's law, the cost of deep sequencing is bound to come down in the coming years, but in the present framework, the less expensive option is shallow depth of sequencing also known as low-pass whole genome sequencing (LPGS).

Low-pass genome sequencing is set to bring about a paradigm shift in the field of cytogenetics. It provides the single nucleotide resolution of NGS which helps in accurate mapping of the genes disrupted by chromosomal rearrangements. Because of its finer mapping, it has the potential of identifying new breakpoints and possibly new genetic etiologies. Since the entire process is automated, it has high throughput, quick turn-around time, low error rate and can work with low input of DNA. Instituting LPGS does not require any special machine. It can work on the infrastructure of previous NGS technology and uses the same output files as NGS. Additional software to read those files needs to be installed. The cost therefore drastically comes down as for a low depth of sequencing, multiple samples can be processed together. These qualities make it suitable for use in prenatal diagnosis where time, money and amount of sample are crucial. The various issues about LPGS, and the advantages and disadvantages of LPGS are discussed below.

Table 1 Definitions of terms used

Term	Definition
Copy number variants (CNV)	Deletions or duplications involving large segments of the genome typically more than 1kb in size
Chromosomal microarray	Microarray consists of a glass or silica chip which contains multiple, short single-stranded DNA sequences (probes) spanning the entire genome from normal humans. The patient's sample (target DNA) is allowed to hybridize with the complimentary probes and the fluorescence generated is read by a computer indicating loss or gain of chromosomal segments.
Cryptic breakpoint	Deletion of a gene due to an apparently balanced complex chromosomal rearrangement which involves more than 2 chromosome breaks.
Low-level mosaicism	Mosaicism seen in less than 20-25% of cells. It is difficult to distinguish it from technical noise.
Loss of heterozygosity	Both the chromosomes have the same allele i.e., they are homozygous. Such regions are usually benign and indicate a common founder ancestor, but in many cases they may harbor autosomal recessive genes with pathogenic sequence variations.
Adapter ligation	Process of attaching ssDNA to DNA fragments which acts like a barcode for the multiple fragments and also helps in amplification
Read length	Number of base pairs sequenced from a DNA fragment. Commonly available NGS platforms offer a read length of 150-200 base pairs
Depth of sequencing	Number of times a particular nucleotide is represented in a particular sequencing platform. It should be at least 10X. Most sequencing platforms offer a depth of sequencing of 30X.
GC rich regions	Areas in the genome where guanine and cytosine form >60% of bases. Such regions do not undergo amplification easily, hence may be underrepresented. Some of these regions may contain important genes which thereby may not be sequenced.
Pseudogenes	Any genomic sequence similar to a protein coding sequence but without any functional product of its own. Variations in the pseudogene are not commonly associated with diseases.
Quantitative PCR	PCR technique which quantifies the product generated in every cycle
Sanger sequencing	Gold standard method of sequencing where after amplification, DNA copies which differ by one nucleotide are fractionated according to size by gel electrophoresis and the fluorescence signals are recorded and interpreted to produce a linear base sequence

How deep is low-pass genome sequencing?

There is no clear consensus of how much depth would be considered as low depth. Majority of studies consider an average depth of coverage of <1X (Dong et al., 2016) as low-pass genome sequencing. However, a few consider it to range from 1X to 5X (Chaubey et al., 2020). Data from the 1000 genome project showed that a depth of at least 8X is required for reliable call of single nucleotide variations (SNVs).

Methods of LPGS

Multiple bioinformatic tools for detecting CNV are now in built in the NGS platforms. These tools use one of the following approaches:

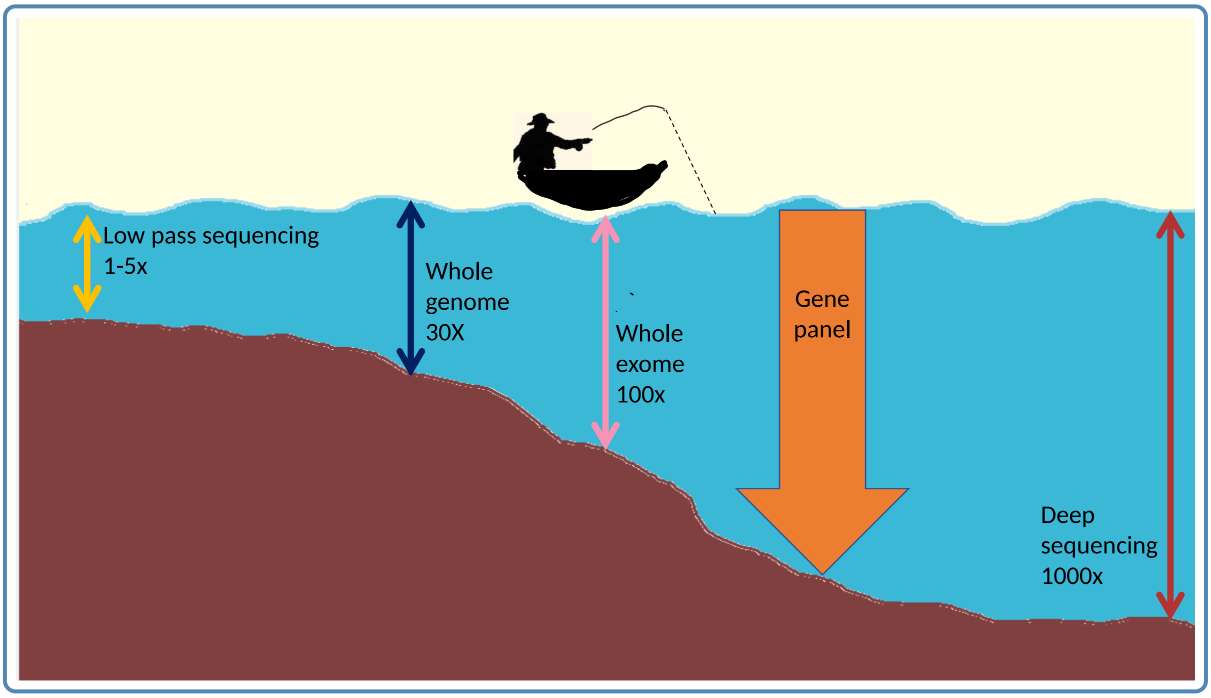

Figure 1 Representation of depth of sequencing in different types of sequencing.

1. Paired-end mapping: As mentioned above, NGS involves fragmentation of DNA to generate several short DNA fragments known as reads. Paired end refers to 2 ends of the same read. Sequencing is done from both the ends of the fragment and these paired reads are then aligned to a reference genome. There is a length of DNA sequence between the two ends which is not sequenced (known as insert size). If there is no major loss or gain of DNA in the fragment of DNA which has been aligned to the reference genome, then the pairs will map concordantly with the reference genome. If the paired ends map too far apart on the reference genome, it is likely that there is some deletion in the sample genome. Likewise, if the pairs map too close, then an insertion in the sample genome can be suspected (**Figure 2**).

2. Split-read approach: This approach also requires paired-end reads but one of the paired-end reads must map to an area containing the breakpoint. This read which maps to the area with the breakpoint is further spilt into multiple reads which are further aligned to the reference genome (**Figure 3**). This method helps in higher resolution of structural variants. Paired-end and split-read approaches are good for identifying the precise breakpoints but are not good to identify copy number variants. False positive and false negative results may also arise if the breakpoint lies in introns or in areas with low coverage (due to GC rich regions, bias in sampling, repetitive regions, poor DNA quality, pseudogenes, or mutation in the mapped region). Also, smaller the read size, more the chances of it being assigned to a different genomic position.

3. Depth of coverage approach: This is the most commonly used method as depth of coverage information is embedded in the NGS platforms. It is based on the fact that coverage is related to copy number. This method assumes that depth of sequencing of a particular region corresponds to its initial copy number. So, the relative depth is compared across the sample and areas of low depth when compared to the average genomic read depth are presumed to have a copy number loss and those with a higher read depth with copy number gain. Areas of genome with natural low coverage due to reasons mentioned above may present with false negatives (**Figure 4**). To overcome this bias, normalizing the coverage across sample and use of ratio is used. However, for samples with cell-free DNA used for non-invasive prenatal screening (NIPS) or tumour markers where target DNA is already less, increasing the depth of sequencing may be the

