BIOLOGY REPORTS

Published: 09 August 2010

The application of massively parallel sequencing technologies in diagnostics

Andreas Dahl[1,2]*, Florian Mertes[1], Bernd Timmermann[3] and Hans Lehrach[1]

Addresses: [1]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany; [2]Deep Sequencing Group – SFB655, Biotechnology Center TU Dresden, Tatzberg 47-49, 01307 Dresden, Germany; [3]Next Generation Sequencing Group, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany

* Corresponding author: Andreas Dahl (andreas.dahl@biotec.tu-dresden.de)

F1000 Biology Reports 2010, **2**:59 (doi:10.3410/B2-59)

Abstract

Massively parallel sequencing (MPS) is rapidly evolving and is starting to be utilized by the clinical field as well as diagnostics. We describe major recent advances that have come about as a result of the application of MPS in the biomedical field and the first approaches in medical genetics that have made use of MPS. Without any doubt, MPS has proven to be a very powerful technique. To unravel the capabilities of MPS for patient care, the most important aspect for the acceptance of MPS within clinics and diagnostics is to guarantee that the large amount of data undergoes vitally important analyses and interpretation and is securely managed.

Introduction and context

The evolution of the so-called next- or second-generation sequencing technologies has been extremely rapid during the last 4 years. This is supposed to continue as the third generation of sequencers starts to enter the market with new approaches and devices that enable longer read lengths and shorter analysis time. Companies such as Pacific Biosciences (Menlo Park, CA, USA) promise to enable sequencing of samples in a few minutes to address the needs of diagnostics. However, the second-generation platforms such as the FLX from Roche (Basel, Switzerland), the Genome Analyzer from Illumina (San Diego, CA, USA), and the SOLiD platform from Applied Biosystems (Foster City, CA, USA) have proven to be very powerful tools for genomic analyses and have transformed biomedical science by opening up fascinating new opportunities. So far, the majority of projects performed on these platforms have been in the basic or biomedical research context. This might change by the recently achieved and more streamlined workflows for all commercially-available next-generation sequencing platforms. To appeal to smaller labs and the diagnostics market, the three leading manufacturers (Roche, Illumina, and Applied Biosystems) started to launch smaller-scaled devices. These instruments enable more flexibility and shorter processing times at reduced throughput.

Massively parallel sequencing (MPS) enables simultaneous screening of thousands of loci for disease-causing mutations, structural rearrangements, or epigenetic changes. On the RNA level, mutational analysis, post-transcriptional modifications, and the profiling of abundant transcripts become possible in one experiment. MPS is, for example, by far the best technique to analyze allele-specific expression or splicing, RNA editing, or to follow the exact genotype of the sequences involved in copy number variation (CNV). The strength of this technology is to allow investigators to look at several aspects in one single experiment. This allows many complex analyses to be simplified and probably will lead to the replacement of other technologies such as microarrays, which have been shown to be useful in the diagnostic setting (for instance, in the classification of cancer types) [1-3]. In contrast to microarray-based techniques, sequencing is essentially digital and therefore can provide almost unlimited accuracy (depending on the sequencing depth). However, the prospects of these enormous capacities have to be well

balanced with the requirements associated with daily routine use in diagnostics. In particular, a simple workflow with few requirements on sample preparation and barcoding, a reasonable time per operation, easy bioinformatics, and (last but not least) manageable costs for the instrument, its operation, and data management are obligatory. Other questions frequently discussed in the community, such as the required read length or the read quality, depend greatly on the planned application.

Major recent advances

In regard to the biological target, the analysis of genetic variation such as mutational analysis and structural variants is of major interest in diagnostics. The classical way of detecting causative mutations has been amplicon sequencing. An application of amplicon sequencing on MPS was shown by Varley and Mitra [4], who combined a polymerase chain reaction (PCR)-based enrichment approach to extract 94 exons from six genes that cause cancer when mutated in the germline (*TP53*, *APC*, *MLH1*, *RB1*, *BRCA1*, and *VHL*) using the 454 sequencing approach. But the technology itself is not limited to a few exons or amplicons. An example of genome-wide analysis was shown by Lupski and colleagues [5]. To unravel the causative mutations in a patient with Charcot-Marie-Tooth neuropathy, whole-genome sequencing was performed and clinically relevant variants were identified in the causative alleles, providing diagnostic information for the care of these patients [5]. Nevertheless, in the near future (1-3 years), it is more likely that analyses will be performed on target regions in diagnostics. The regions might differ from kilobases to megabases, and focus on specific loci, but with the improvement of data reduction and computation-aided analysis tools, the size of regions will increase in the coming years and finally we might end up on the genome-wide scale. However, recent work has shown the usefulness of targeted MPS. Choi and colleagues [6] made an unanticipated genetic diagnosis of congenital chloride diarrhea in a patient referred with a suspected diagnosis of Bartter syndrome. The molecular diagnosis was based on the finding of a homozygous missense mutation and could be confirmed by clinical follow-up [6]. Ng *et al.* [7] successfully identified a candidate gene causing the Miller syndrome by using exome enrichment followed by MPS and successive mutation analysis. An example of a further candidate for the targeted approach is hypertrophic cardiomyopathy (HCM), a heterogeneous autosomal dominant cardiac disorder with a prevalence of 1 in 500. So far, more than 450 different pathogenic mutations in at least 16 genes have been identified with alternative techniques. The large allelic and genetic heterogeneity of HCM requires high-throughput, rapid, and affordable mutation detection technologies, which can be provided by MPS [8].