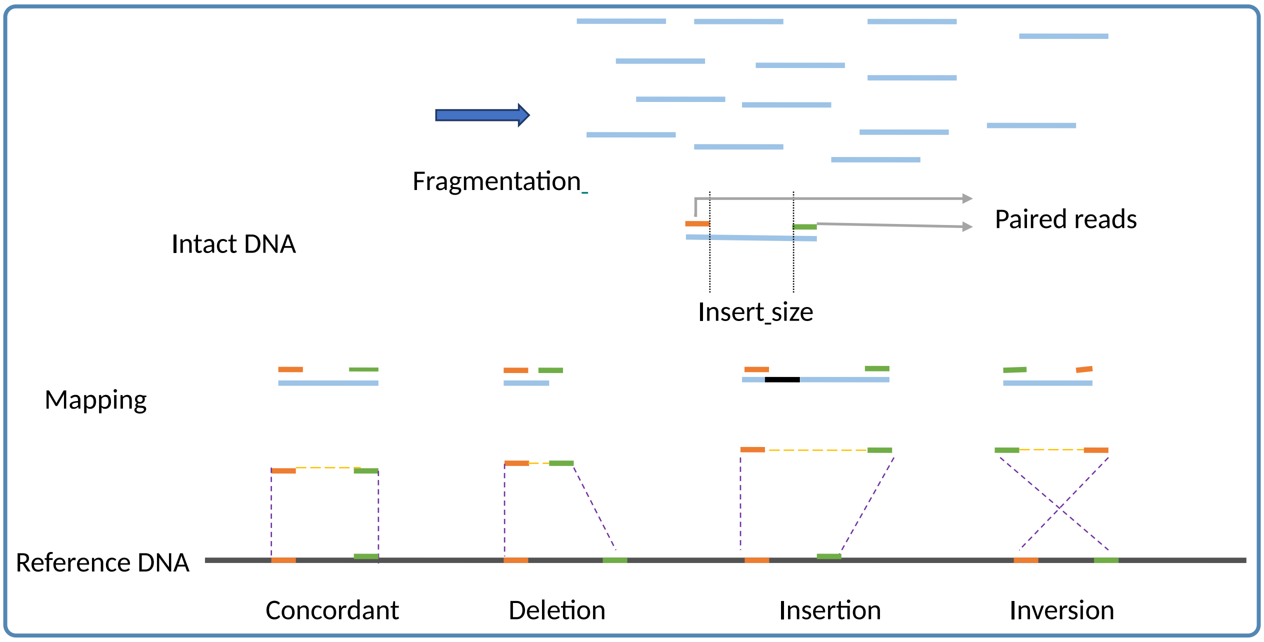

Figure 2 Paired-end mapping: the orange and green coloured bars are the paired end reads which sequence the fragment of DNA from both sides. These paired reads are then aligned to the reference genome. Discrepancy in alignment lead to suspicion of copy number changes.

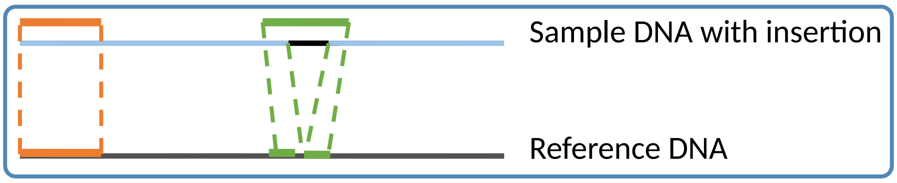

Figure 3 Split-read approach: paired-end mapping is done but the read which maps the breakpoint is further split into small reads which are again sequenced to identify the breakpoint accurately. This figure shows the split-read approach for an insertion in the sample DNA. The green read which maps the insertion on the sample DNA is further split into 2 reads to map the breakpoint precisely.

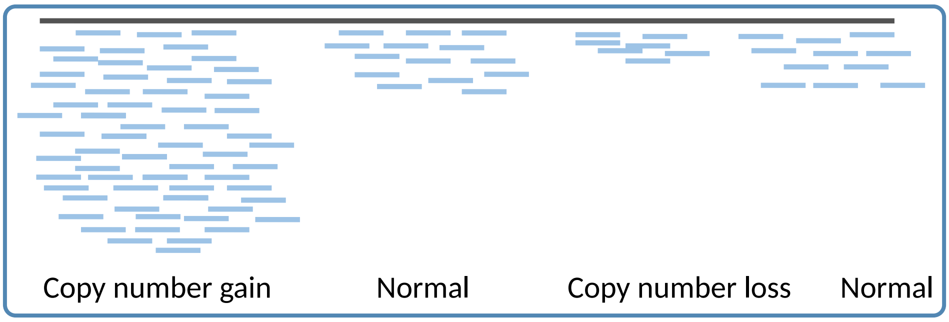

Figure 4 Depth of coverage approach: the gray bar represents the reference genome, and the light blue bars represent the sample fragmented DNA. Areas with more depth are presumed to have copy number gain and areas with low depth are suspected to have a copy number loss.