A significant advantage of MPS is the ability to detect even rare variants, which are not represented in the common single-nucleotide polymorphisms (SNPs) usually scored by array-based SNP genotyping techniques. Rare variants were recently shown to be significantly correlated with the risk for schizophrenia, whereas no significant association could be found for common variants [9]. The comprehensive assessment of variants for a patient in a single experiment can be used to determine dose requirements and the susceptibility to adverse drug effects of current and future novel drugs. However, sequencing, in parallel, will provide information on CNVs and translocations and therefore will help to get a much more complete picture of all potentially relevant changes in the genome. MPS also allows easy integration of information on the transcriptome (expression levels, splicing, and RNA editing), DNA methylations (e.g., by the use of MeDIP [methylated DNA immunoprecipitation]-based procedures), or protein-DNA interactions (chromatin immunoprecipitation sequencing [ChIP-seq]) on the same material. This is particularly interesting for cancer treatment. Experiences from sequencing tumor genomes have shown that tumors typically have tens of thousands of somatic changes, making every tumor different and therefore making the response of every patient to a particular treatment an individual response. It is now well understood that tumors of identical clinical classification may require very different treatments [10]. The genetic instability and clonal evolution of cancer genomes lead to very heterogeneous tissues. This can result in the misleading interpretation of data coming from array-based or PCR-based analyses. At a sufficient depth of coverage, MPS enables the quantification of even low abundant sequences and thus their accurate detection.

Diagnosis should be as non-invasive a process as possible. MPS represents a new approach that is potentially applicable to non-invasive diagnosis in all body fluids. This was shown by Chiu *et al.* [11] for maternal plasma samples, which were screened for fetal chromosomal aneuploidies. For epigenetic changes of cell-free DNA in blood, a similar approach is applicable as well, as it has been shown that these mutations may act as diagnostic or early detection/risk markers for cancer [12]. In this context, targeted approaches [13] and genome-wide approaches [14] have been demonstrated to access the methylation status. Alternatively, genome-wide histone modifications have been identified by combining ChIP with MPS [15].

For detection and quantification of viral or bacterial populations, MPS provides the ability to identify rare or even currently unknown microorganisms by their

sequence. Recent examples have shown a combination of targeted enrichment of an informative genomic sequence followed by deep sequencing. Claesson and colleagues [16] used the V4 and V6 regions of 16S ribosomal RNA genes in bacterial DNA to decipher the microbial spectrum in the human intestinal tract. Holtz *et al*. [17] identified a new picornavirus by phylogenetic analysis of deep sequencing data of a sample derived from a patient. Wang and colleagues [18] used ultra-deep pyrosequencing to detect minor sequence variants in HIV-1 protease and reverse transcriptase genes from clinical plasma samples. With appropriate analysis, ultra-deep sequencing is a promising method for characterizing genetic diversity and detecting minor yet clinically relevant variants in biological samples with complex genetic populations. For a wide range of applications, the short-read-delivering technologies (not more than 201 base pairs) are well suited and are advantageous in regard to throughput and costs per experiment. But to distinguish between different species with a high degree of homology or to detect structural variants, longer reads are required. Paired-end and mate-pair sequencing can help to circumvent this [19]. But the complex sample preparation procedure and the required high amounts of input DNA make mate-pair sequencing in particular improper for routine diagnostics.

Next-generation sequencers produce an enormous amount of data. Currently, the instruments have a weekly data output of approximately 400 gigabytes to 1 terabyte. Large genome centers are prepared to deal with this, but the majority of diagnostics laboratories are not. Tremendous computing and storage capacities are still needed at the moment. Cloud computing is discussed as one possible way to circumvent investments of zillions of dollars from the diagnostics community in information technology infrastructure [20]. Langmead and colleagues [21] recently reported the development of software that uses cloud computing to enable the analysis of a human genome within one day. Currently, the transfer of data is still limiting, facing potentially hundreds of gigabytes of data from a single experiment. However, the concerted improvements of data size reduction and data transfer capacities might solve this in the near future. For the handling of patient data, data safety is of particular concern. Standards for the privacy and security of health-related data have to be established. Superior in this respect is market leader Amazon (Seattle, WA, USA). The company committed itself to data security by its compliance with the Health Insurance Portability Act [20]. Besides the technical issues that need to be solved, there are ethical questions. Patients will gain a huge amount of information describing potential risks and genetic predispositions. Future advances in medical research could mean that people end up discovering things that they might not have wanted to know, an issue that needs to be resolved both ethically and legally.

Future directions

By looking at the requirements of smaller and diagnostics laboratories, companies have started to launch smaller devices and more streamlined workflows, which enable the continuous analysis of many samples within a relatively short period of time. We expect that, in the next 1-3 years, targeted enrichment will enable the set up of disease-focused applications. This, in combination with barcoding, gives the flexibility needed in the diagnostics setting. On the other hand, we can expect sequencing costs to drop further, sequencing speed to increase dramatically, and third-generation sequencing techniques, based (for example) on fluorescence detection (Pacific Biosciences), alternative detection systems (Ion Torrent, Guilford, CT, USA), or nanopore-based techniques (Oxford Nanopore Technologies Ltd, Oxford, UK), to allow the routine determination of the sequence of entire genomes at low cost in less than an hour. Therefore, at sufficiently low cost and sufficiently high speed, whole-genome/transcriptome sequencing, in the long run, might be as cost-effective as enrichment-based strategies.

In the context of diagnostics, simpler sample preparation is needed as are new methods to handle data and to assess statistical significance without immense bioinformatics support in the clinical routine. Quality metrics will help to access technical reproducibility, accuracies of raw base calls, or systematic error patterns. Cloud computing providing standard analysis pipelines might be an option for data analysis, but the data safety issue has to be solved technically as well as legally. Data storage and transfer have to be of major concern in the near future so that we are prepared for the application of this technology in diagnostics. This will crucially influence the arrival of MPS in diagnostics.

Methods for data storage, management, and analysis and the suitable workflow to fit into diagnostics are technologically-oriented developments needed to tightly link MPS to diagnostics routine. A major difficulty, however, will be to efficiently transform the large amounts of sequence information, which are increasingly easy to generate, into clinically relevant information. For this, few systematic approaches are seen in the field. One example is the Treat 1000 project [22], which is going to develop individually optimized treatments for patients with cancer. Modeling tools will be used to generate models of the drug response of individual patients on the basis of data obtained by deep sequencing of the genome and transcriptome of the tumor and the genome of the patient.

Although such tools are promising, we still need additional biological knowledge of complex disorders in order to develop robust models into which the obtained comprehensive data are fed. It is hoped that the safe interpretation and prediction that result from this will lead to a personalized medical care that earns the trust of the patient.

Abbreviations

ChIP, chromatin immunoprecipitation; CNV, copy number variation; HCM, hypertrophic cardiomyopathy; MPS, massively parallel sequencing; PCR, polymerase chain reaction; SNP, single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

References

1. de Tayrac M, Etcheverry A, Aubry M, Saïkali S, Hamlat A, Quillien V, Le Treut A, Galibert MD, Mosser J: **Integrative genome-wide analysis reveals a robust genomic glioblastoma signature associated with copy number driving changes in gene expression.** *Genes Chromosomes Cancer* 2009, **48:**55-68.
2. Sargent R, Jones D, Abruzzo LV, Yao H, Bonderover J, Cisneros M, Wierda WG, Keating MJ, Luthra R: **Customized oligonucleotide array-based comparative genomic hybridization as a clinical assay for genomic profiling of chronic lymphocytic leukemia.** *J Mol Diagn* 2009, **11:**25-34.
3. Schwaenen C, Viardot A, Berger H, Barth TF, Bentink S, Döhner H, Enz M, Feller AC, Hansmann ML, Hummel M, Kestler HA, Klapper W, Kreuz M, Lenze D, Loeffler M, Möller P, Müller-Hermelink HK, Ott G, Rosolowski M, Rosenwald A, Ruf S, Siebert R, Spang R, Stein H, Truemper L, Lichter P, Bentz M, Wessendorf S; Molecular Mechanisms in Malignant Lymphomas Network Project of the Deutsche Krebshilfe: **Microarray-based genomic profiling reveals novel genomic aberrations in follicular lymphoma which associate with patient survival and gene expression status.** *Genes Chromosomes Cancer* 2009, **48:**39-54.
4. Varley KE, Mitra RD: **Nested Patch PCR for highly multiplexed amplification of genomic loci.** *Cold Spring Harb Protoc* 2009, **7:** pdb.prot5252.
5. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.** *N Engl J Med* 2010, **362:**1181-91.