Table 2 Review of Literature studying the utility of LPGS and CMA

Sl.	Authors	Depth of sequencing	Type of sample	Sample size	Concordance with CMA	Yield
1.	Dong et al., 2017	0.25X	POC Stillbirth Prenatal Postnatal	198 37 149 186	188 (95%) 34 (91.9%) 141 (94.6%) 186 (100.0%)	
2.	Wang et al., 2020	0.25X	Prenatal	1023		13.5% VUS-5.2%
3.	Deleye et al., 2015	0.3-0.4X	Trophectoderm biopsy in preimplantation genetic diagnosis (PGD) in translocation carriers	47 blastocysts (5 normal and 42 abnormal)	100%	
4.	Chaubey et al., 2020	5X	Variable	331 33	100%	17.2%
5.	Wang et al., 2020	1-2X	ID/DD + congenital anomalies/ autism/ no anomaly	95		16.84%
6.	Ye et al., 2020	0.5-1.9X	NIPT in singleton pregnancies	873	67.31% (for CNV>2Mb-81.58%; for CNV<2Mb-21.43%)	
7.	Chau et al., 2020	0.25X	Varied	532		22.4%

POC – Products of conception; ID – Intellectual disability; DD – Developmental delay; CNV – Copy number variation; NIPT – Non-invasive prenatal testing

only way to detect breakpoint and copy number with good sensitivity. This is the reason why one of the studies done for CNV in NIPS with LPGS had low sensitivity when the size of CNV was <2Mb (**Table 2**). Apart from increasing the time and cost, needless to say, increasing the depth would undermine the very purpose of LPGS.

Advantages of LPGS

One may note that the diagnostic yield from LPGS is not drastically different from that of CMA. But LPGS, with its high precision and accuracy allows fine mapping of deletions and duplications. Other advantages include:

1. Identifying cryptic CNVs located in regions with insufficient probe coverage on CMA platforms (Xiao et al., 2020)
2. Increased sensitivity in detecting low-level mosaicism (Wang et al., 2020)
3. It is more useful in prenatal cases due to: i. shorter turn-around time; ii. more accuracy iii. lesser cost; iv. higher resolution compared to CMA; v. lesser quantity of DNA required - CMA requires a larger quantity of DNA (300 ng) compared to LPGS (50 ng); and vi. reduced technical repeat rate from 4.6% for CMA to 0.5%.

Disadvantages of LPGS

The disadvantages of LPGS can mostly be attributed to its use of short read length, which is an integral part of second-generation sequencing. Other disadvantages of LPGS are its inability to detect triploidy and breakpoints in balanced translocation (Chaubey et al.,2020). Lastly, no reference standard exists to benchmark CNV calls from LPGS. Hence, it is difficult to compare studies using LPGS due to variable choice in methods of analysis and platforms used.

Validation of CNV calls from LPGS

Single nucleotide variants or small deletions/duplications reported in NGS are validated by Sanger sequencing especially in case of presence of pseudogene or poor read depth of the variant. For LPGS, validation of CNV calls may depend upon the sensitivity of the NGS method used to identify the smallest size of CNV, location of CNV in the genome and its correlation with population and disease databases. Nevertheless, CMA is currently the best option to validate CNV calls from LPGS. For CNVs involving single exon, quantitative PCR or Sanger sequencing may be used and for CNVs involving more than >1 exon to the entire gene, MLPA (multiplex ligation probe amplification) may be used. Fluorescence in situ hybridization (FISH) may be used for CNVs of approximately 100kb or more in size.

Current status

Various studies over the last few years done at varying low sequencing depths have found LPGS to have good concordance (67-100%) with CMA in prenatal and post-natal samples. The yield of LPGS is similar to cytogenetic microarray but its ability to identify copy number variants in areas without probe, better delineation of breakpoints, technical ease and low cost per sample is much better than cytogenetic microarray. However, at present, it has lower sensitivity for inversions, balanced translocations, loss of heterozygosity and small size of CNV.