   F1000 Factor 8.6 *Exceptional*
   Evaluated by Sue Malcolm 15 Mar 2010, Anthony Antonellis 15 Mar 2010, Thomas Friedman 23 Mar 2010, Klaus-Armin Nave 24 Mar 2010, Andrea Ballabio 16 Apr 2010
6. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP: **Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.** *Proc Natl Acad Sci U S A* 2009, **106:**19096-101.
7. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: **Exome sequencing identifies the cause of a mendelian disorder.** *Nat Genet* 2010, **42:**30-5.

   F1000 Factor 8.3 *Exceptional*
   Evaluated by Stephen Scherer 25 Jan 2010, Steven Salzberg 03 Feb 2010, Michele Ramsay 17 Feb 2010, Terri Beaty 02 Jul 2010
8. Fokstuen S, Lyle R, Munoz A, Gehrig C, Lerch R, Perrot A, Osterziel KJ, Geier C, Beghetti M, Mach F, Sztajzel J, Sigwart U, Antonarakis SE, Blouin JL: **A DNA resequencing array for pathogenic mutation detection in hypertrophic cardiomyopathy.** *Hum Mutat* 2008, **29:**879-85.
9. Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL, Shianna KV, Yoon W, Kasperaviciūte D, Gennarelli M, Strittmatter WJ, Bonvicini C, Rossi G, Jayathilake K, Cola PA, McEvoy JP, Keefe RS, Fisher EM, St Jean PL, Giegling I, Hartmann AM, Möller HJ, Ruppert A, Fraser G, Crombie C, Middleton LT, St Clair D, Roses AD, Muglia P, Francks C, *et al.*: **A genome-wide investigation of SNPs and CNVs in schizophrenia.** *PLoS Genet* 2009, **5:**e1000373.
10. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, *et al.*: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318:**1108-13.

    F1000 Factor 6.5 *Must Read*
    Evaluated by J Steven Leeder 27 Nov 2007, John Nemunaitis 15 Aug 2008, Kai Zinn 11 Dec 2007
11. Chiu RW, Chan KC, Gao Y, Lau VY, Zheng W, Leung TY, Foo CH, Xie B, Tsui NB, Lun FM, Zee BC, Lau TK, Cantor CR, Lo YM: **Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma.** *Proc Natl Acad Sci U S A* 2008, **105:**20458-63.
12. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S, Jones A, Lechner M, Beck S, Jacobs IJ, Widschwendter M: **An epigenetic signature in peripheral blood predicts active ovarian cancer.** *PloS One* 2009, **4:**e8274.
13. Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang MQ, Ye K, Bhattacharjee A, Brizuela L, McCombie WR, Wigler M, Hannon GJ, Hicks JB: **High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing.** *Genome Res* 2009, **19:**1593-605.
14. Bormann Chung CA, Boyd VL, McKernan KJ, Fu Y, Monighetti C, Peckham HE, Barker M: **Whole methylome analysis by ultra-deep sequencing using two-base encoding.** *PloS One* 2010, **5:**e9320.
15. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129:**823-37.

    F1000 Factor 7.0 *Must Read*
    Evaluated by Steven Henikoff 22 May 2007, Xing Wang Deng 05 Jun 2007, Michael Meisterernst 19 Jun 2007, Deyou Zheng 29 Jun 2007, Magdalena Zernicka-Goetz 15 Jan 2008
16. Claesson MJ, O'Sullivan O, Wang Q, Nikkilä J, Marchesi JR, Smidt H, de Vos WM, Ross RP, O'Toole PW: **Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine.** *PloS One* 2009, **4:**e6669.
17. Holtz LR, Finkbeiner SR, Kirkwood CD, Wang D: **Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea.** *Virol J* 2008, **5:**159.
18. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW: **Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance.** *Genome Res* 2007, **17:**1195-201.
19. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: **Paired-end mapping reveals extensive structural variation in the human genome.** *Science* 2007, **318:**420-6.

    F1000 Factor 6.4 *Must Read*
    Evaluated by Kenneth Zaret 26 Nov 2007, Heng Zhu 24 Apr 2008
20. Sansom C: **Up in a cloud?** *Nat Biotechnol* 2010, **28:**13-5.
21. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL: **Searching for SNPs with cloud computing.** *Genome Biol* 2009, **10:**R134.
22. **Treat 1000 homepage.** [www.treat1000.org]