Discussion

The available studies have reaffirmed the fact that the yield of LPGS is as good as CMA in varied types of samples (**Table 2**). The study on 1023 prenatal samples by Wang et al. (2020) showed that LPGS not only identified all 124 numerical disorders or pathogenic or likely pathogenic (P/LP) CNVs detected by CMA in 121 cases (11.8%, 121/1,023), but also defined 17 additional and clinically relevant P/LP CNVs in 17 cases (1.7%, 17/1,023). In addition, LPGS significantly reduced the technical repeat rate from 4.6% (47/1,023) for CMA to 0.5% (5/1,023) and required less DNA (50 ng) as input. Small but relevant CNVs detected in the study by LPGS include a 31.2-kb cryptic hemizygous deletion in the male fetus involving the 42nd exon of *DMD*, a 19.3-kb homozygous deletion characteristic for Southeast Asian (SEA) type alpha thalassemia. Another case from the same study by Wang et al. had a 298.7-kb maternally inherited heterozygous deletion involving exons 1–8 of *FBN2* in the fetus. Variants in *FBN2* cause ventricular septal defect and the mother had ventricular septal defect. Among the 16 deletions not detected by CMA, the reason was attributed to insufficient probe coverage in the target regions on the CMA platform. LPGS detected one case with low level of mosaicism for partial duplication of chromosome 8 (Wang et al., 2020). But the technique requires good quality DNA. Poor DNA quality has low concordance with CMA results for the same level of sequencing as seen in cases with fetal demise. Also, techniques like NIPT which work with very low amount of DNA, may also have lower sensitivity when working at low depth of sequencing (Xiaoqing et al., 2020). Although LPGS is presumed to be genome-wide and probe-free, there are certain regions of the genome which may not be well represented on sequencing. CNV if present in these areas may require validation by other methods. With rise in number of cases for sequencing, variants of uncertain significance are bound to increase. The study by Wang et al. (2020) not only detected all variations of unknown significance (VOUS) identified by CMA, but also revealed an additional six VOUS in six cases.

At present, population and disease databases for CNV calls from LPGS are yet to be functional. Hence the diagnostic utility of LPGS is yet to reach its zenith.

Conclusion

With reliable standards and availability of guidelines, traditional cytogenetics still holds supremacy in detecting triploidy and balanced translocations. But due to its high resolution, CMA is now considered as the tier I test for cytogenetic analysis. With improving sequencing and bioinformatics algorithms, LPGS may soon become a standard test in clinical settings. The available robust data has proved its reliability as compared to CMA and also added advantages of detecting mosaicism, cryptic breakpoints and better coverage of the genome than CMA. Though not included yet, a young contender for the throne for chromosomal analysis would be the third generation of sequencing which uses long read sequencing. This would allow us to overcome the bias due to short read length which is inherent to second-generation sequencing.

References

1. Chau MH, et al. Low-pass genome sequencing: a validated method in clinical cytogenetics. Hum Genet. 2020; 139: 1403–1415.
2. Chaubey A, et al. Low-pass genome sequencing: validation and diagnostic utility from 409 clinical cases of low-pass genome sequencing for the detection of copy number variants to replace constitutional microarray. J Mol Diagn. 2020; 22: 823–840.
3. Deleye L, et al. Shallow whole genome sequencing is well suited for the detection of chromosomal aberrations in human blastocysts. Fertil Steril. 2015; 104:1276–1285.
4. Dong Z, et al. Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. Genet Med. 2016; 18: 940–948.
5. Friedman JM. High-resolution array genomic hybridization in prenatal diagnosis. Published in Affiliation with the International Society for Prenatal Diagnosis. Prenat Diagn. 2009; 29: 20–28.
6. Fu F, et al. Whole exome sequencing as a diagnostic adjunct to clinical testing in fetuses with structural abnormalities. Ultrasound Obstet Gynecol. 2018; 51: 493–502.
7. Martin CL, Warburton D. Detection of chromosomal aberrations in clinical practice: from karyotype to genome sequence. Annu Rev Genomics Hum Genet. 2015; 16: 309–326.
8. Sadek AA, Mohamed MA. Yield of karyotyping in children with developmental delay and/or dysmorphic features in Sohag University Hospital, Upper Egypt. Egypt J Med Hum Genet. 2018; 19: 253–259.
9. Valsesia A, et al. The growing importance of CNVs: new insights for detection and clinical interpretation. Front Genet. 2013; 4: 92.
10. Wang H, et al. Low-pass genome sequencing versus chromosomal microarray analysis: implementation in prenatal diagnosis. Genet Med. 2020; 22: 500–510.
11. Wang J, et al. The diagnostic yield of intellectual disability: combined whole genome low-coverage sequencing and medical exome sequencing. BMC Med Genomics. 2020; 13: 1–5.
12. Xiao B, et al. Whole genome low-coverage sequencing concurrently detecting copy number variations and their underlying complex chromosomal rearrangements by systematic breakpoint mapping in intellectual deficiency/developmental delay patients. Front Genet. 2020; 11: 616.
13. Ye X, et al. Identification of copy number variants by NGS-based NIPT at low sequencing depth. Eur J Obstet Gynecol Reprod Biol. 2021; 256: 297–301.